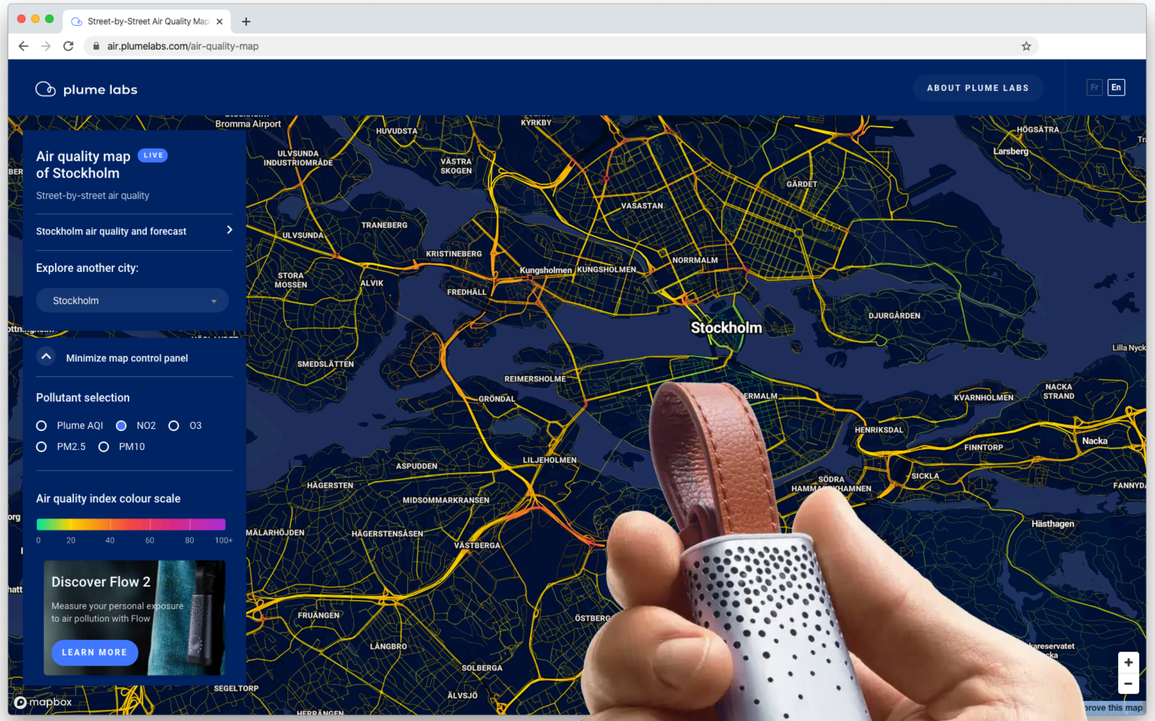
Map, Model, Measure: **AI for Biomolecules**

Romain Lacombe <rlacombe@stanford.edu>
Stanford Chemical Engineering

About me

- Physics/Math undergrad (France) and Engineering Systems MS (MIT)
- Climate tech founder: **Plume Labs**, powers 1 in 4 smartphones globally
- Acquired by AccuWeather, lead AI for weather and climate team
- **Stanford ChemE**: MS (HCP)
- **AI for Science:** climate, materials, and biomolecular engineering



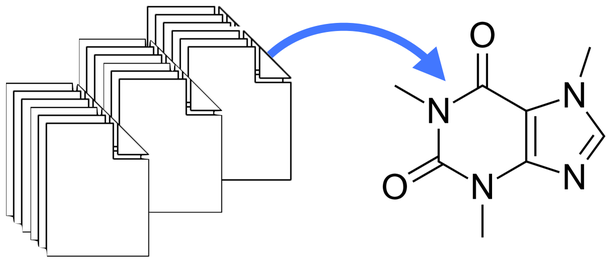
Plume Labs:
Street-level air quality map
Flow 1 air quality sensor

More on my work:
romainlacombe.com

AI for Biomolecules: 3 papers

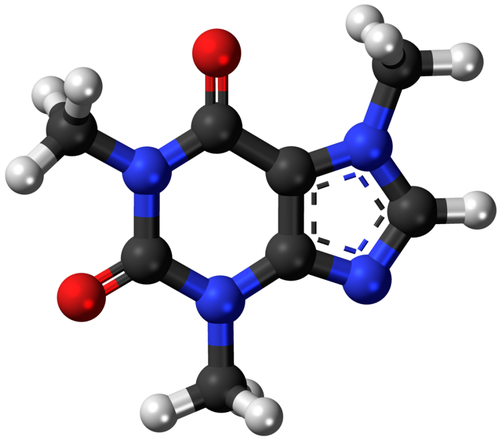
Map:

Can we predict properties of molecules from science papers?



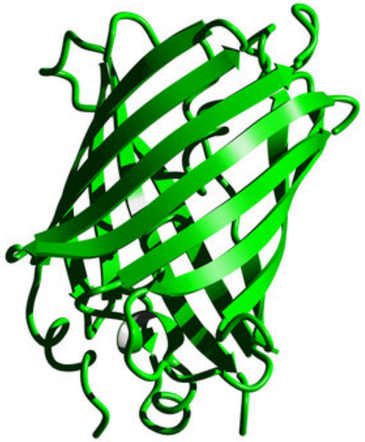
Model:

Can we accelerate generation of molecule conformations?



Measure:

How good are protein models out of their evolutionary domain?



Map, Model, Measure: AI for Biomolecules

- Extracting Molecular Properties from Natural language with Multimodal Contrastive Learning GNNs LMs evaluation
  ICML 2023 Computational Biology
  ACS Fall 2023 AI for Organic Chemistry Workshop

- Accelerating the Generation of Molecular Conformations with Progressive Distillation of Equivariant Latent Diffusion Models GNNs diffusion models
  ICLR 2024 – Generative and Experimental Perspectives for Biomolecular Design

- Non-Canonical Crosslinks Confound Evolutionary Protein Structure Models
  Experimental Design for AI in Science 2025 protein models evaluation

Map, Model, Measure: AI for Biomolecules >>

Extracting Molecular Properties from Natural Language with Multimodal Contrastive Learning

Romain Lacombe, Andrew Gaut, Jeff He, David Lüdeke, Kateryna Pistunova

ICML 2023 Computational Biology Workshop
ACS Fall 2023 AI for Organic Chemistry Workshop

[arXiv 2307.12996] GNNs LMs evaluation

Stanford University

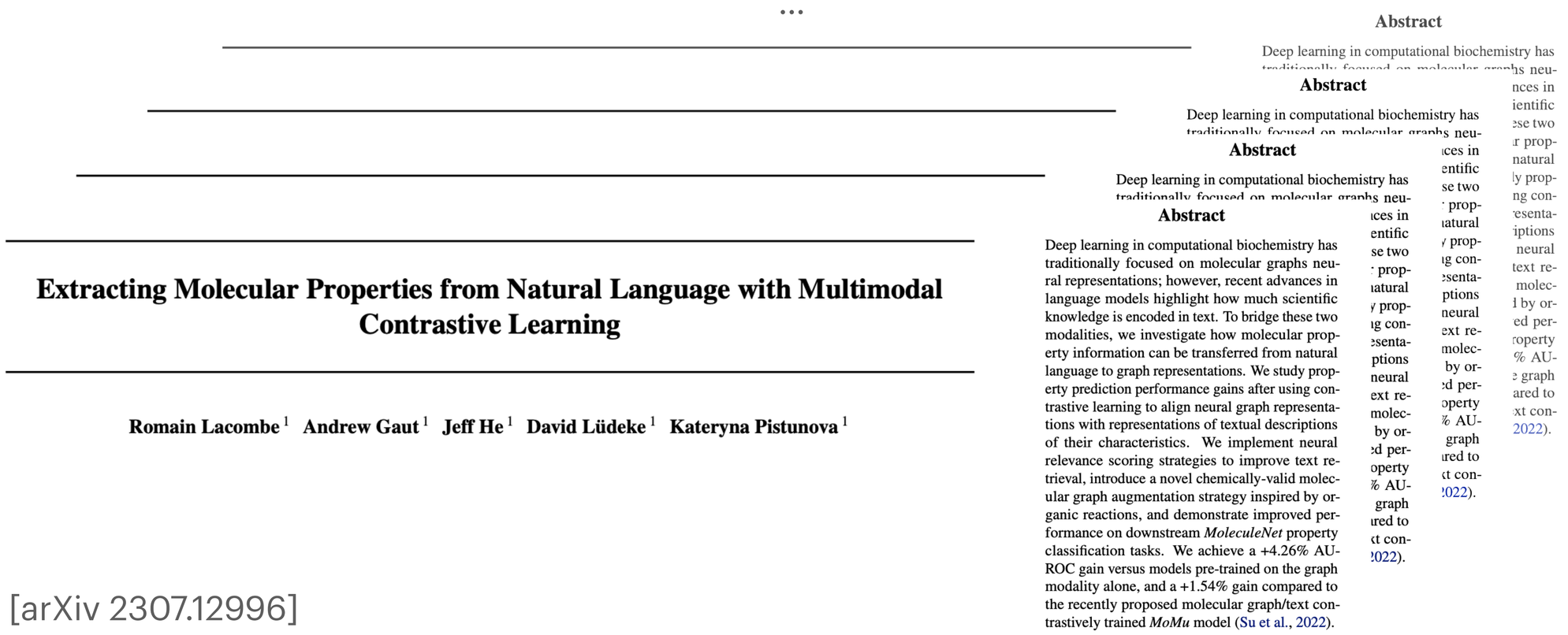
ICML International Conference On Machine Learning

ACS Chemistry for Life® American Chemical Society

Can AI learn chemistry from science papers?

Extracting molecular properties from text?

Treasure trove of collective knowledge now accessible.



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]

[arXiv 2307.12996]

Abstract

Deep learning in computational biochemistry has traditionally focused on molecular graphs neural representations; however, recent advances in language models highlight how much scientific knowledge is encoded in text. To bridge these two modalities, we investigate how molecular property information can be transferred from natural language to graph representations. We study property prediction performance gains after using contrastive learning to align neural graph representations with representations of textual descriptions of their characteristics. We implement neural relevance scoring strategies to improve text retrieval, introduce a novel chemically-valid molecular graph augmentation strategy inspired by organic reactions, and demonstrate improved performance on downstream *MoleculeNet* property classification tasks. We achieve a +4.26% AUROC gain versus models pre-trained on the graph modality alone, and a +1.54% gain compared to the recently proposed molecular graph/text contrastively trained *MoMu* model (Su et al., 2022).

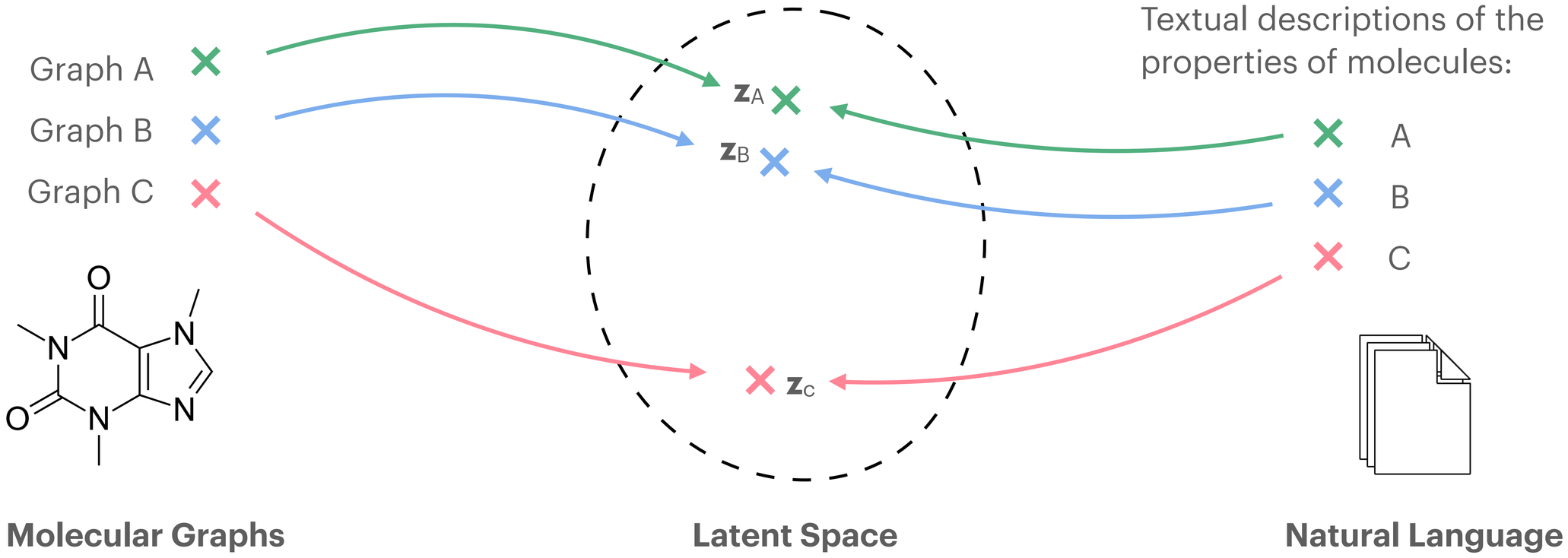
Contrastive learning

Self-supervised learning of molecule representations.

- Tasks in ML for chemistry require **deep molecular graph representations**
- GNNs can be trained to learn effective representations through **contrastive learning:**

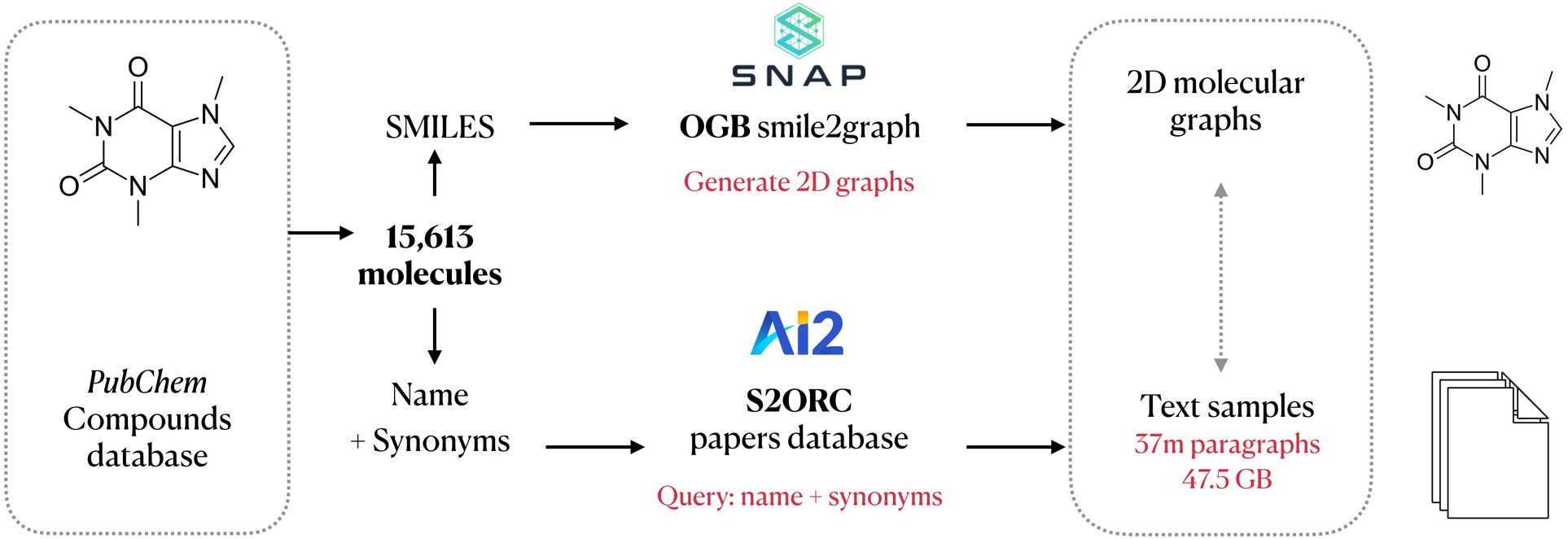
Multimodal contrastive learning

Align graph and text representations in latent space.



Joint dataset of molecules and papers

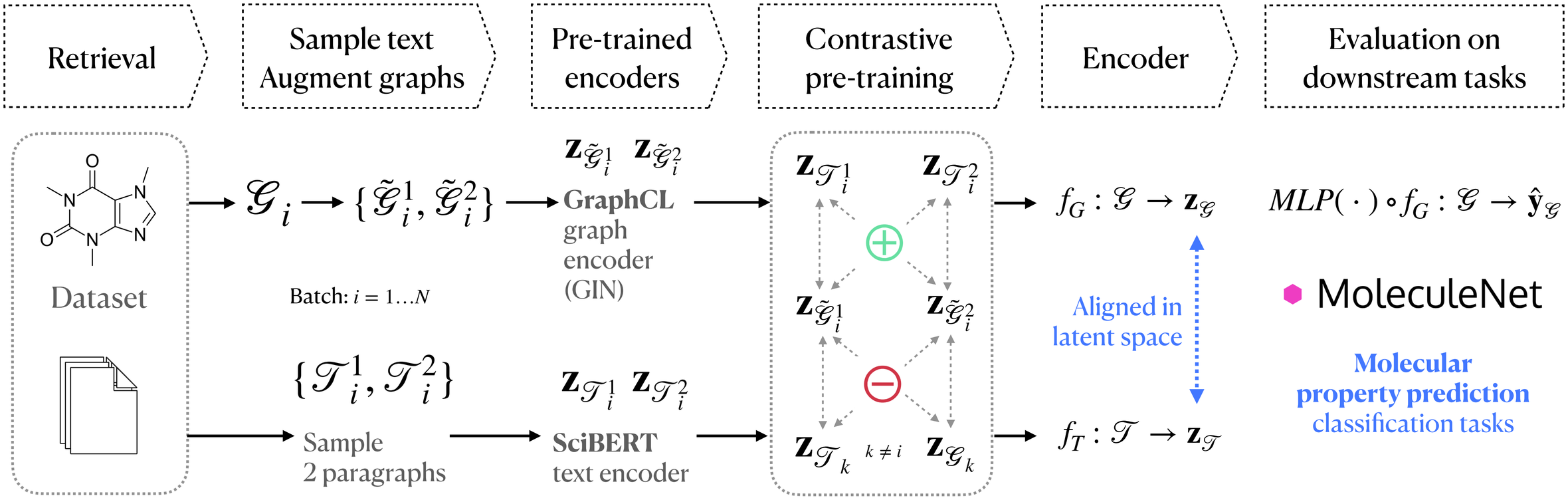
PubChem molecules and PubMed papers



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

Aligning graph and text representations

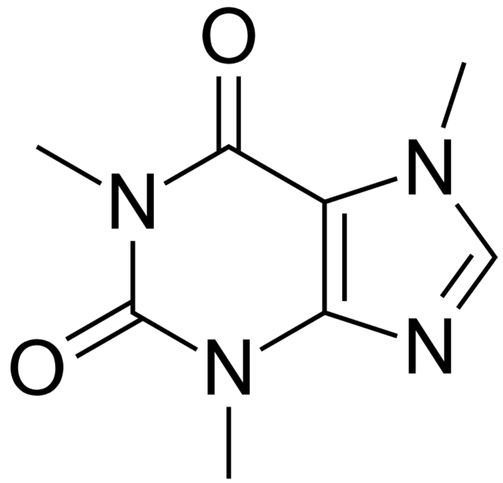
Using multimodal contrastive learning.



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

Su et al. 2022: https://arxiv.org/abs/2209.05481
Liu et al. 2022: https://arxiv.org/abs/2212.10789

Could we generate molecules from text?



Text prompt
`make me coffee’

Molecular graph
Caffeine ☕

Answer: yes!

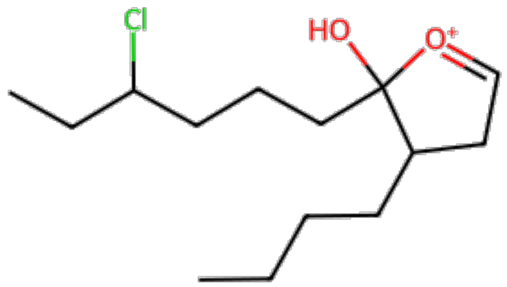
But not very well.

Prompt

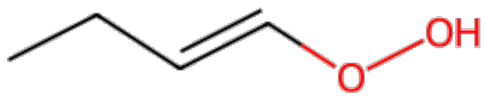
“This molecule has a hydroxyl group and a carbonyl group”

“This molecule is hazardous for health”

Generation



Has hydroxyl, but no carbonyl group (furan cycle)



⚠️ 1-Hydroperoxybut-2-ene: **unstable and explosive** (!!)

How can we improve performance?

How can we *measure* improvements?

Experiment: evaluation

MoleculeNet benchmark.

Evaluate graph representations on 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.

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

MoleculeNet

Wu et al. 2017: https://arxiv.org/abs/1703.00564
Su et al. 2022: https://arxiv.org/abs/2209.05481
Liu et al. 2022: https://arxiv.org/abs/2212.10789

Results

Graph only

Graph +natural language

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

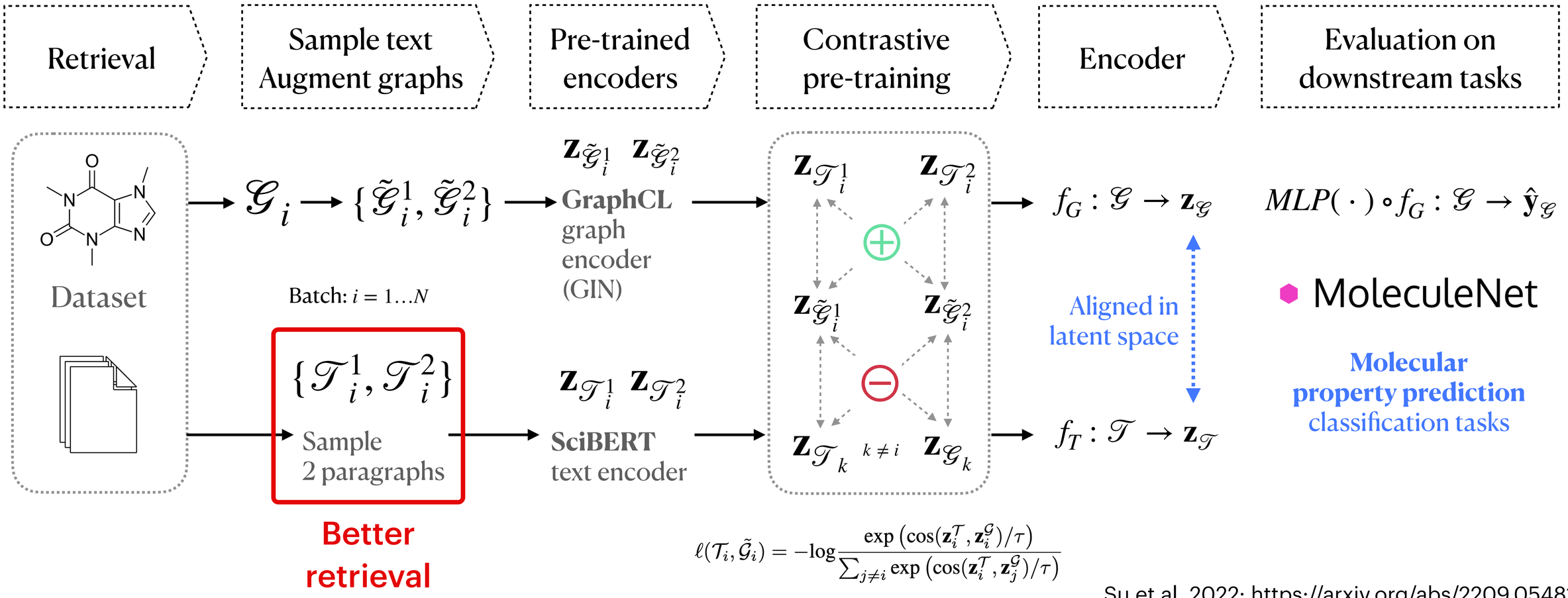
Graph only: GraphCL self-supervised

MoMu: GraphCL & SciBERT (Su et al. 2022)
Pruned: shorter paragraph (control for noise)
Relevant: only paragraphs with name of molecule + top 20 synonyms

Can we better select text paragraphs?

Improve text retrieval with LMs

Better sampling should extract better information.



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

Su et al. 2022: https://arxiv.org/abs/2209.05481
Liu et al. 2022: https://arxiv.org/abs/2212.10789

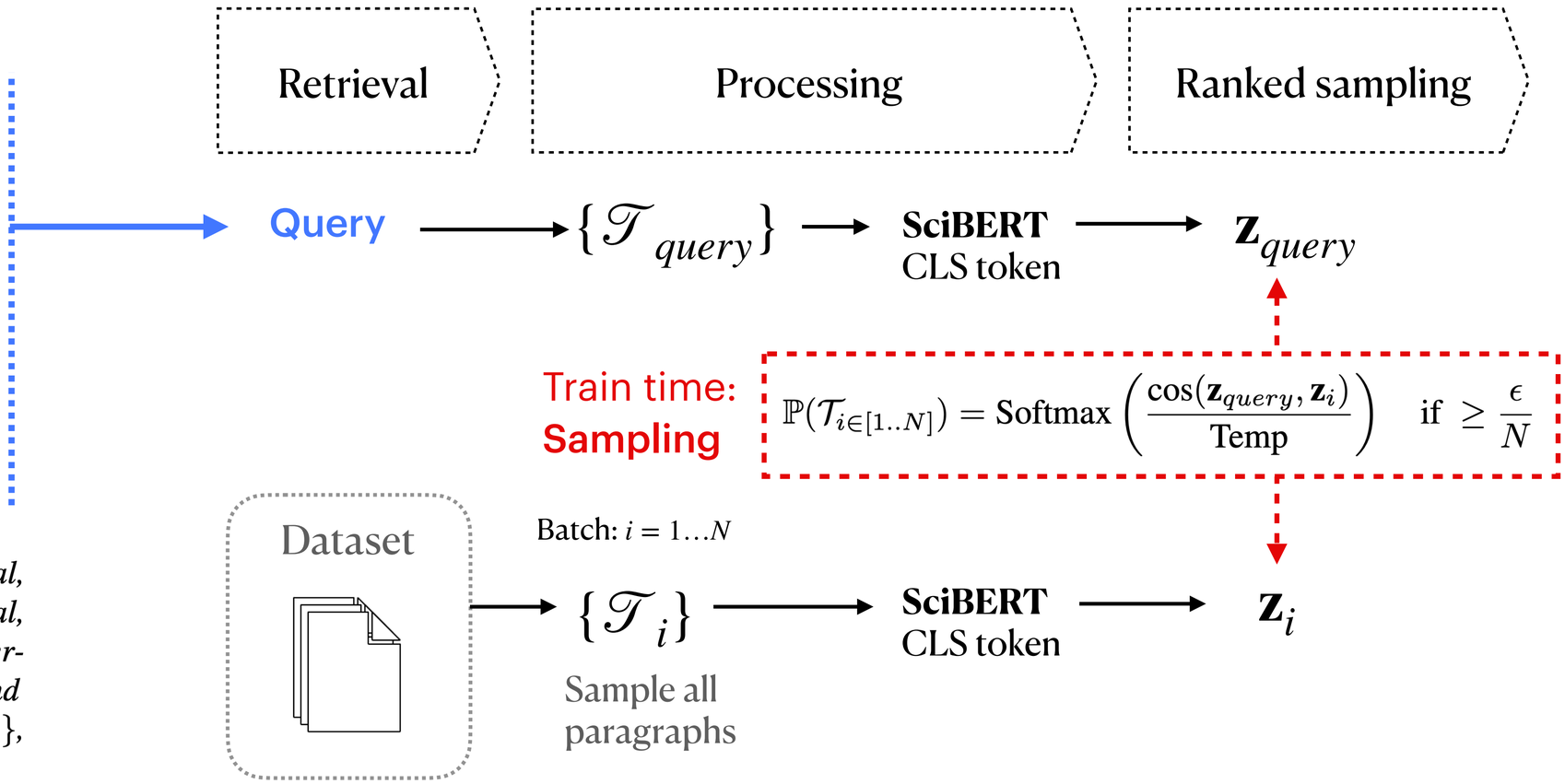
Neural retrieval to improve semantic relevance

Epsilon sampling (Hewitt et al.) for top cosine similarity sentences.

Query for cosine similarity:

- **Mean:** cosine similarity with mean of CLS tokens for the top 20 synonyms
- **Max:** max cosine sim with any of top 20 synonyms CLS token
- **Sentence:** cosine sim with CLS token of a query in natural language:

"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$_N$}."



Hewitt et al. 2022: https://arxiv.org/abs/2210.15191

Results: neural retrieval improves performance

Graph only

Graph +natural language

Improved retrieval

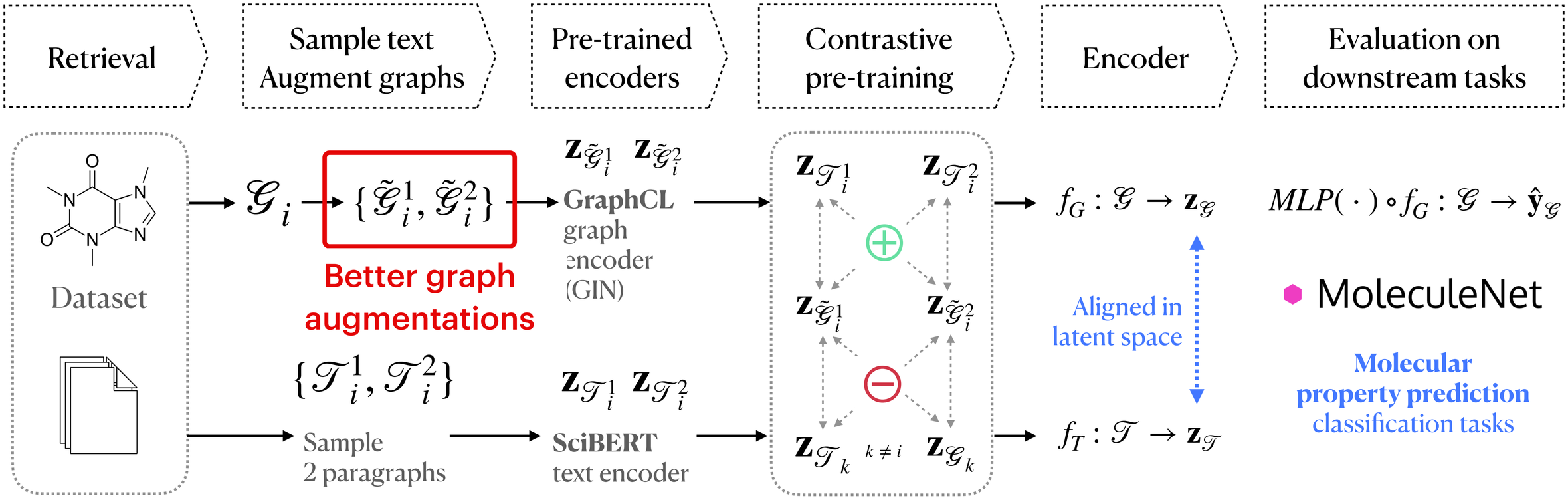
Experiment	BACE	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV
Graph only pre-training	70	65.8	74	63.4	57.3	58	71.8
Baseline (*MoMu*)	70.31 ±3.67	68.04 ±1.67	74.6 ±0.68	63.27 ±0.53	59.39 ±0.51	61.09 ±1.1	**75.66 ±0.55**
Baseline (pruned)	71.14 ±1.93	67.86 ±2.1	74.77 ±0.37	62.71 ±1.3	59.31 ±0.72	61.17 ±1.39	75.18 ±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

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.

Better inductive bias for chemistry?

Improve graph augmentations

More principled augmentations should learn better representations.



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

Su et al. 2022: https://arxiv.org/abs/2209.05481
Liu et al. 2022: https://arxiv.org/abs/2212.10789

Baseline: random graph augmentations

Baseline for molecular representations learning

- **GraphCL** (You et al. 2020) contrastive pre-training uses 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

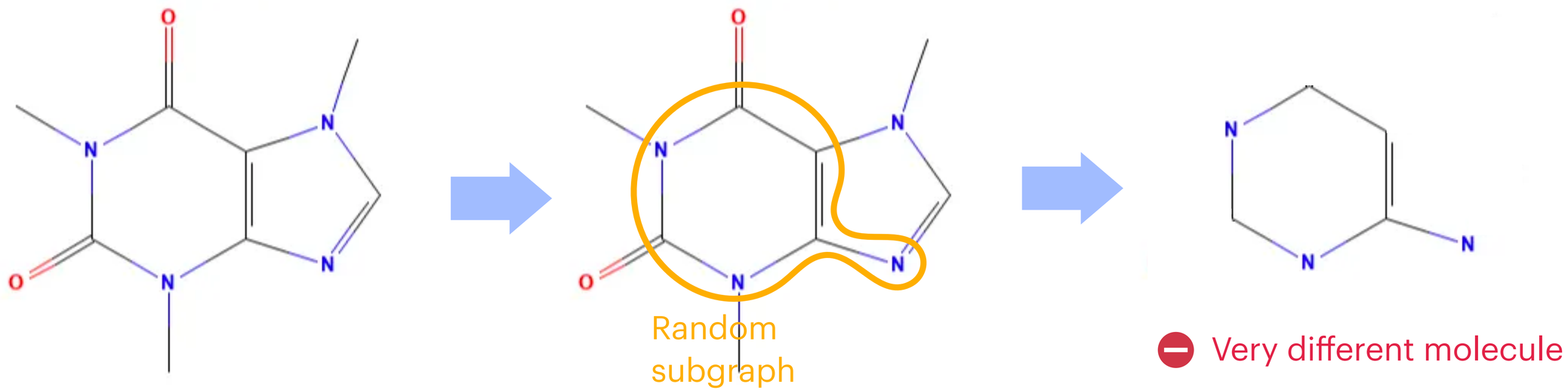
No guarantee that augmented graphs are valid molecules!

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

Random graph augmentations are suboptimal

Small changes lead to large differences in chemical space.

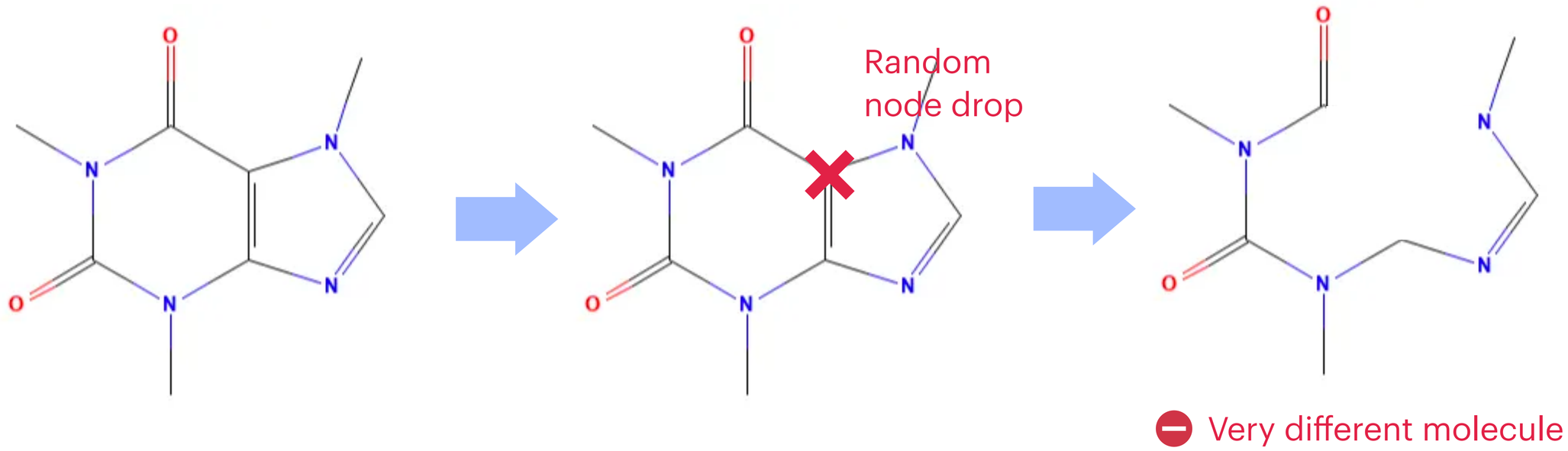
- *Ex:* random subgraph.



Random graph augmentations are suboptimal

Small changes lead to large differences in chemical space.

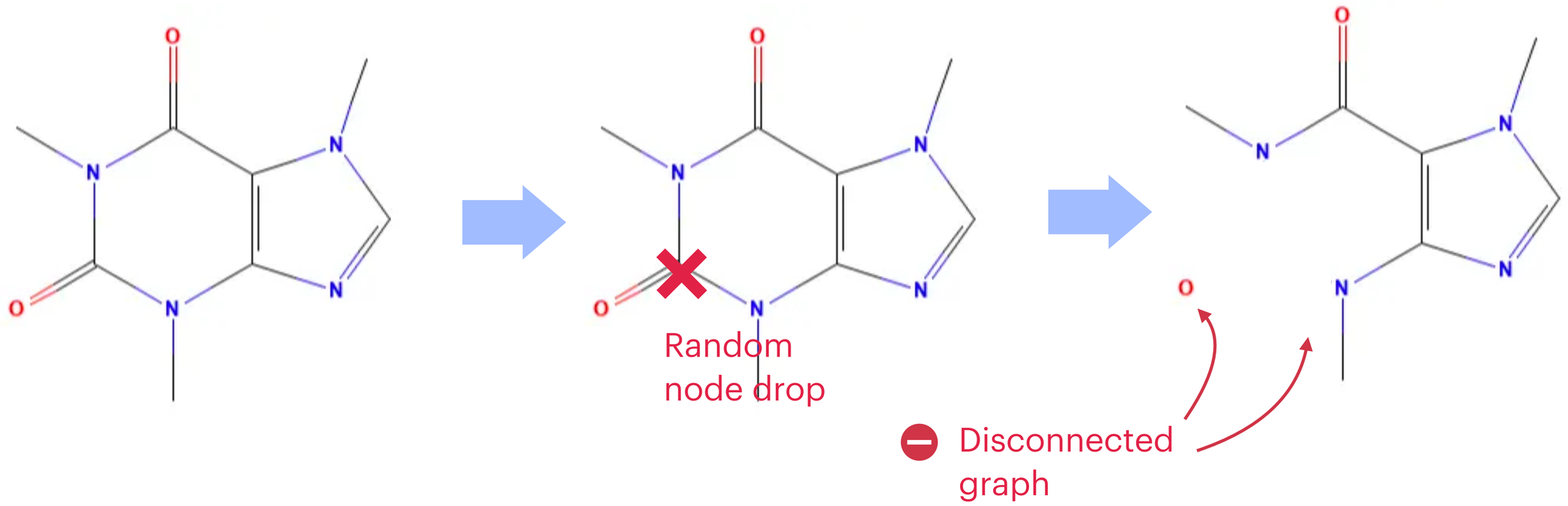
- *Ex:* drop random atom.



Random graph augmentations are suboptimal

Small changes lead to large differences in chemical space.

- *Ex:* drop random atom.

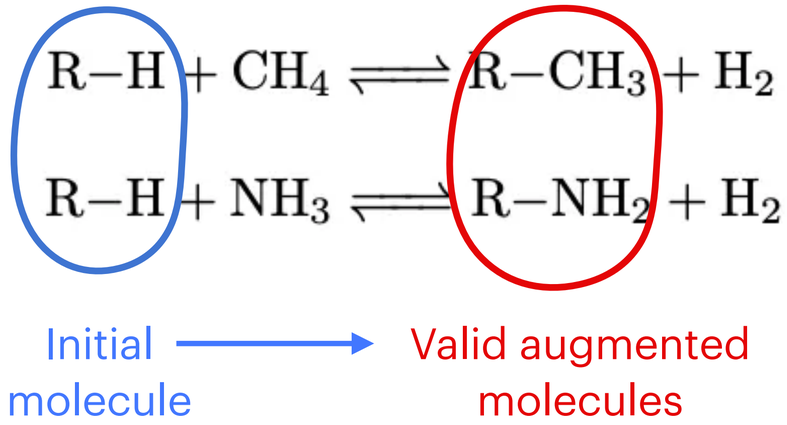


Idea: use chemical reactions!

Nature already provides principled graph augmentations.

Idea: use addition/elimination organic reactions!

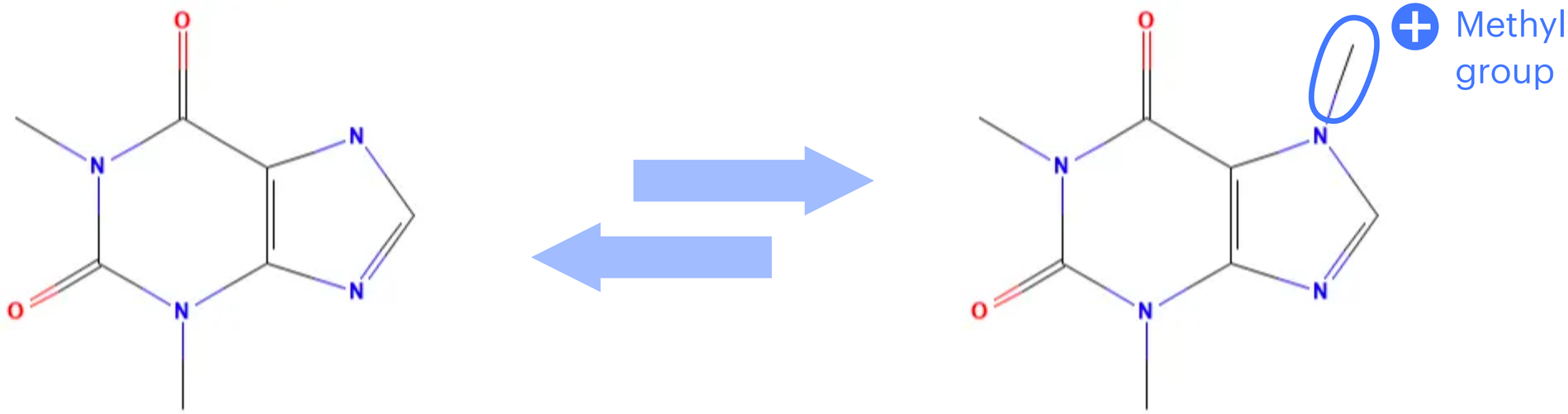
Transform initial molecular graph into better behaved augmentations through valid chemical reactions!



Idea: use chemical reactions!

Nature already provides principled graph augmentations.

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

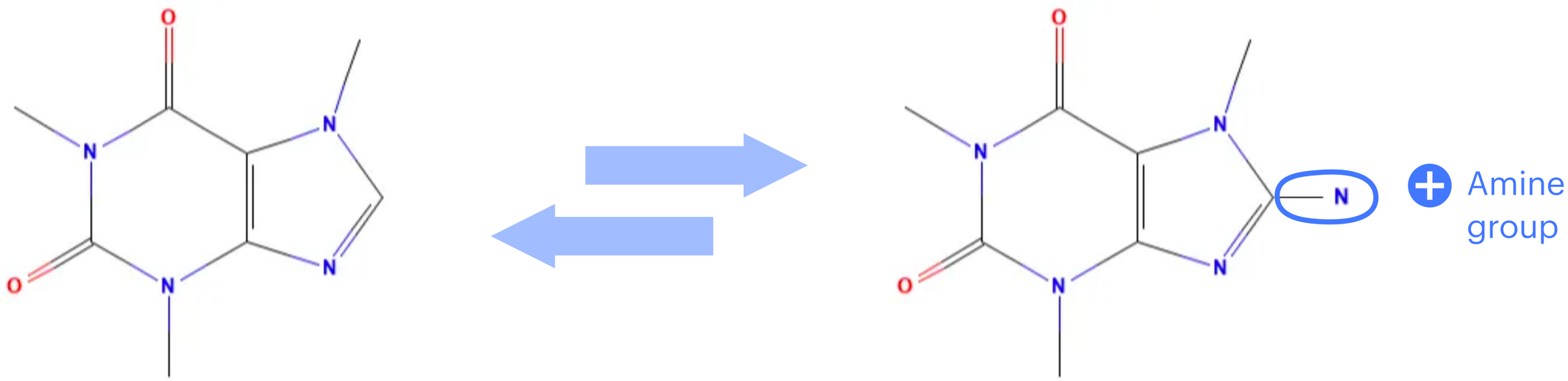


Valid + close to original molecule

Idea: use chemical reactions!

Nature already provides principled graph augmentations.

- *Ex:* amination/de-amination. $\mathrm{R{-}H + NH_3 \rightleftharpoons R{-}NH_2 + H_2}$



Valid + close to original molecule

Final results

Graph only

Graph +natural language

Improved retrieval

Better graph augmentations

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.

Take aways

Conclusions and future work

- **Improved retrieval helps extract information from natural language.**
  - AUROC performance metric for molecular property prediction improves by an average of **+4.26% across MoleculeNet classification tasks**
- Principled graph augmentations inspired from chemistry improve inductive bias for molecular representation learning
  - What other inductive biases from nature can we capture?

Questions?

Extracting Molecular Properties from Natural Language with Multimodal Contrastive Learning

ICML 2023 Computational Biology Workshop | **ACS Fall 2023** AI for Organic Chemistry Workshop

[arXiv 2307.12996] GNNs LMs evaluation

Stanford University

ICML International Conference On Machine Learning

ACS Chemistry for Life® American Chemical Society

Map, Model, Measure: AI for Biomolecules >>

Accelerating the Generation of Molecular Conformations with Progressive Distillation of Equivariant Latent Diffusion Models

Romain Lacombe, Neal Vaidya

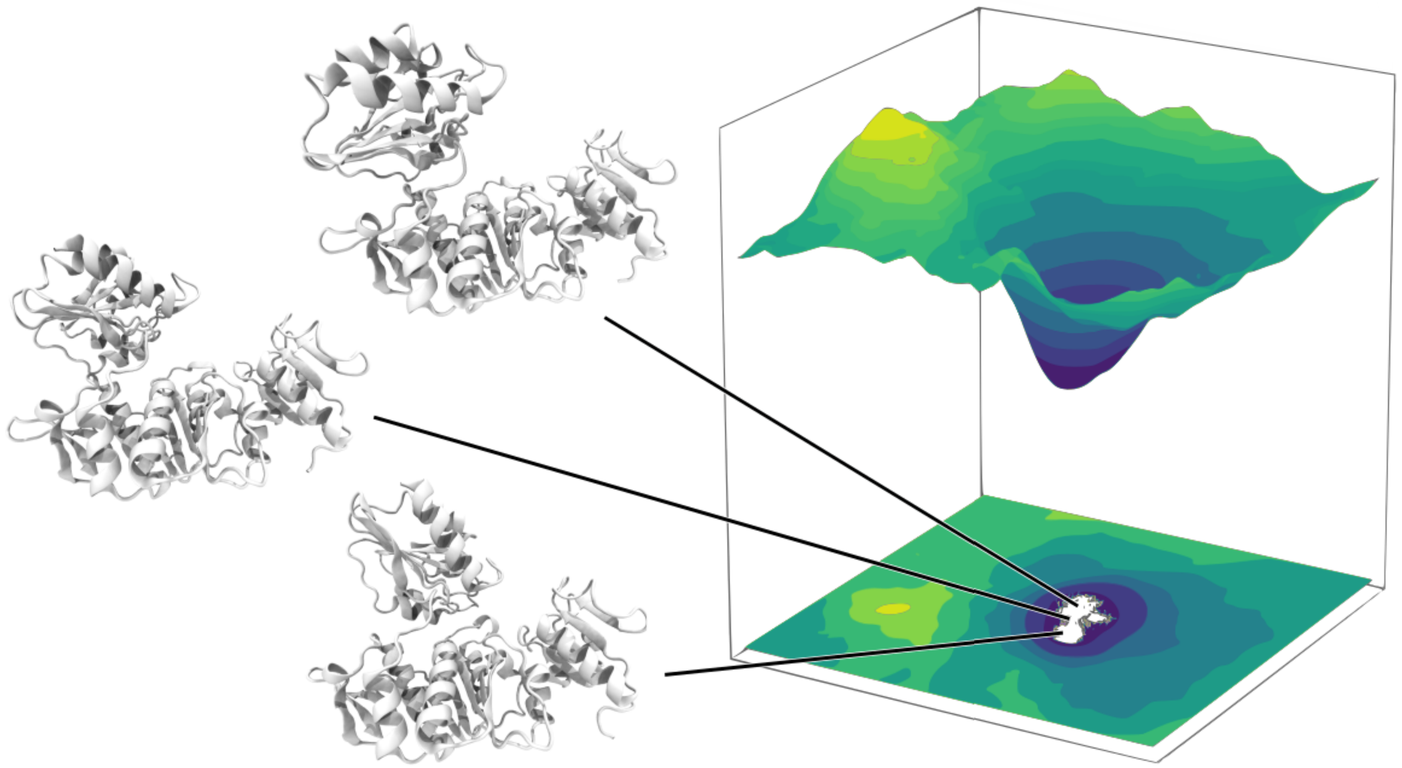
ICLR 2024 Generative & Experimental Perspectives for Biomolecular Design

[arXiv 2404.13491v1] GNNs diffusion models

Stanford University NVIDIA ICLR GEM @ ICLR 2024

“Molecular structures are fake news” —**Aviv Korman**, Dror Lab

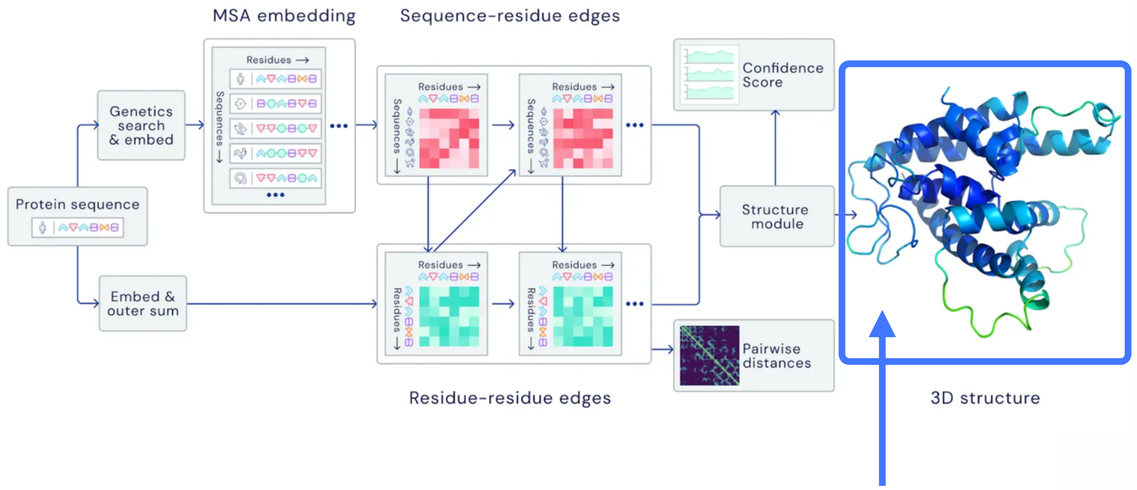
We should be thinking of structures as distributions.



We should predict structure *distributions*

But still out of reach today.

SOTA models are trained on PDB structures from crystallography or cryo-EM

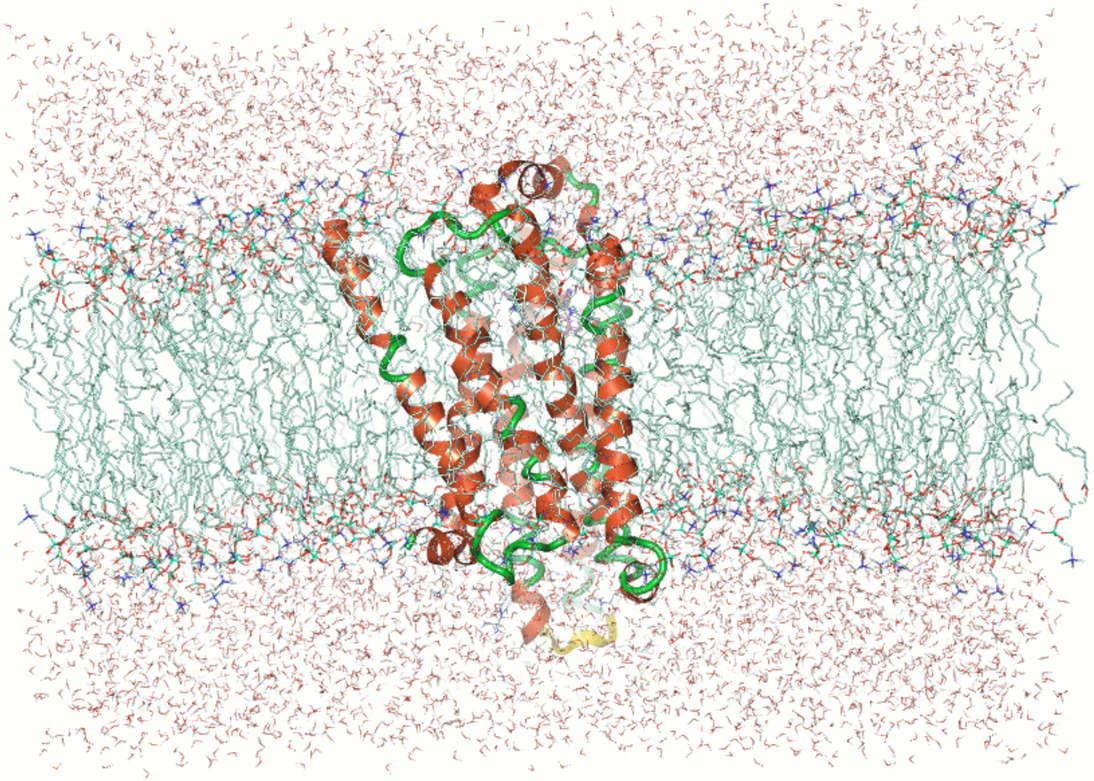


PDB structure ground truth

DeepMind

Jumper et al. (2022)

MD is slow, expensive, and may not sample all conformation space

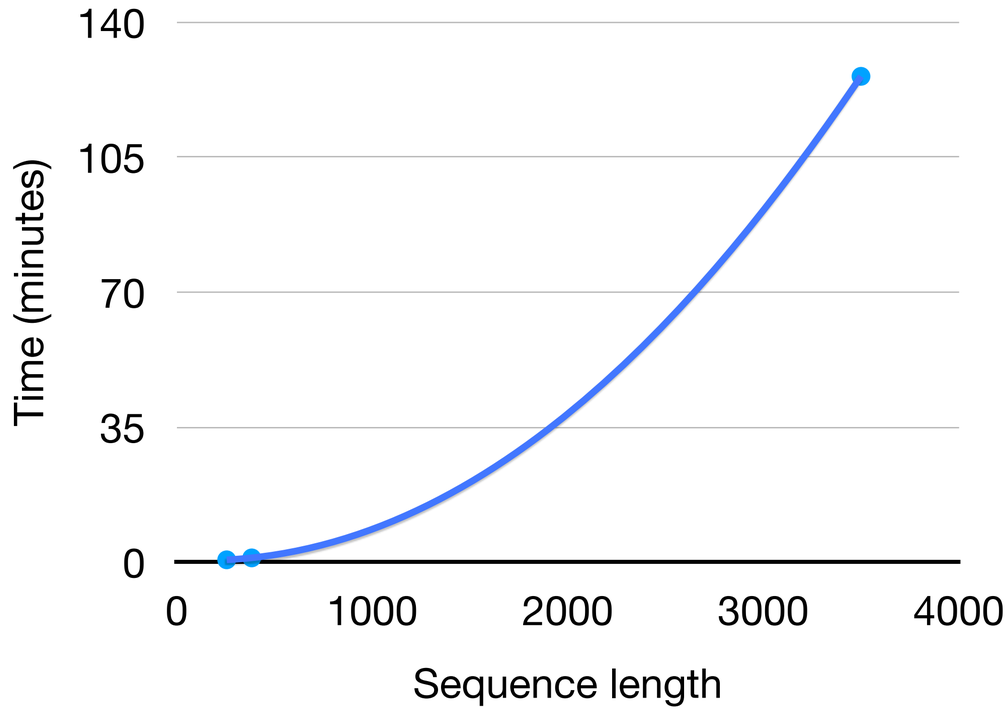


Can we accelerate generation today?

Necessary step towards MC distribution sampling.

AlphaFold 2, V100 GPU, single structure (no ensembling):

- 0.6 min for a 256-residue chain
- 1.1 min for a 384-residue chain
- 2.1 h for a 2,500-residue chain

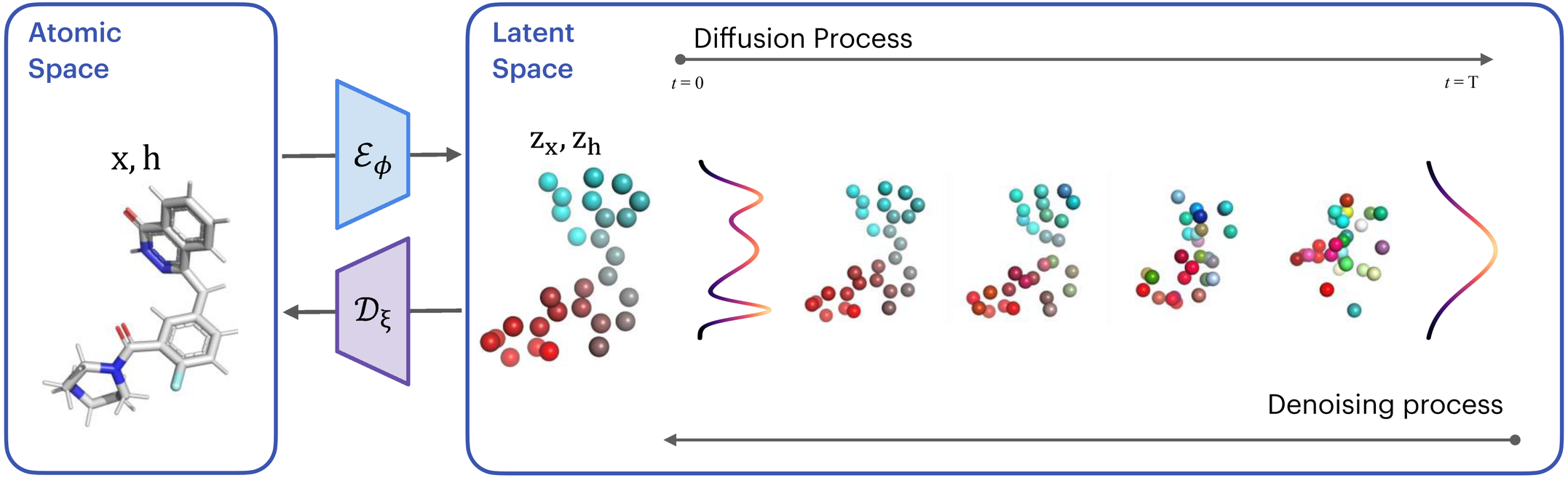


DeepMind

Jumper et al. (2022)

Setup: molecular generation with GeoLDM

Geometrically equivariant latent diffusion model.

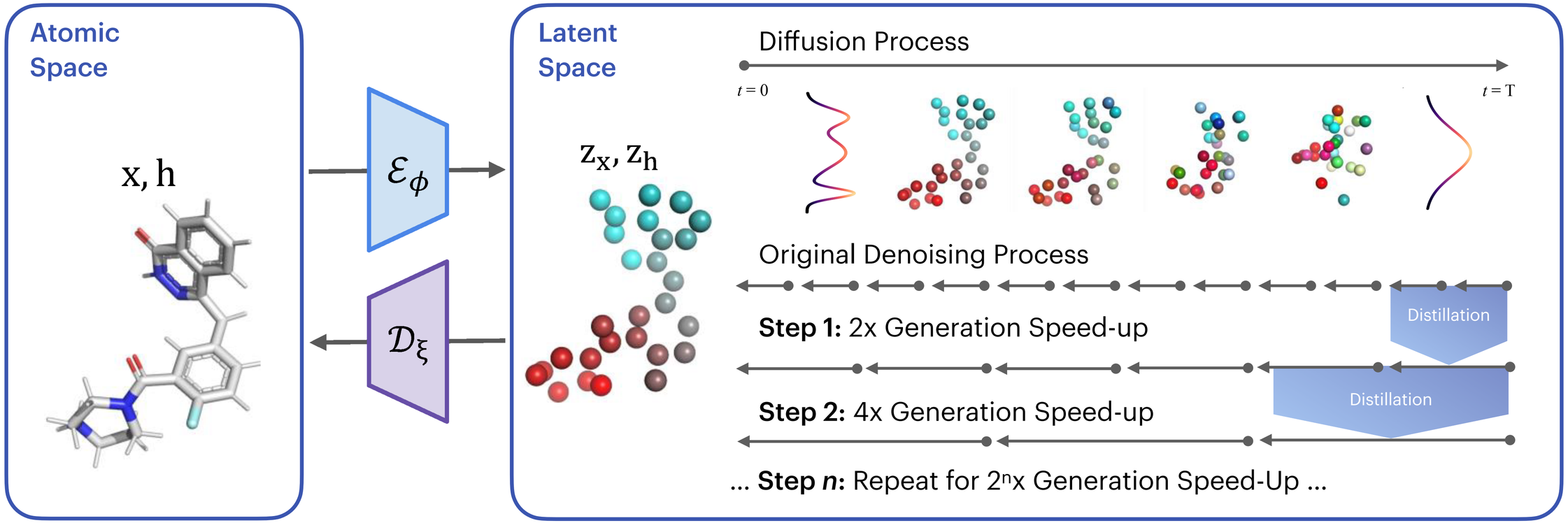


Xu et al. (2023): Geometric Latent Diffusion Models for 3D Molecule Generation

Idea: progressive distillations of the denoising process

Structure models rely on evolution and the PDB

AlphaFold 3

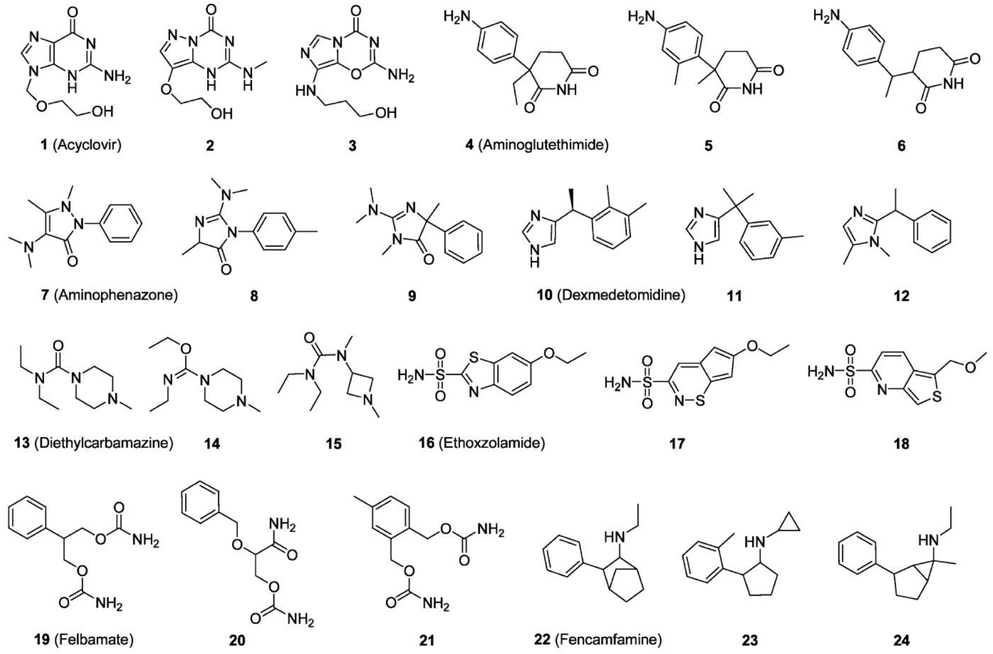


Experiments

Training set

QM9 dataset of 3D molecular structures

- **~134 K molecules:** Small organic compounds of up to 9 heavy atoms (C, N, O, F) with valence hydrogens
- **3D geometries:** equilibrium structures
- **Gold-standard benchmark:** Widely used for training and evaluating DFT and QC ML models



Ramakrishnan et al. (2014) Quantum chemistry structures and properties of 134 kilo molecules.

Experiment: baseline vs progressive distillation

Train over larger steps directly vs progressively

Steps size:

- **GeoLDM** (baseline): 1000 denoising steps per generation
- **Larger steps**: train on 100 steps per generation directly
- **Distill**: train to take 2x larger steps recursively

Experiment: DDIM vs DDPM

Stochastic vs implicit deterministic denoising solver

Solvers:

- **DDPM:** stochastic denoising steps

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z, \quad z \sim \mathcal{N}(0, I)$$

- **DDIM:** deterministic implicit denoising formula

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(x_t, t)$$

Ho et al. (2020): DDPM. Song et al. (2021): DDIM.

Results

Progressive distillation maintains quality

8× speed-up and only a 1 point drop in molecular stability.

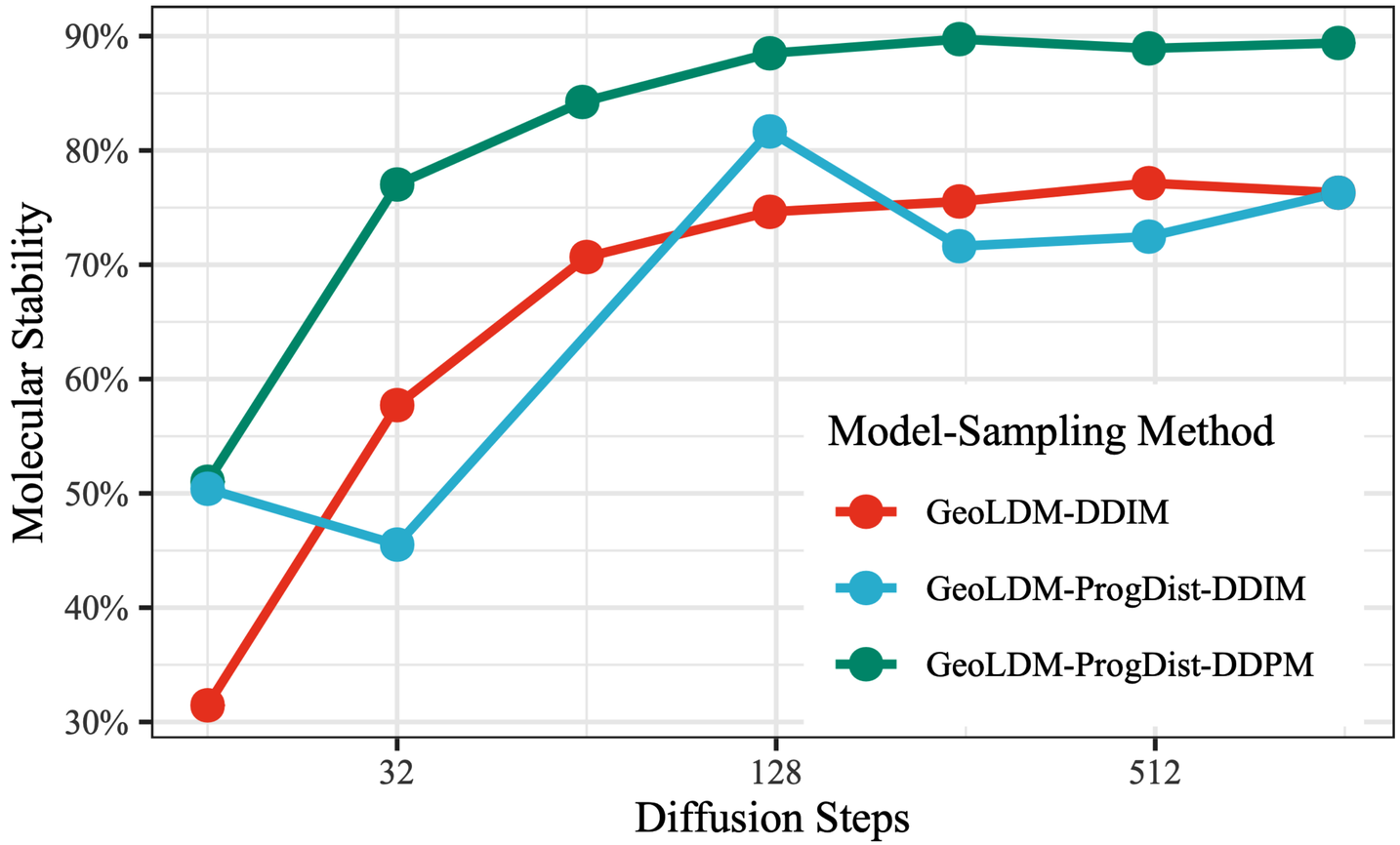
Baseline

Progressive distillation

Model	Sampling Method	Steps	Speed (sec^{-1})	Mol Sta %	Valid %	Valid & Unique %
GeoLDM	DDPM	1000	3.70	89.4	93.8	92.7
		100	33.30	55.8	70.6	79.7
GeoLDM	DDIM	1000	3.59	76.3	87	86.1
		125	28.30	74.6	85.3	84.1
		16	196.69	31.4	53.0	52.7
GeoLDM-PD	DDPM	125	28.28	88.4	93.3	91.6
		16	196.51	51.0	73.2	72.3
GeoLDM-PD	DDIM	125	28.28	81.6	91.7	83.6
		16	196.51	50.4	73.4	72.6

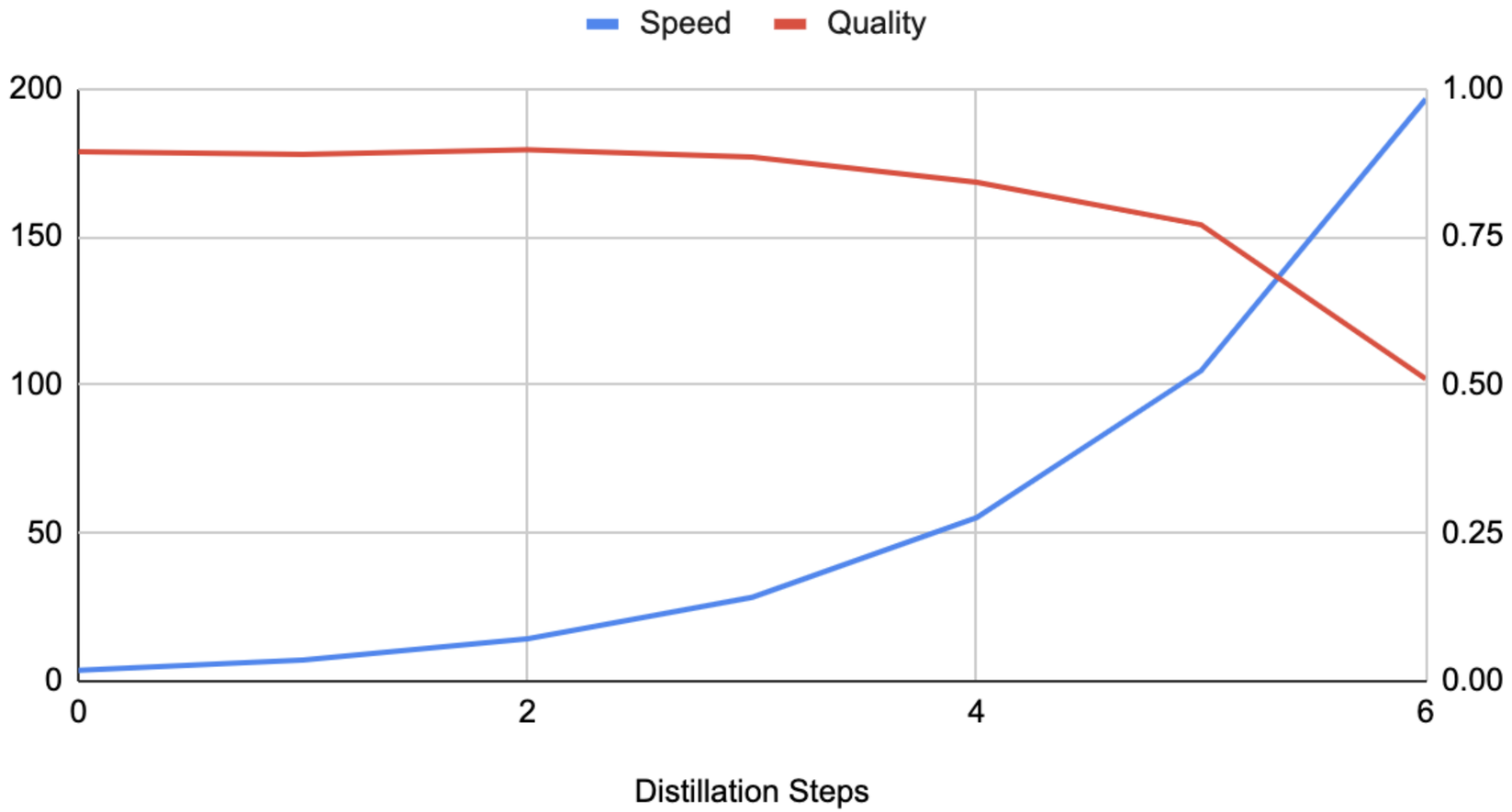
Quality maintained after 3-4 distillation steps

DDPM-PD @ 125 steps speeds up 8x with equivalent stability.



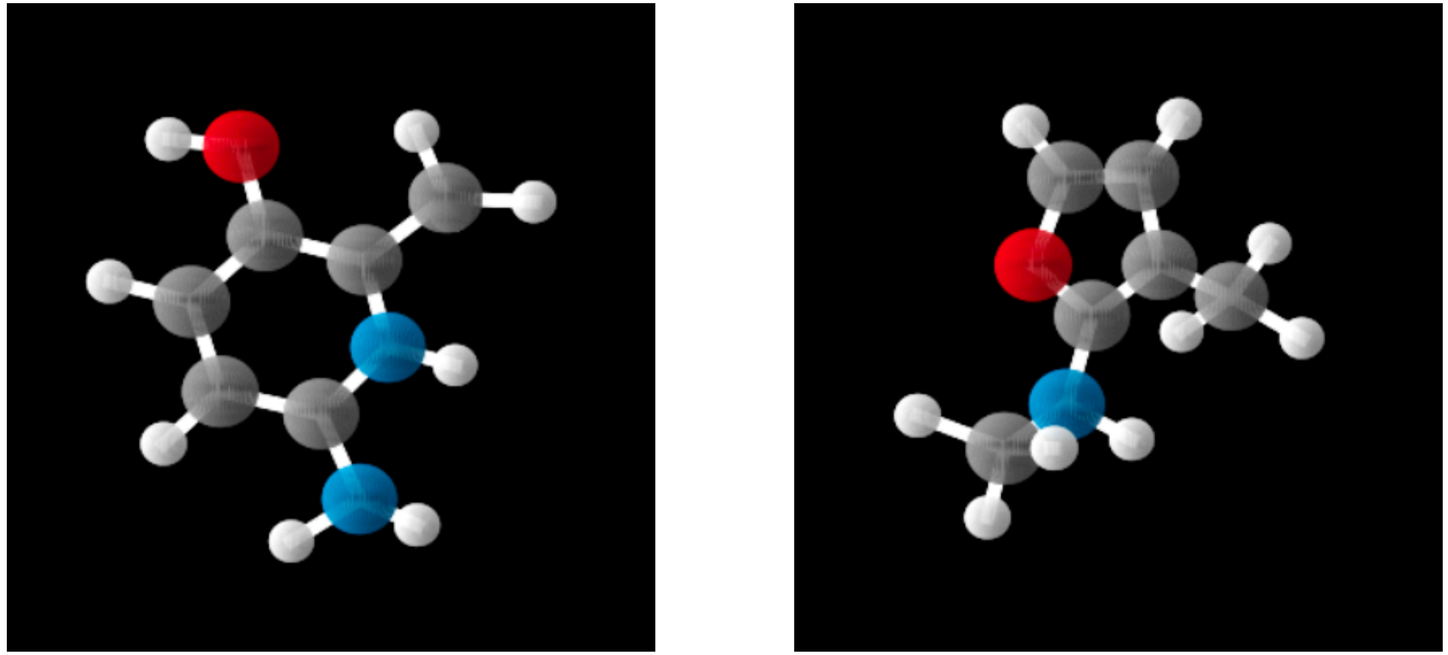
Significant speed gains before quality loss

But quality drops off after 4 steps of progressive distillation.



Examples of generated molecules

Stable conformations aligned with QM9 distribution

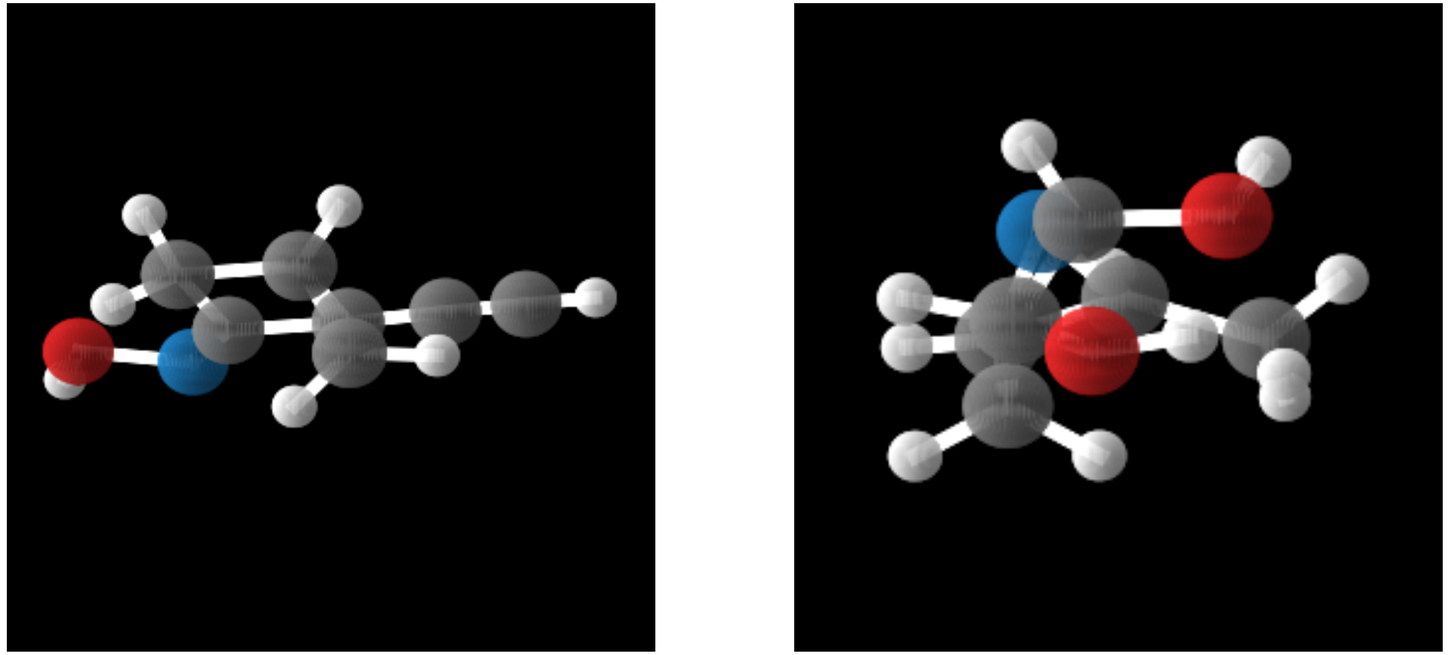


Sample stable conformations

3 distillations | 125 steps | 8x speed-up

Examples of generated molecules

Unstable conformations outside the QM9 distribution



Sample unstable conformations

6 distillations | 16 steps | 64x speed-up

Take aways

Conclusions and future work

- **Progressive distillation leads to significant gains in generation speed while maintaining conformation quality.**
  - **8x speed-up gains with comparable quality for DDPM-PD**
- Future work: What applications does this speed-up open?
  - Scale up high-throughput screening
  - Large or multi-domain proteins: progressive distillation of AlphaFold3/Boltz-1?
  - Other ways to speed up inference? e.g. consistency models (one-step generation)

Map, Model, Measure: AI for Biomolecules >>

Questions?

Accelerating the Generation of Molecular Conformations with Progressive Distillation of Equivariant Latent Diffusion Models

ICLR 2024 Generative and Experimental Perspectives for Biomolecular Design

[arXiv 2404.13491v1] GNNs diffusion models

Stanford University NVIDIA ICLR GEM @ ICLR 2024

Map, Model, Measure: AI for Biomolecules >>

Non-Canonical Crosslinks Confound Evolutionary Protein Structure Models

Romain Lacombe

Experimental Design in AI for Science workshop, 2025

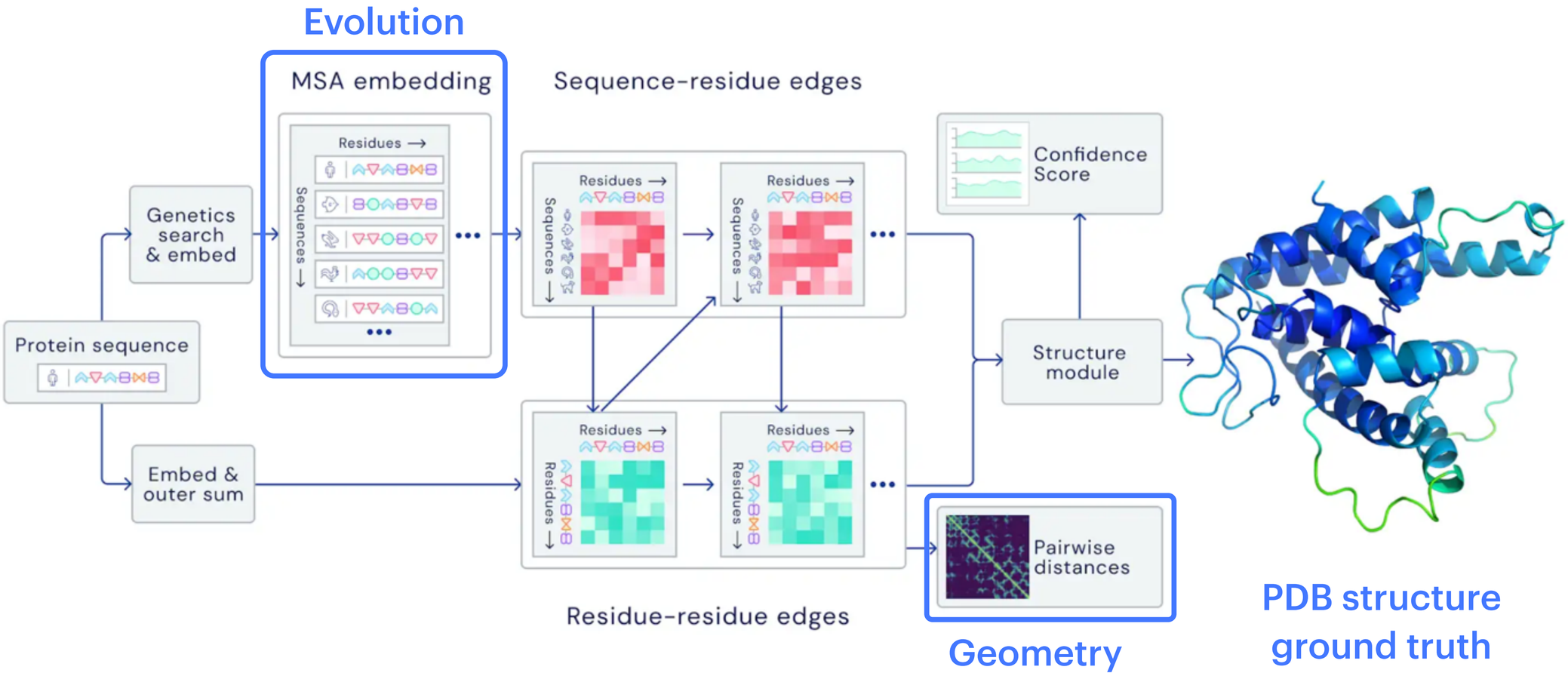
[bioRxiv: 2025.03.17.643596v1] protein models evaluation

Stanford University CIFAR

Structure predictors rely on evolutionary and structural data.

Structure models rely on evolution and the PDB

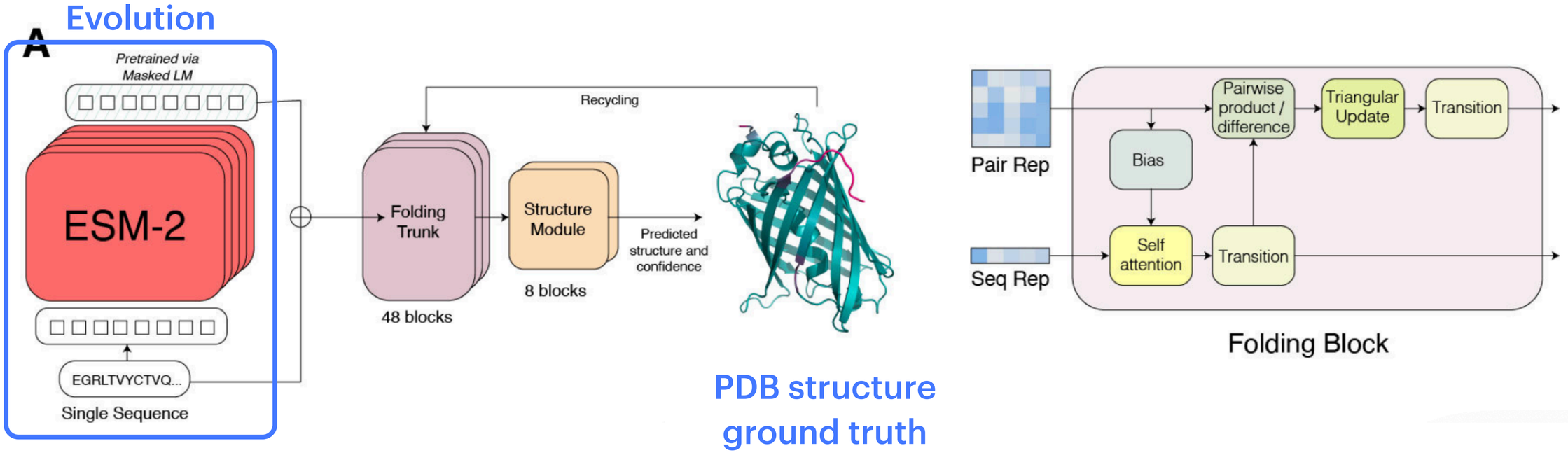
AlphaFold 3



DeepMind

Structure models rely on evolution and the PDB

ESMFold

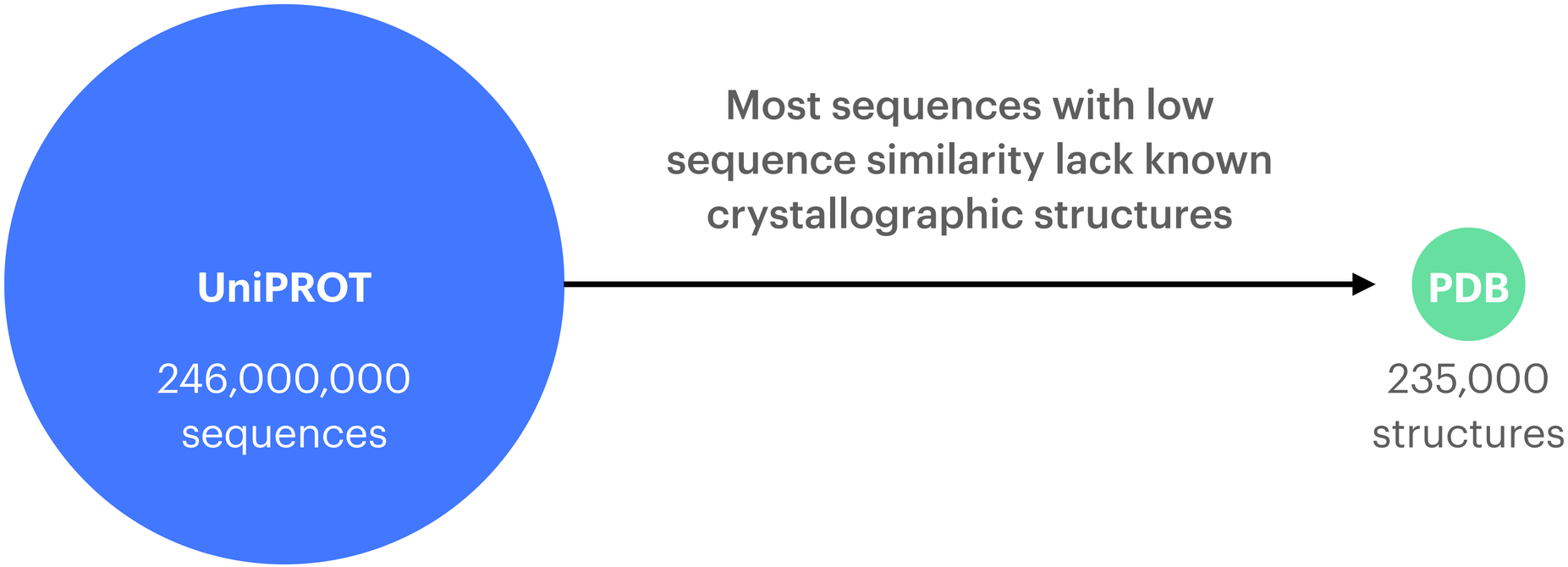


Meta AI

How do structure predictors perform out-of-domain?

How do they perform out-of-domain?

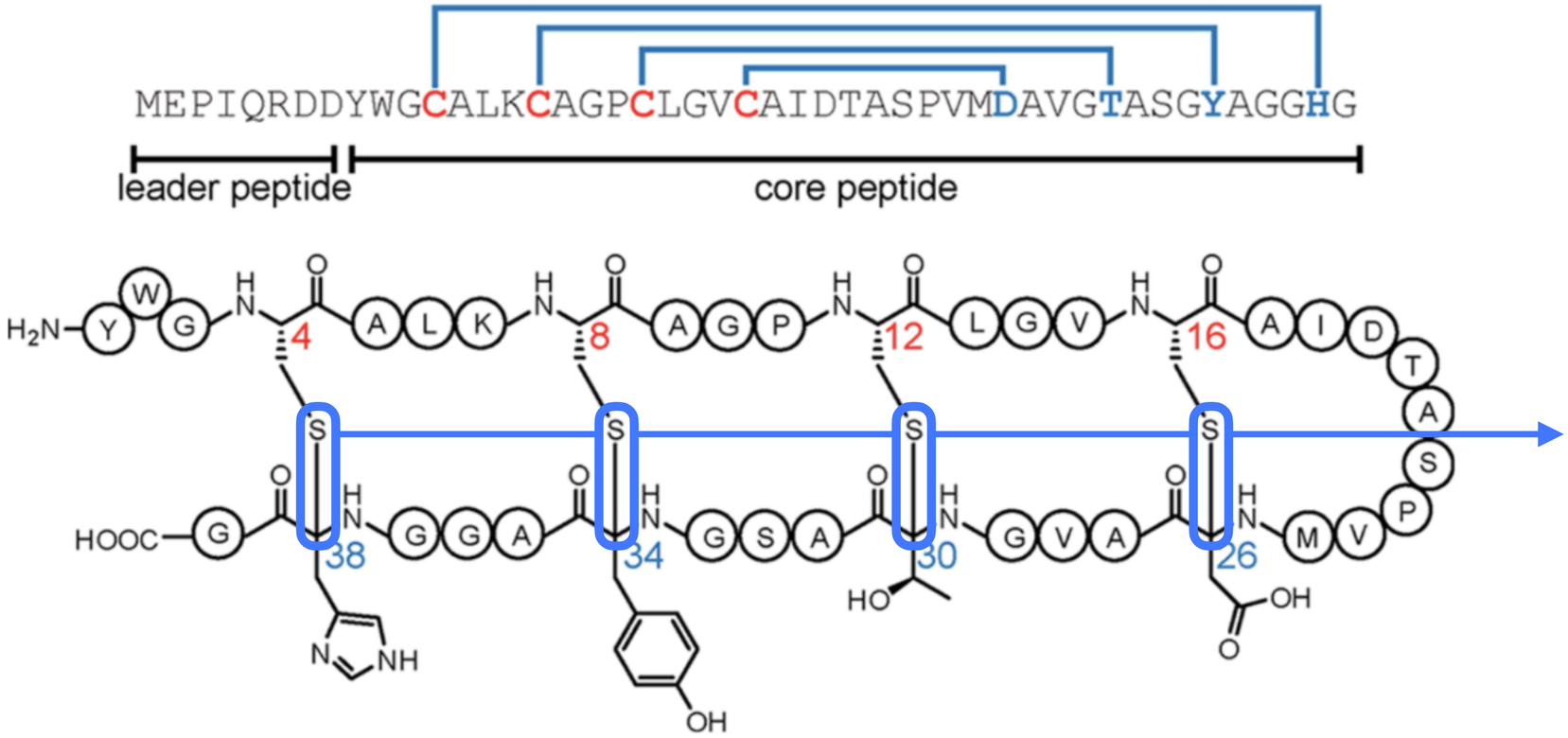
Problem: lack of structural ground truth.



Idea: crosslink geometries as ground truth for predictions!

Enter sactipeptides

Rare class of proteins with sulfur-to-α-carbon crosslinks



Strong geometric constraint:
S-to-α-carbon bonds have a
known length of 1.82 Å

Natural experiment

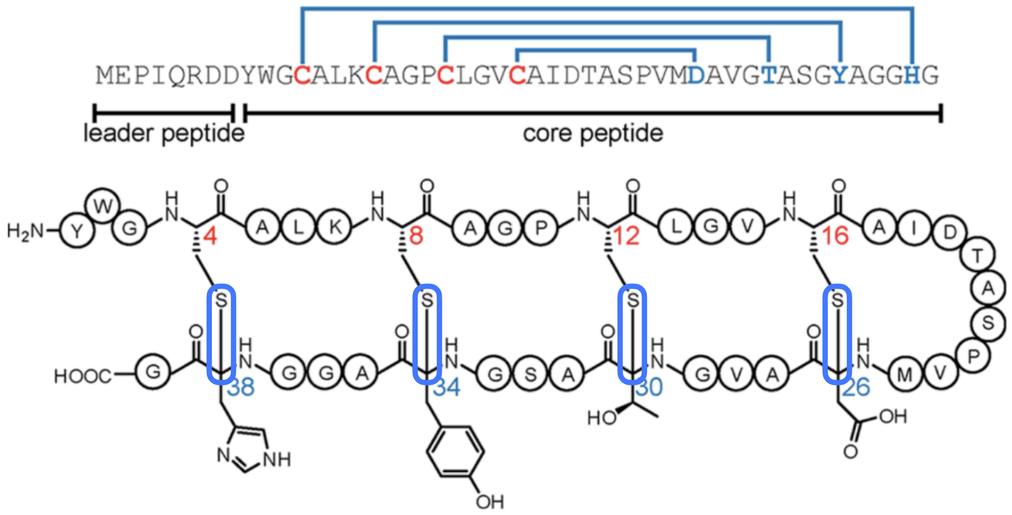
10 peptides with known crosslinks, only 5 in the PDB

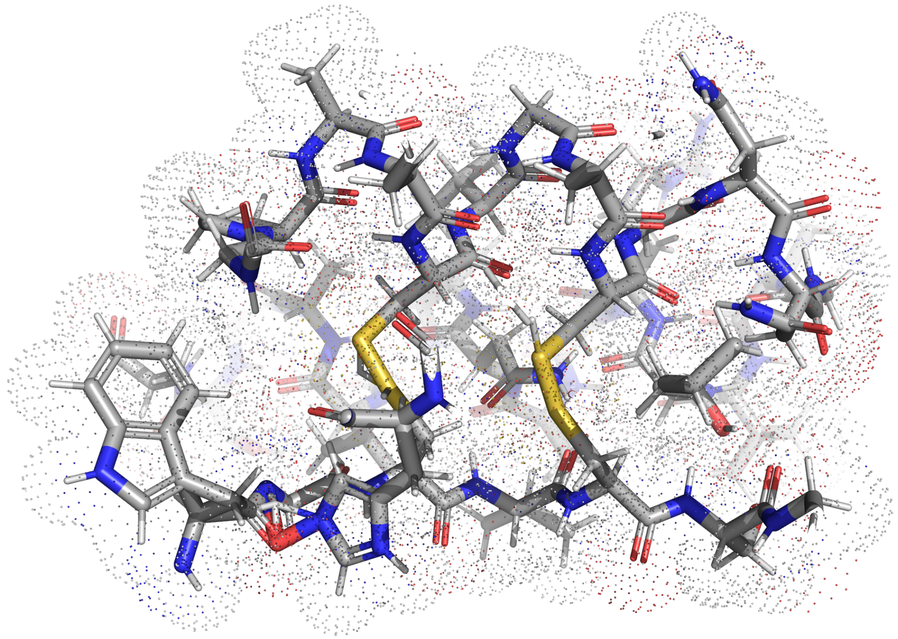
	Sactipeptide	Length	PDB ID	Cross-links
Known structure	Ruminococcin C1	44	6T33	C3→N16, C5→A12, C22→K42, C26→R34
	Subtilosin A	35	1PXQ	C4→F31, C7→T28, C13→F22
	Thurincin H	31	2LBZ	C4→S28, C7→T25, C10→T22, C13→N19
	Thuricin CD α	30	2L9X	C5→T28, C9→T25, C13→S21
	Thuricin CD β	30	2LA0	C5→Y28, C9→A25, C13→T21
Unknown structure	Huazacin	40	—	C4→H38, C8→Y34, C12→T30, C16→D26
	Hyicin 4244	35	—	C4→F31, C7→T28, C13→F22
	Skf A	26	—	C4→M12
	Streptosactin	14	—	C4→S7, C10→G13
	QmpA	13	—	C6→D4, C10→D8

Geometric ground truth

Benchmarking structure predictors

How closely do predicted structures fit crosslink geometry?





Ground truth: S-to-α-carbon bonds of **known length: 1.82 Å**

Benchmark: predicted distance of **each S atom to bound α-carbon**

Evaluation metrics

Adapting structure prediction metrics to crosslinks geometry.

Global Distance Test – Total Score:

$$\text{GDT-TS} = \frac{1}{4} \sum_{D \in \{1,2,4,8\}} \%\{\text{sactibond} \mid d(\text{S}, \text{C}_\alpha^{\text{target}}) \leq D + 1.8 \text{ Å}\}$$

Root Mean Square Distance:

$$\text{RMSD} = \sqrt{\sum_{\text{S, C} \in \text{sactibonds}} ||d(\text{S}, \text{C}_\alpha^{\text{target}}) - 1.8 \text{ Å}||^2}$$

Results: benchmarking 6 SOTA structure prediction models

Results

Structure predictors generalize poorly beyond evolutionary priors.

Benchmarking 6 SOTA models:

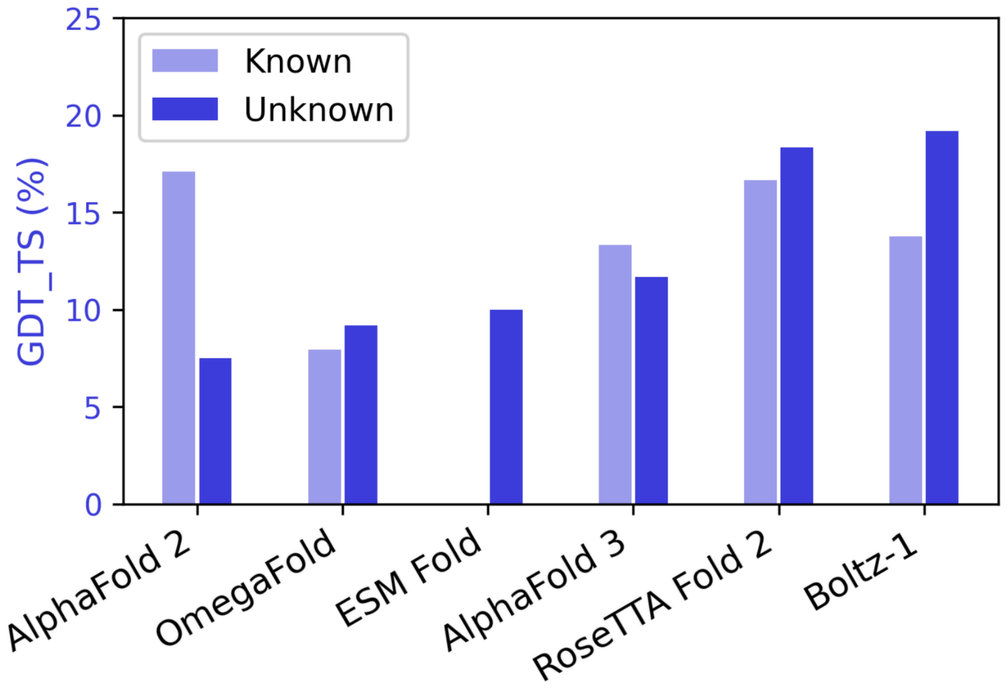
Model	Authors	Ensembling	GDT-TS		RMSD	
			Known	Unknown	Known	Unknown
AlphaFold 2	DeepMind	Yes	17.1 %	7.5 %	10.7 Å	12.7 Å
AlphaFold 3	DeepMind	Yes	13.3 %	11.7 %	16.9 Å	12.0 Å
Boltz-1	MIT	No	13.7 %	**19.2 %**	9.7 Å	**6.9 Å**
ESMFold	Meta	No	0.0 %	10.0 %	17.8 Å	12.3 Å
OmegaFold	Tencent AI	No	7.9 %	9.2 %	9.2 Å	9.1 Å
RoseTTAFold 2	Baker Lab	Yes	16.7 %	18.3 %	8.1 Å	7.9 Å
Average	–	–	11.45 %	12.65 %	12.1 Å	10.1 Å

(+) Higher == better (–) Lower == better

Results

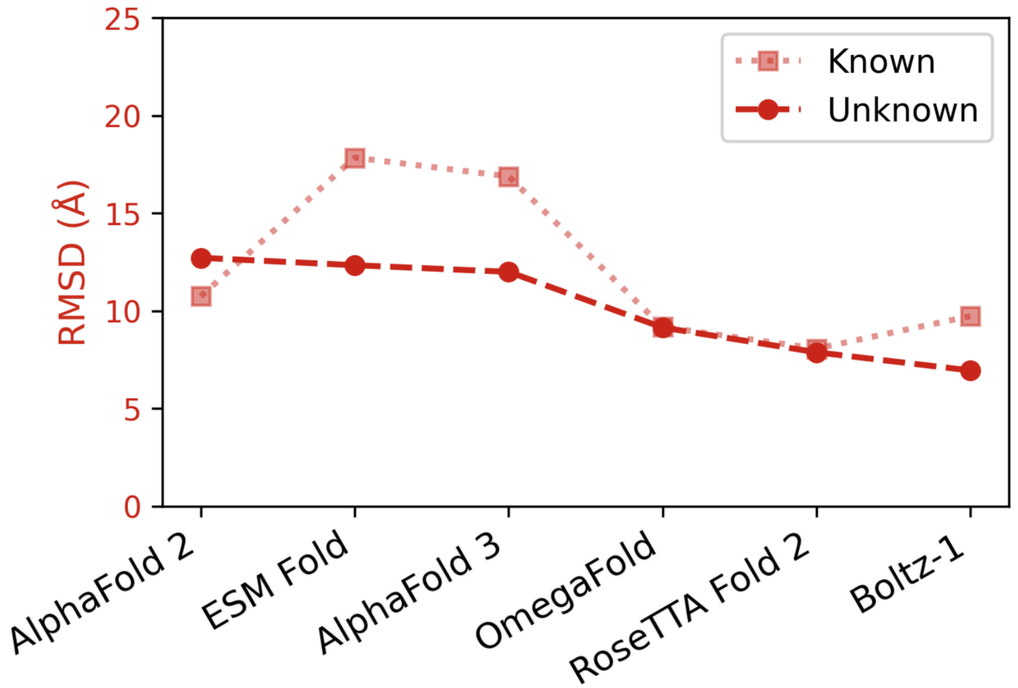
Structure predictors generalize poorly beyond evolutionary priors.

Global Distance Test – Total Score:



(+) Higher == better

Root Mean Square Distance:



(–) Lower == better

Take aways

Conclusions and future work

- **Structure predictors generalize poorly beyond evolutionary priors.**
- This limits their usefulness for mutational screening, *de novo* design, etc.
- Can we train on more low evolutionary depth examples?
  - More non-canonical crosslinks? Macrocycles? Limited by PDB data
- Can we inject physics into protein structure predictors?
  - Free energy in loss?
  - Blend MD and denoising steps?

Map, Model, Measure: AI for Biomolecules >>

Questions?

Non-Canonical Crosslinks Confound Evolutionary Protein Structure Models

Experimental Design in AI for Science, 2025

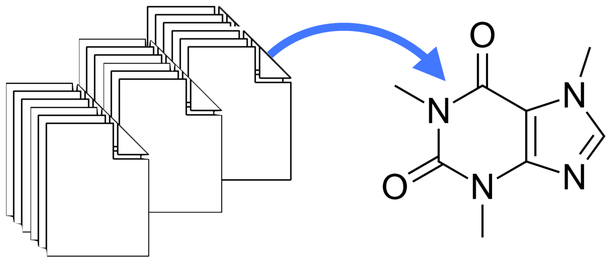
[bioRxiv: 2025.03.17.643596v1] protein models evaluation

Stanford University CIFAR

AI for Biomolecules: 3 papers

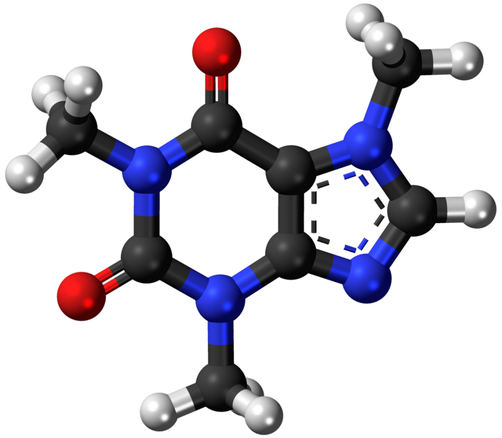
Map:

Can we extract properties of molecules from science papers?



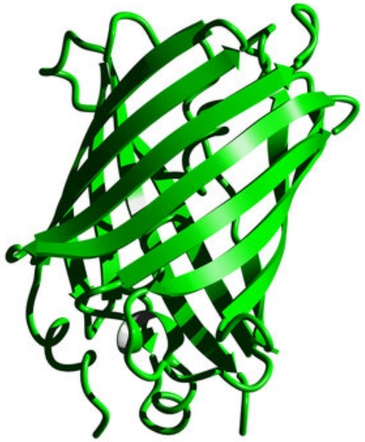
Model:

Can we accelerate sampling of molecule conformations?



Measure:

How good are protein models out of their evolutionary domain?



Thank you!

Co-authors:

Romain Lacombe
Stanford ChemE

David Lüdeke
Stanford CS

Andrew Gaut
Stanford CS

Kateryna Pistunova
Stanford Physics

Jeff He
Stanford CS

Neil Vaidya
NVIDIA

Chris Manning
Stanford CS

Stefano Ermon
Stanford CS

Will Van Treuren
Interface Biosciences

Support:

Stanford Computer Science

HKUST

NVIDIA

ICML International Conference On Machine Learning

ACS Chemistry for Life®

ICLR

GEM © ICLR 2024

CIFAR

GeoLDM
Minkai Xu, Alexander Powers, Ron Dror, Stefano Ermon, Jure Leskovec

MoMu Anyi Rao et al.

NVIDIA GPU Cluster (GeoLDM distillation)

ICML 2023 Computational Biology Workshop

ACS Fall 2023 AI in Organic Chemistry Workshop

ICLR 2024 Generative and experimental approaches to biomolecular design workshop

Experimental Design: AI for Science Workshop 2025

Interface Bio (sactipeptides expertise)

Thank you!

Romain Lacombe <rlacombe@stanford.edu>
Stanford Chemical Engineering