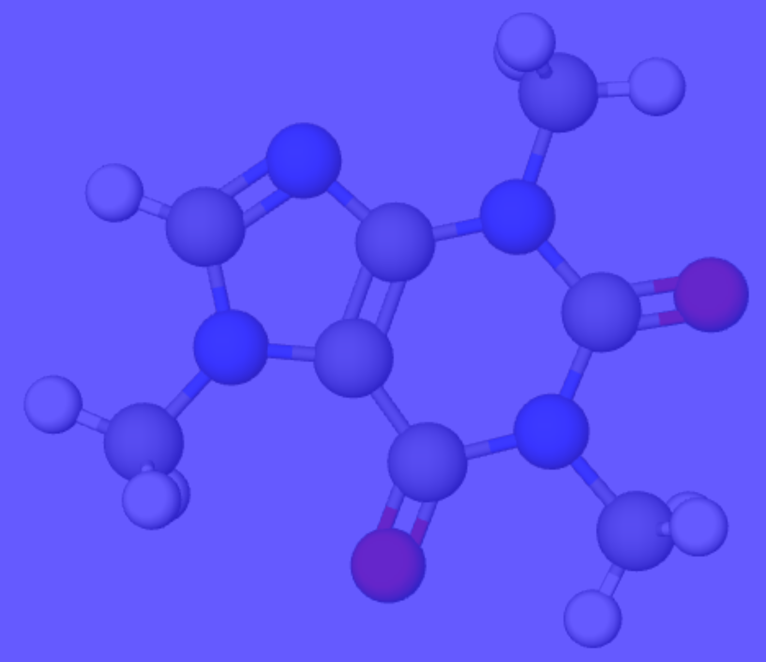
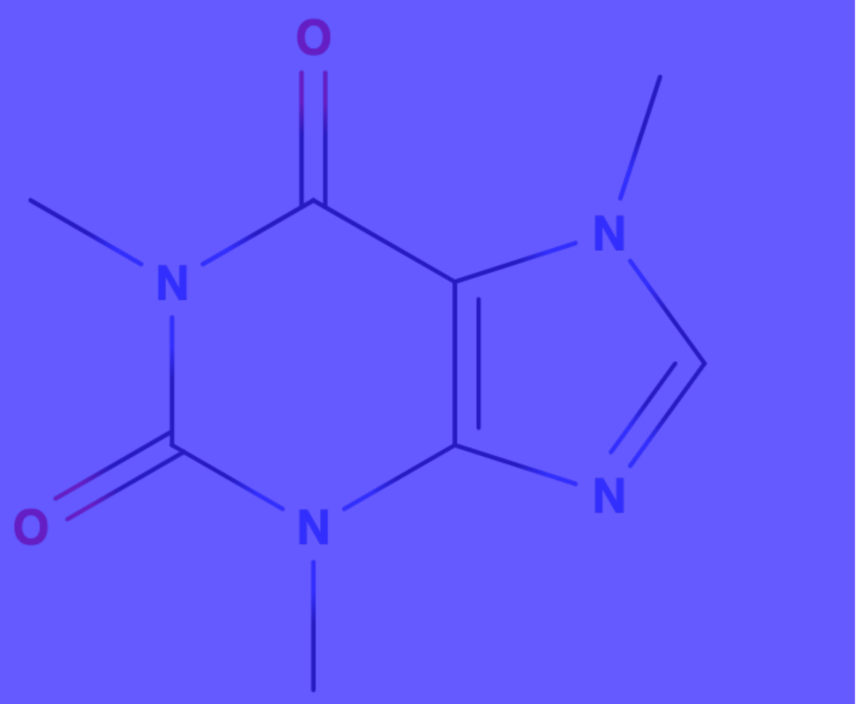
Molecule Representation Learning:
A Perspective from Topology, Geometry, and Textual Description

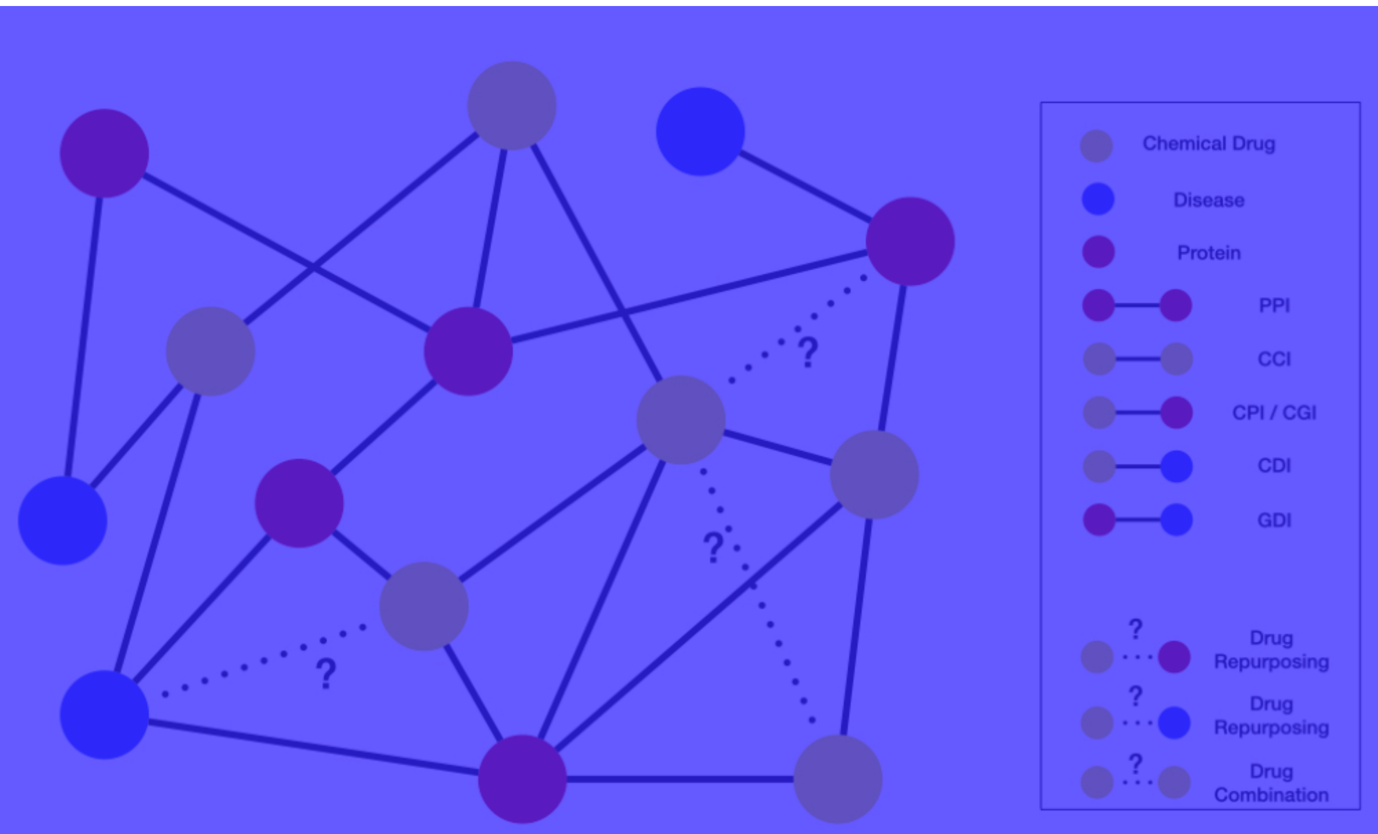
Shengchao Liu, Mila-UdeM



3D Molecular Graph
3D GNN



2D Molecular Graph
2D GNN



Biological Knowledge Graph (KG)
GNN

Internal Structure

External Knowledge

Molecule Data Structure & Representation

String (SMILES, SELFIES)
CNN, RNN, LM

OC(=O)C1=CC=CC=C1O

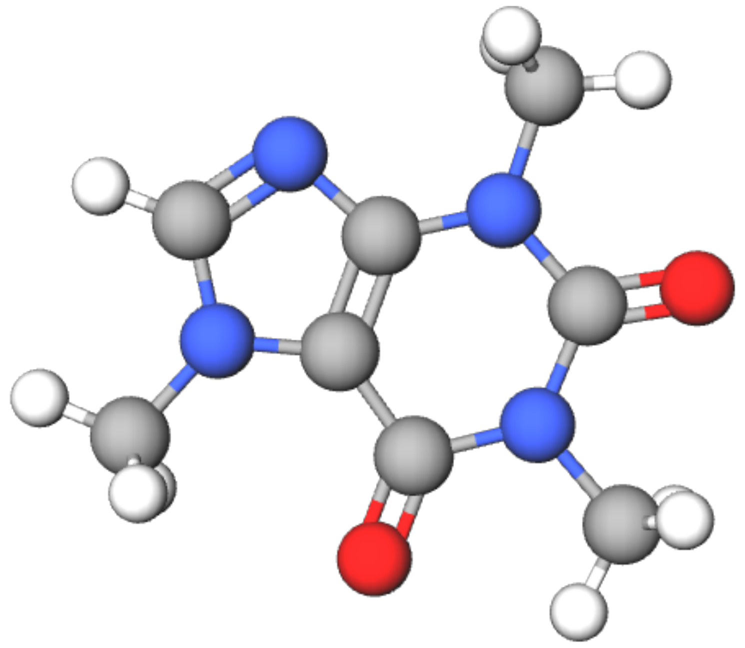
Fingerprint
RF, XGB, MLP

0001100….00100…1100

Textual Description
RNN, LM

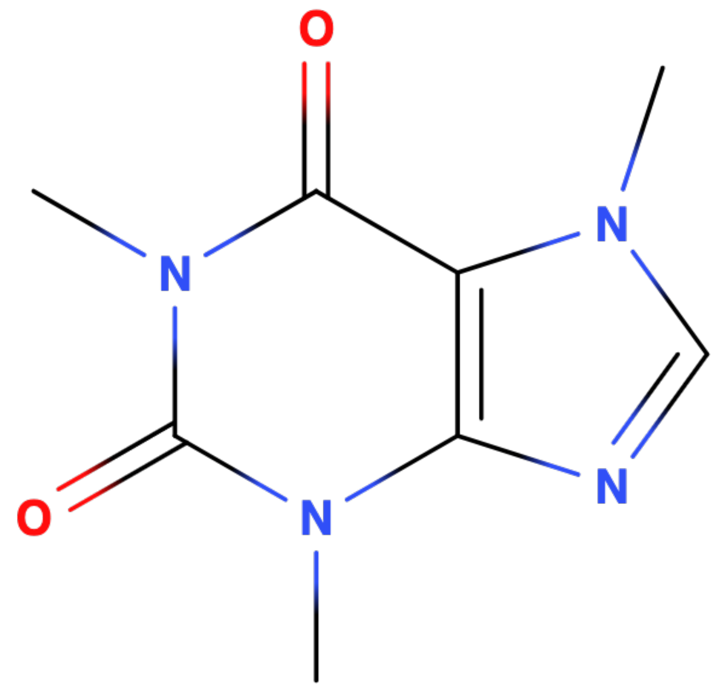
Biomedical Text

Salicylic acid is a phenolic compound present in the plants, where it plays a central role in the development of resistance to pathogen infection.



3D Molecular Graph

3D GNN



2D Molecular Graph

2D GNN

(1) GraphMVP (2D-3D SSL)
ICLR’22

(2) GeoSSL (3D SSL)
ICLR’23

Biomedical Text

Salicylic acid is a phenolic compound present in the plants, where it plays a central role in the development of resistance to pathogen infection.

Textual Description

RNN, LM

(3) MoleculeSTM (Foundation Model)
In submission

String (SMILES, SELFIES)

CNN, RNN, LM

OC(=O)C1=CC=CC=C1O

GraphMVP: Pre-training Molecular Graph Representation with 3D Geometry

ICLR 2022

Shengchao Liu, *Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, Jian Tang*

Motivation & Problem Definition

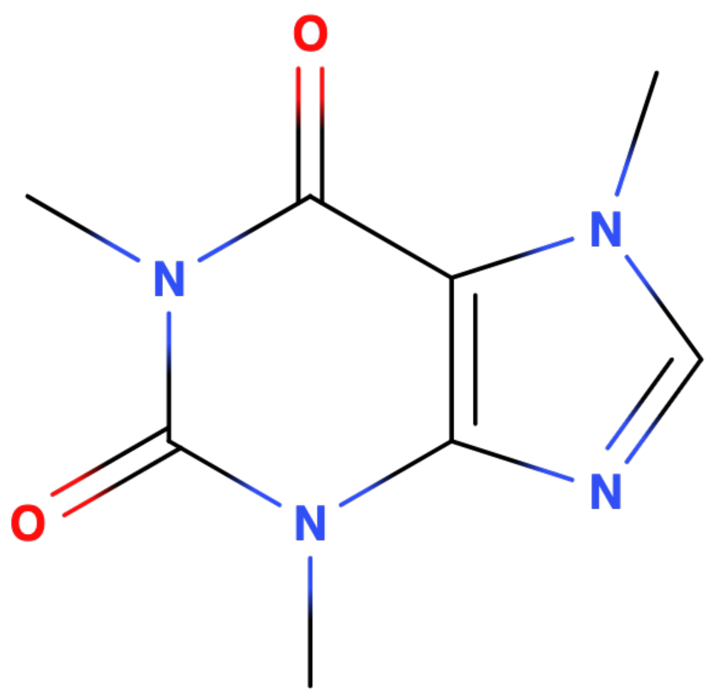
Ultimate goal:

- Molecular property prediction on target (downstream) tasks.
- MoleculeNet [1]: only 2D topology for molecular graph is available.

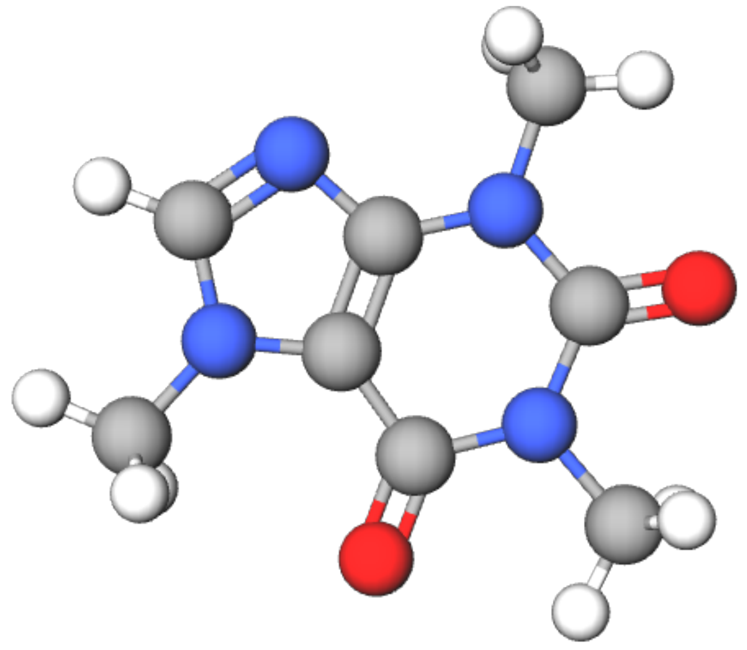
[1] Wu, Zhenqin, et al. "MoleculeNet: a benchmark for molecular machine learning." *Chemical science* 9.2 (2018): 513-530.

Community has put more efforts in gathering large-scale 3D geometry datasets.

- GEOM (250K, June 2020)



2D Molecular Graph

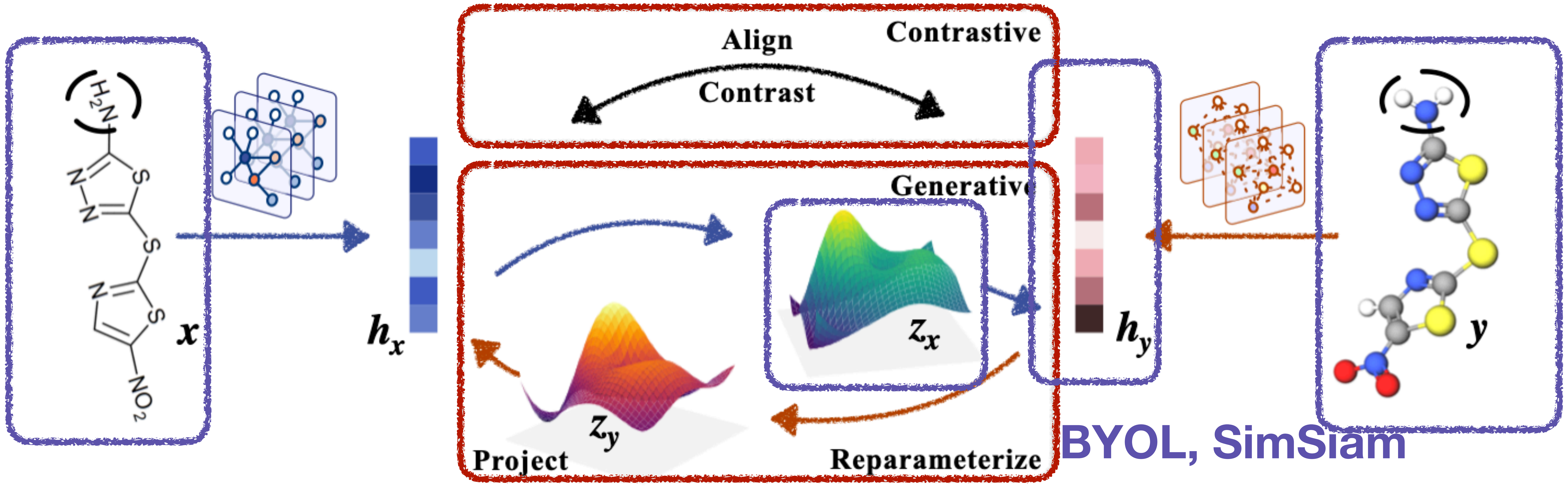


3D Molecular Graph

GraphMVP

$$I(X;Y) \Longrightarrow \boxed{\frac{1}{2}\mathbb{E}_{p(x,y)}[\log p(x|y) + \log p(y|x)]}$$

Energy-Based Model, Noise Contrastive Estimation (EBM-NCE)



BYOL, SimSiam

Variational Representation Reconstruction (VRR)

Follow-up:

Doing reconstruction on the data space is better than that of representation space.

More Datasets

Community has put more efforts in gathering large-scale 3D geometry datasets.

- GEOM (250K, June 2020)
- No 3D downstream tasks in GraphMVP
  - QM9 has ~130K data points.
  - MD17 has 50K-1M conformers.

- **[after submission of GraphMVP]**
  - Molecule3D (3.8M, Aug 2021)
  - PCQM4Mv2 (3.4M, May 2022)

GeoSSL: Molecular Geometry Pretraining with SE(3)-Invariant Denoising Distance Matching

ICLR 2023

Shengchao Liu, *Hongyu Guo, Jian Tang*

Problem Definition

Pure 3D geometric representation exploration.

- Pretraining: a large molecule 3D dataset (1M from Molecule3D [1]).
- Downstream tasks:
  - QM9: quantum mechanics prediction.
  - MD17: force prediction.
  - LBA & LEP: ligand-pocket binding prediction.

[1] Xu, Zhao, et al. "Molecule3D: A Benchmark for Predicting 3D Geometries from Molecular Graphs." *arXiv preprint arXiv:2110.01717* (2021).

Coordinate Perturbation

Coordinate perturbation is **important**!

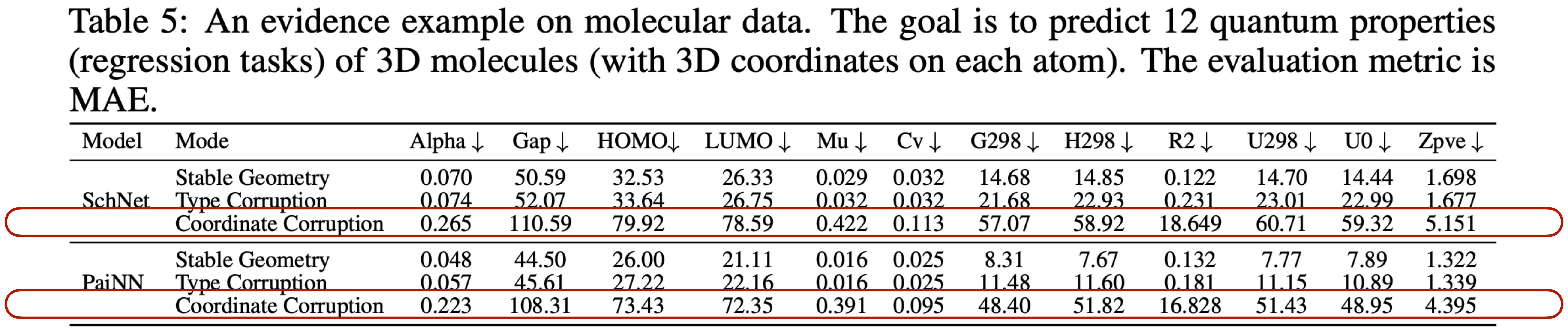
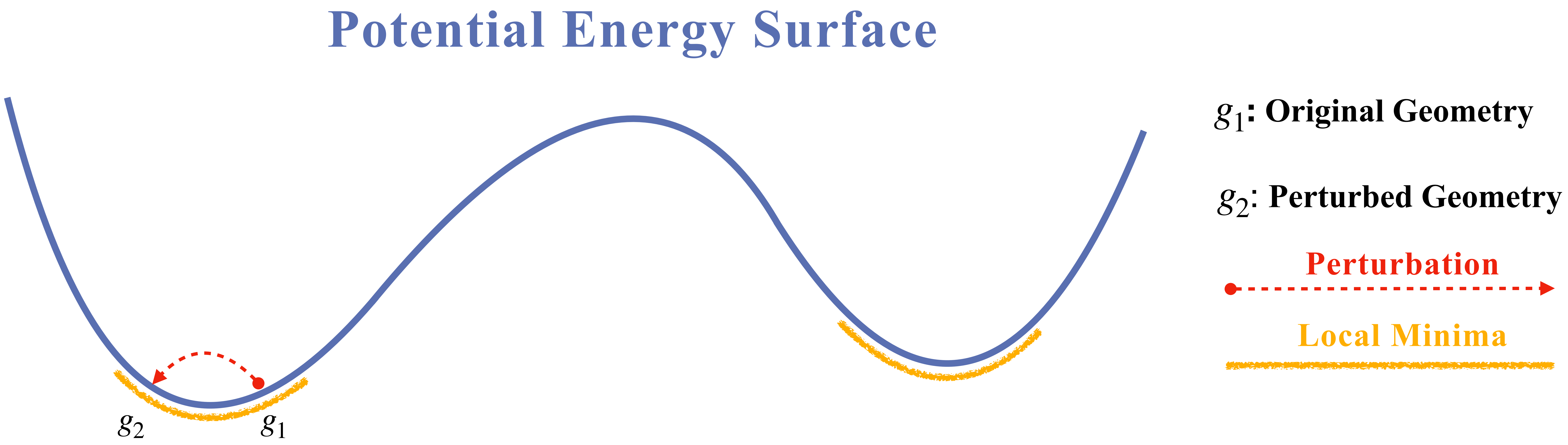


Table 5: An evidence example on molecular data. The goal is to predict 12 quantum properties (regression tasks) of 3D molecules (with 3D coordinates on each atom). The evaluation metric is MAE.

Model	Mode	Alpha ↓	Gap ↓	HOMO↓	LUMO ↓	Mu ↓	Cv ↓	G298 ↓	H298 ↓	R2 ↓	U298 ↓	U0 ↓	Zpve ↓
SchNet	Stable Geometry	0.070	50.59	32.53	26.33	0.029	0.032	14.68	14.85	0.122	14.70	14.44	1.698
	Type Corruption	0.074	52.07	33.64	26.75	0.032	0.032	21.68	22.93	0.231	23.01	22.99	1.677
	Coordinate Corruption	0.265	110.59	79.92	78.59	0.422	0.113	57.07	58.92	18.649	60.71	59.32	5.151
PaiNN	Stable Geometry	0.048	44.50	26.00	21.11	0.016	0.025	8.31	7.67	0.132	7.77	7.89	1.322
	Type Corruption	0.057	45.61	27.22	22.16	0.016	0.025	11.48	11.60	0.181	11.15	10.89	1.339
	Coordinate Corruption	0.223	108.31	73.43	72.35	0.391	0.095	48.40	51.82	16.828	51.43	48.95	4.395



Potential Energy Surface

g_1: **Original Geometry**

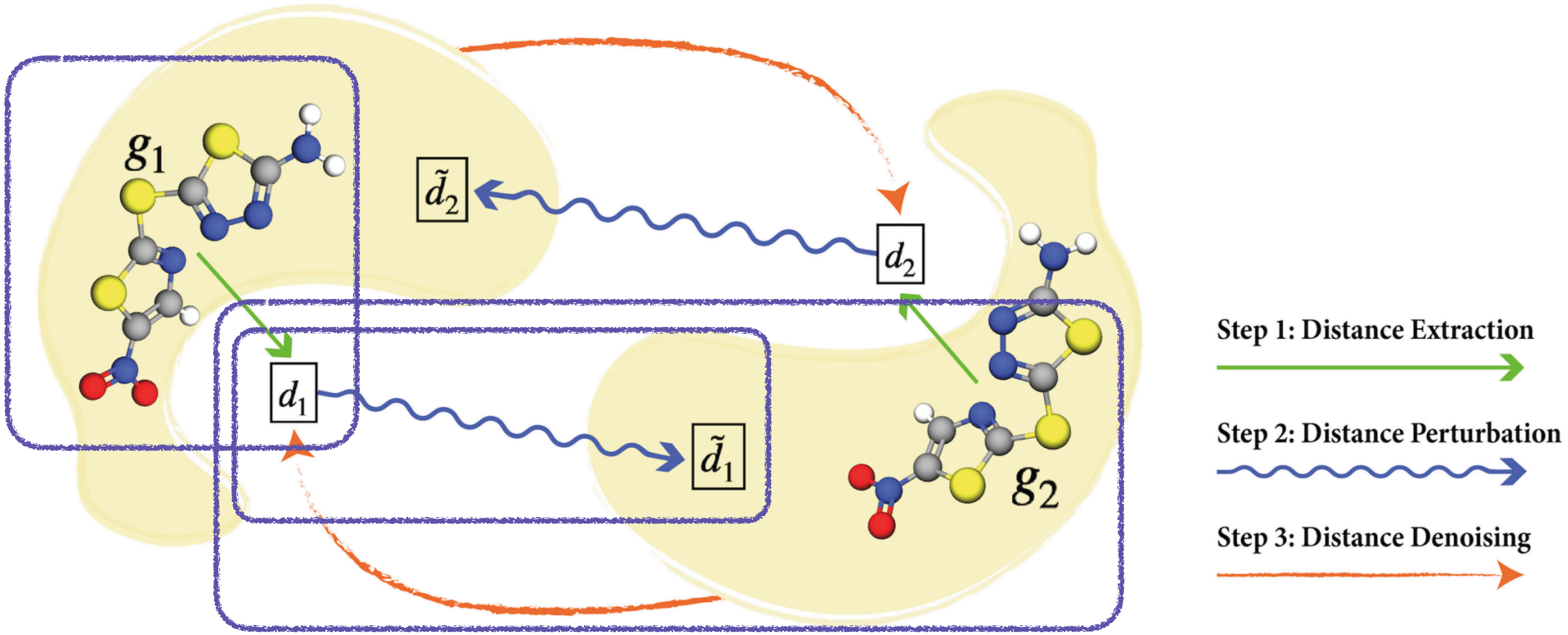
g_2: **Perturbed Geometry**

Perturbation

Local Minima

GeoSSL: SE(3)-Invariant Denoising Pretraining

$$\mathscr{L}_{GeoSSL} = \frac{1}{2}\mathbb{E}_{p(g_1,g_2)}\Big[\log p(g_1 \,|\, g_2)\Big] + \frac{1}{2}\mathbb{E}_{p(g_1,g_2)}\Big[\log p(g_2 \,|\, g_1)\Big]$$



g_1
$\tilde{d}_2$
d_2
d_1
$\tilde{d}_1$
g_2
Step 1: Distance Extraction
Step 2: Distance Perturbation
Step 3: Distance Denoising

MoleculeSTM: Multi-modal Molecule Structure-text Model for Text-based Editing and Retrieval

In Submission

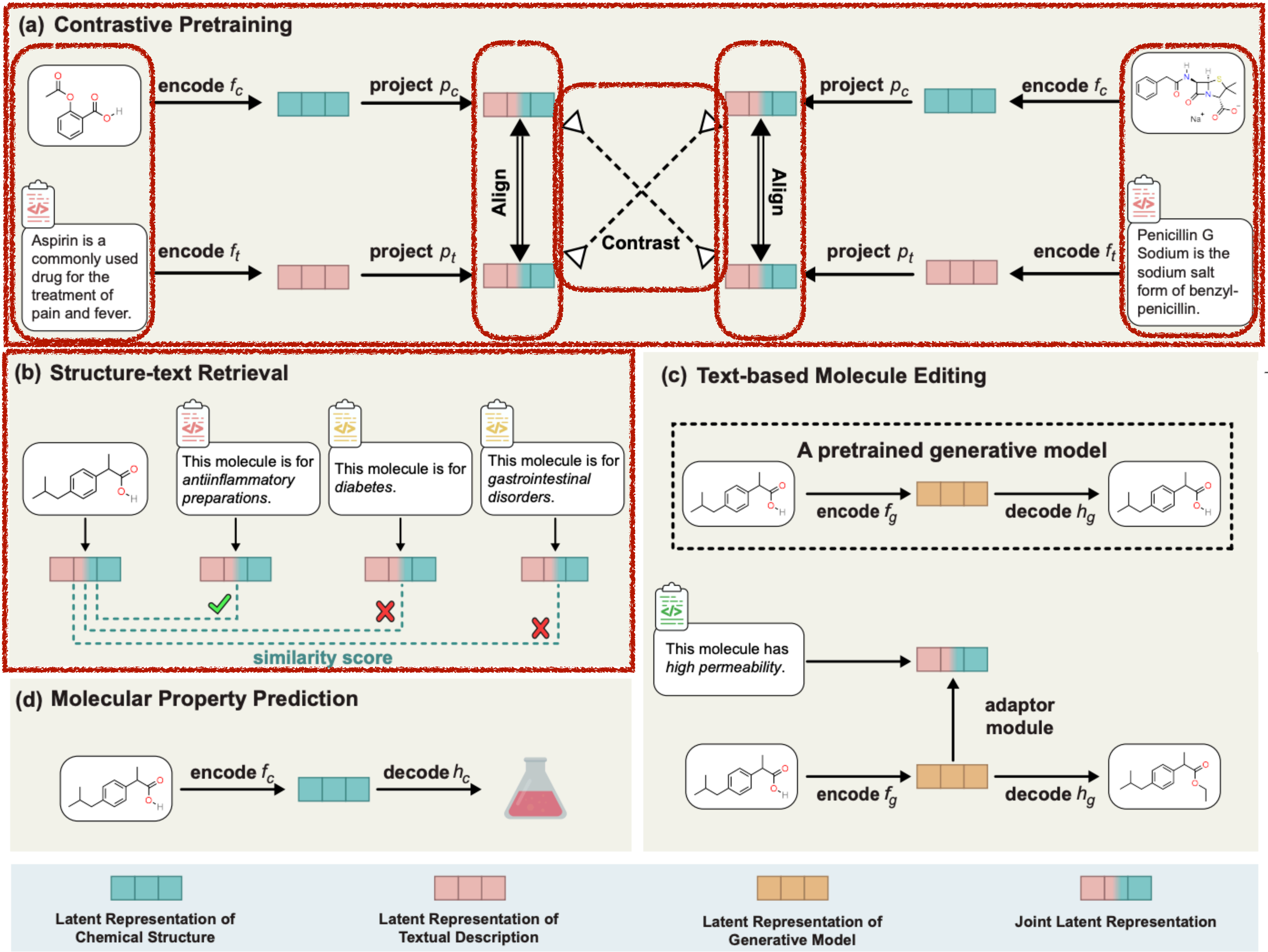
Shengchao Liu, *Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, Anima Anandkumar*

Motivation & Goal

In this work, we want to explore two modalities of molecules:

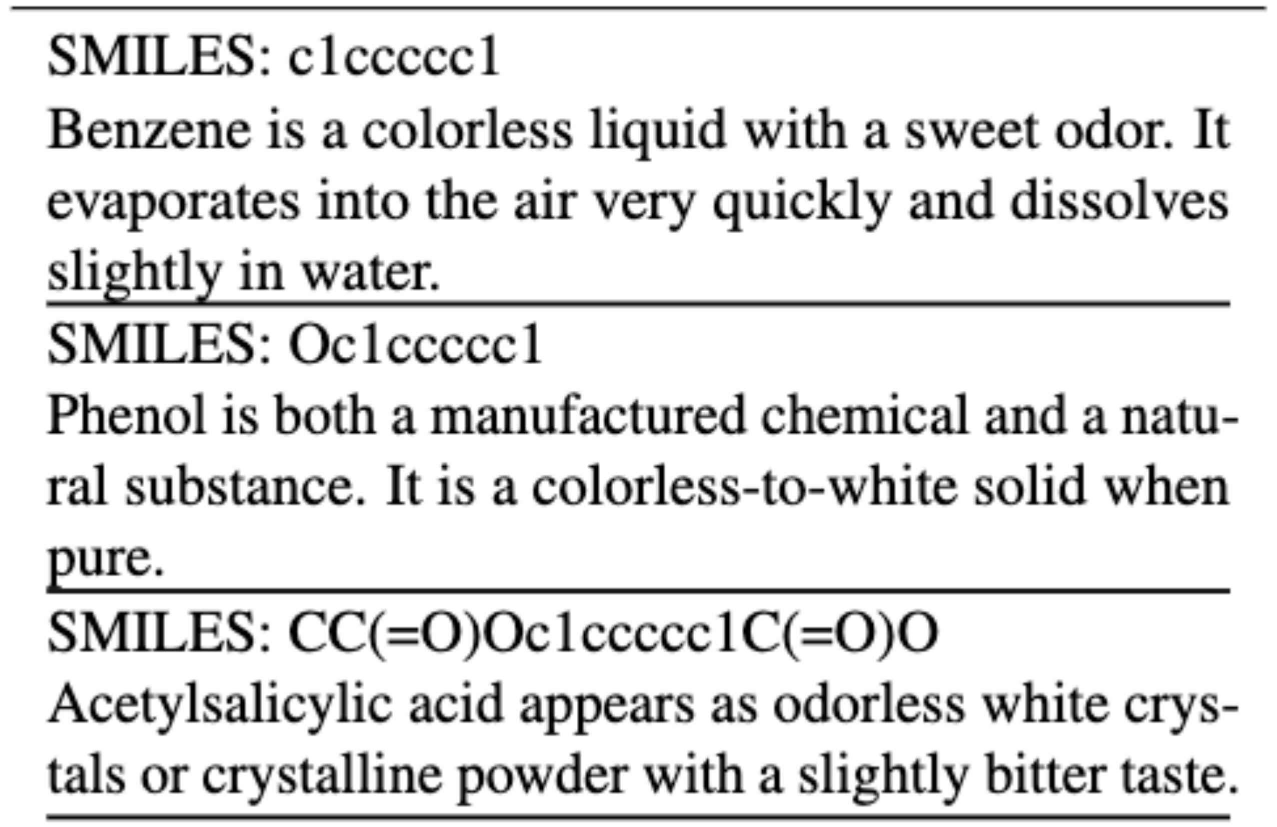
- Chemical structure and domain text.
- Revealing two attributes of natural language.
    - Open vocabulary.
    - Compositionality.

Pipeline


(a) Contrastive Pretraining
encode f_c
project p_c
Align
Contrast
Align
project p_c
encode f_c
Aspirin is a commonly used drug for the treatment of pain and fever.
encode f_t
project p_t
project p_t
encode f_t
Penicillin G Sodium is the sodium salt form of benzyl-penicillin.
(b) Structure-text Retrieval
This molecule is for *antiinflammatory preparations*.
This molecule is for *diabetes*.
This molecule is for *gastrointestinal disorders*.
similarity score
(c) Text-based Molecule Editing
A pretrained generative model
encode f_g
decode h_g
This molecule has *high permeability*.
adaptor module
encode f_g
decode h_g
(d) Molecular Property Prediction
encode f_c
decode h_c
Latent Representation of Chemical Structure
Latent Representation of Textual Description
Latent Representation of Generative Model
Joint Latent Representation

A new dataset PubChemCLIP.

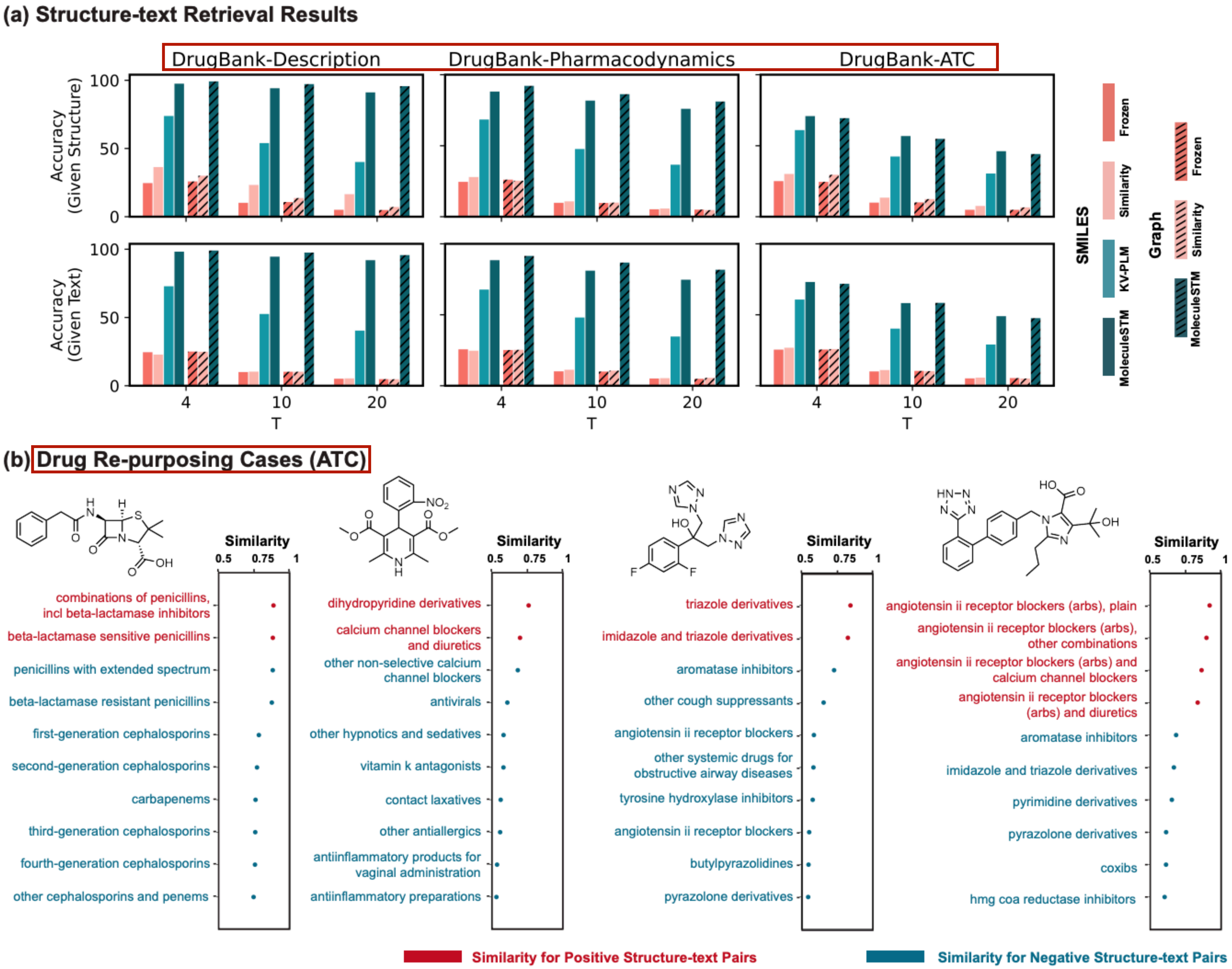
- PubChem has ~110M molecules.
- PubChemCLIP has 280K structure-text pairs.



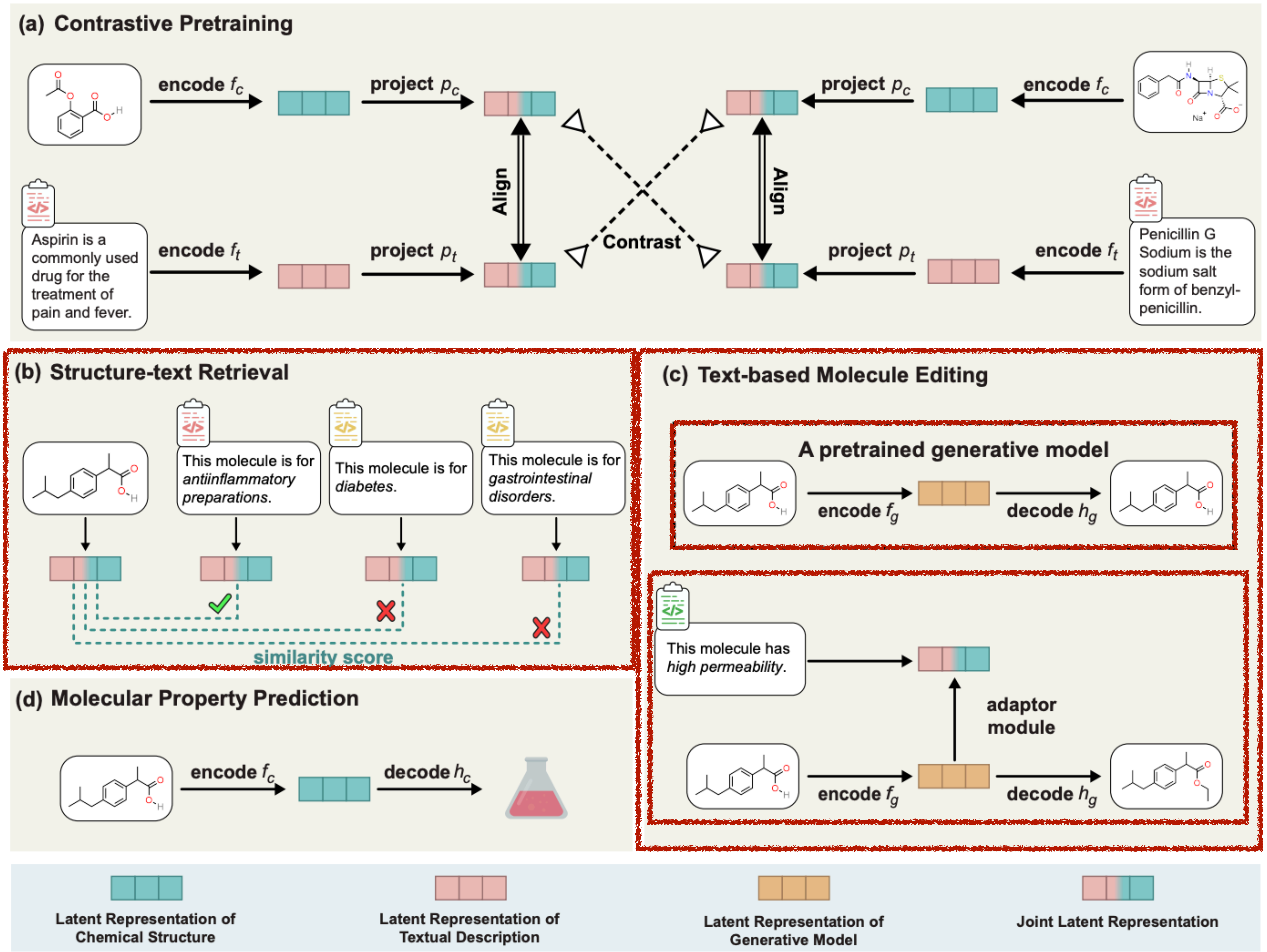
SMILES: c1ccccc1
Benzene is a colorless liquid with a sweet odor. It evaporates into the air very quickly and dissolves slightly in water.

SMILES: Oc1ccccc1
Phenol is both a manufactured chemical and a natural substance. It is a colorless-to-white solid when pure.

SMILES: CC(=O)Oc1ccccc1C(=O)O
Acetylsalicylic acid appears as odorless white crystals or crystalline powder with a slightly bitter taste.


(a) Structure-text Retrieval Results
DrugBank-Description
DrugBank-Pharmacodynamics
DrugBank-ATC
Accuracy (Given Structure)
Accuracy (Given Text)
100
50
0
4
10
20
T
SMILES
Frozen
Similarity
KV-PLM
MoleculeSTM
Graph
Frozen
Similarity
MoleculeSTM
(b) Drug Re-purposing Cases (ATC)
Similarity
0.5
0.75
1
combinations of penicillins, incl beta-lactamase inhibitors
beta-lactamase sensitive penicillins
penicillins with extended spectrum
beta-lactamase resistant penicillins
first-generation cephalosporins
second-generation cephalosporins
carbapenems
third-generation cephalosporins
fourth-generation cephalosporins
other cephalosporins and penems
dihydropyridine derivatives
calcium channel blockers and diuretics
other non-selective calcium channel blockers
antivirals
other hypnotics and sedatives
vitamin k antagonists
contact laxatives
other antiallergics
antiinflammatory products for vaginal administration
antiinflammatory preparations
triazole derivatives
imidazole and triazole derivatives
aromatase inhibitors
other cough suppressants
angiotensin ii receptor blockers
other systemic drugs for obstructive airway diseases
tyrosine hydroxylase inhibitors
angiotensin ii receptor blockers
butylpyrazolidines
pyrazolone derivatives
angiotensin ii receptor blockers (arbs), plain
angiotensin ii receptor blockers (arbs), other combinations
angiotensin ii receptor blockers (arbs) and calcium channel blockers
angiotensin ii receptor blockers (arbs) and diuretics
aromatase inhibitors
imidazole and triazole derivatives
pyrimidine derivatives
pyrazolone derivatives
coxibs
hmg coa reductase inhibitors
Similarity for Positive Structure-text Pairs
Similarity for Negative Structure-text Pairs

Pipeline


(a) Contrastive Pretraining
encode f_c
project p_c
encode f_t
project p_t
Align
Contrast
(b) Structure-text Retrieval
similarity score
(c) Text-based Molecule Editing
A pretrained generative model
encode f_g
decode h_g
adaptor module
(d) Molecular Property Prediction
encode f_c
decode h_c
Latent Representation of Chemical Structure
Latent Representation of Textual Description
Latent Representation of Generative Model
Joint Latent Representation

Aspirin is a commonly used drug for the treatment of pain and fever.

Penicillin G Sodium is the sodium salt form of benzyl-penicillin.

This molecule is for *antiinflammatory preparations*.

This molecule is for *diabetes*.

This molecule is for *gastrointestinal disorders*.

This molecule has *high permeability*.

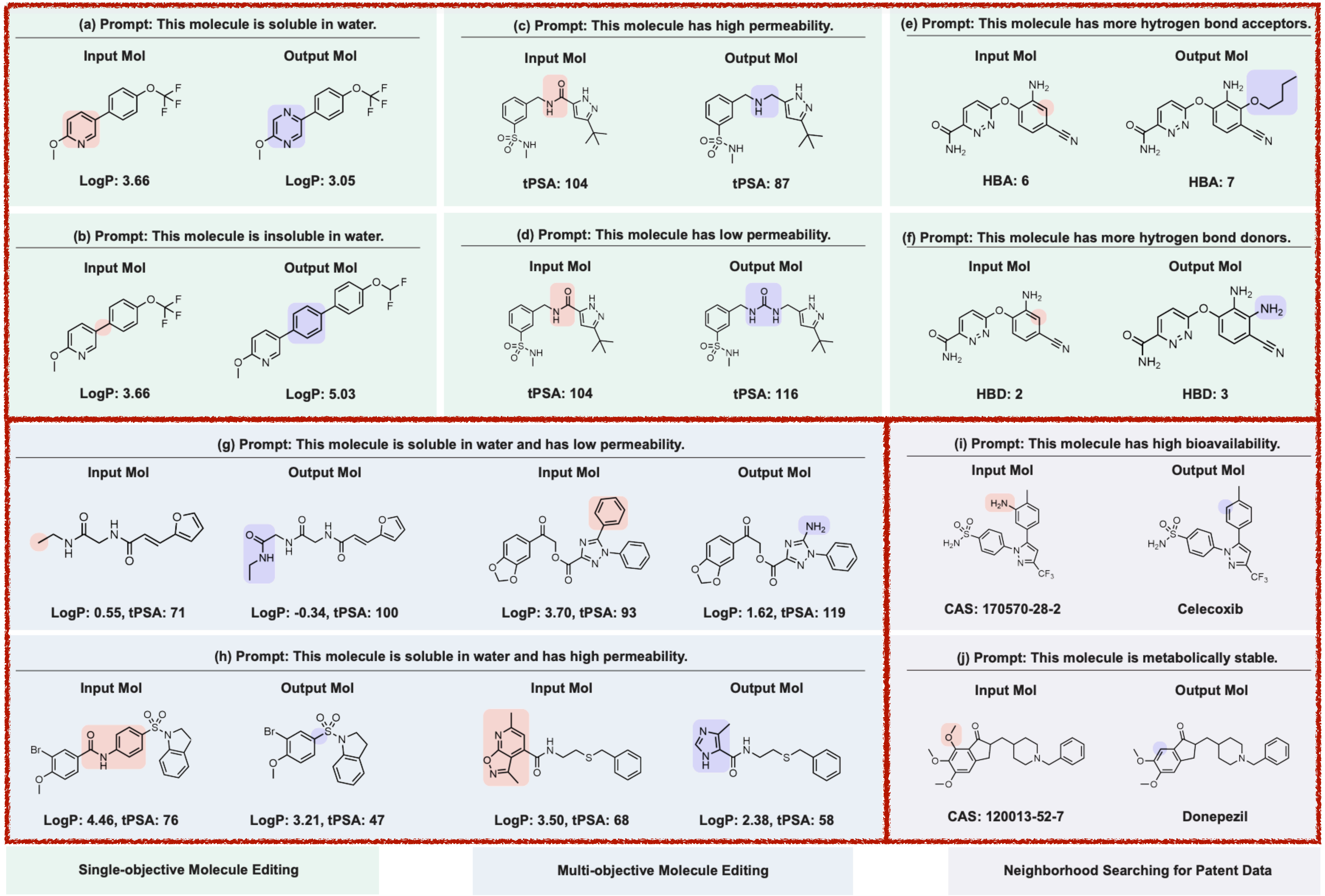
Zero-shot Text-guided Molecule Editing

Phase 1: Space Alignment

Phase 2: Latent Optimization

This molecule has high permeability.

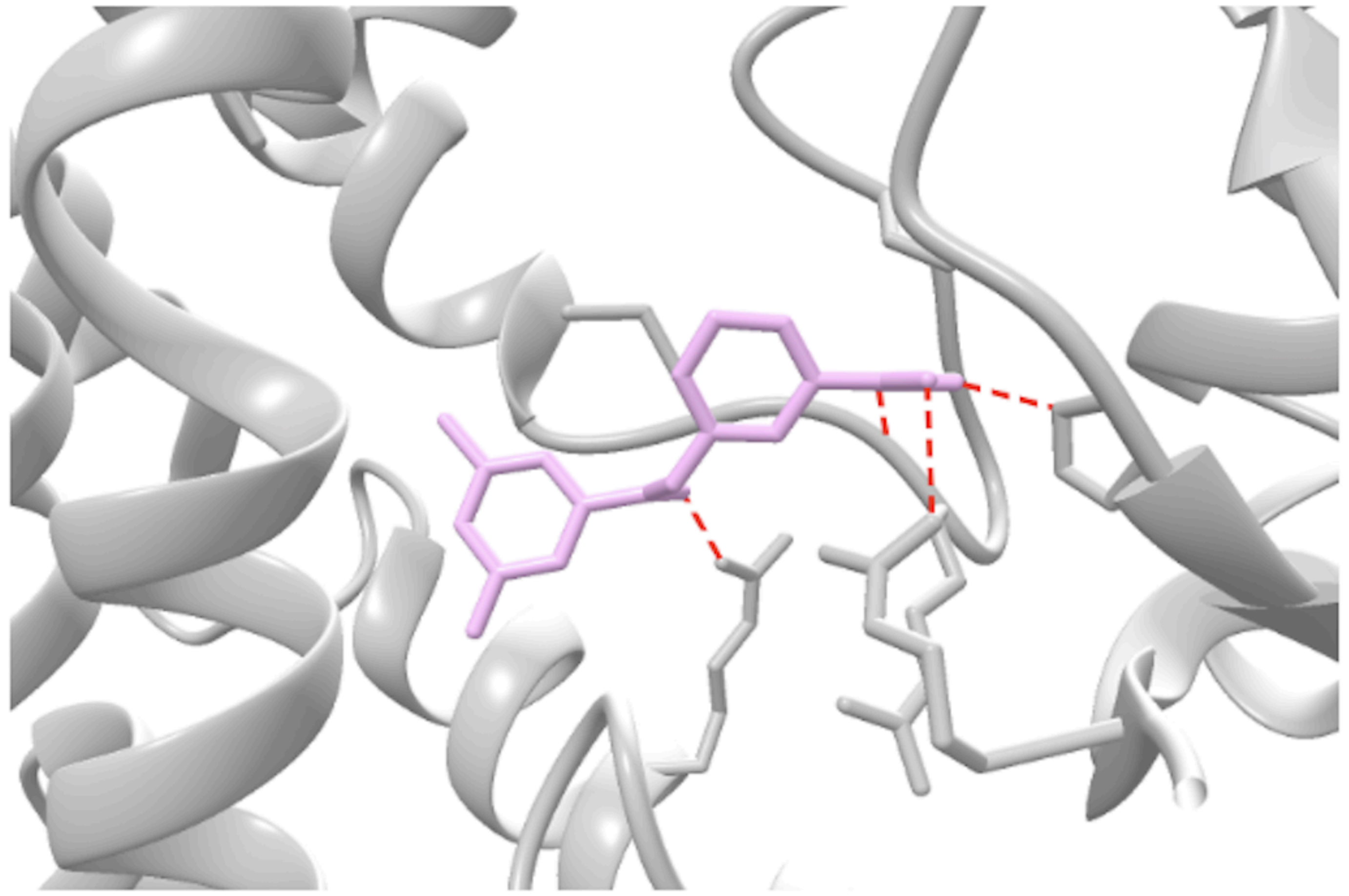
encode
encode
project
align m_{g2f}
encode
similarity
similarity
encode
project
decode


(a) Prompt: This molecule is soluble in water.
Input Mol
Output Mol
LogP: 3.66
LogP: 3.05
(b) Prompt: This molecule is insoluble in water.
Input Mol
Output Mol
LogP: 3.66
LogP: 5.03
(c) Prompt: This molecule has high permeability.
Input Mol
Output Mol
tPSA: 104
tPSA: 87
(d) Prompt: This molecule has low permeability.
Input Mol
Output Mol
tPSA: 104
tPSA: 116
(e) Prompt: This molecule has more hytrogen bond acceptors.
Input Mol
Output Mol
HBA: 6
HBA: 7
(f) Prompt: This molecule has more hytrogen bond donors.
Input Mol
Output Mol
HBD: 2
HBD: 3
(g) Prompt: This molecule is soluble in water and has low permeability.
Input Mol
Output Mol
Input Mol
Output Mol
LogP: 0.55, tPSA: 71
LogP: -0.34, tPSA: 100
LogP: 3.70, tPSA: 93
LogP: 1.62, tPSA: 119
(h) Prompt: This molecule is soluble in water and has high permeability.
Input Mol
Output Mol
Input Mol
Output Mol
LogP: 4.46, tPSA: 76
LogP: 3.21, tPSA: 47
LogP: 3.50, tPSA: 68
LogP: 2.38, tPSA: 58
(i) Prompt: This molecule has high bioavailability.
Input Mol
Output Mol
CAS: 170570-28-2
Celecoxib
(j) Prompt: This molecule is metabolically stable.
Input Mol
Output Mol
CAS: 120013-52-7
Donepezil
Single-objective Molecule Editing
Multi-objective Molecule Editing
Neighborhood Searching for Patent Data

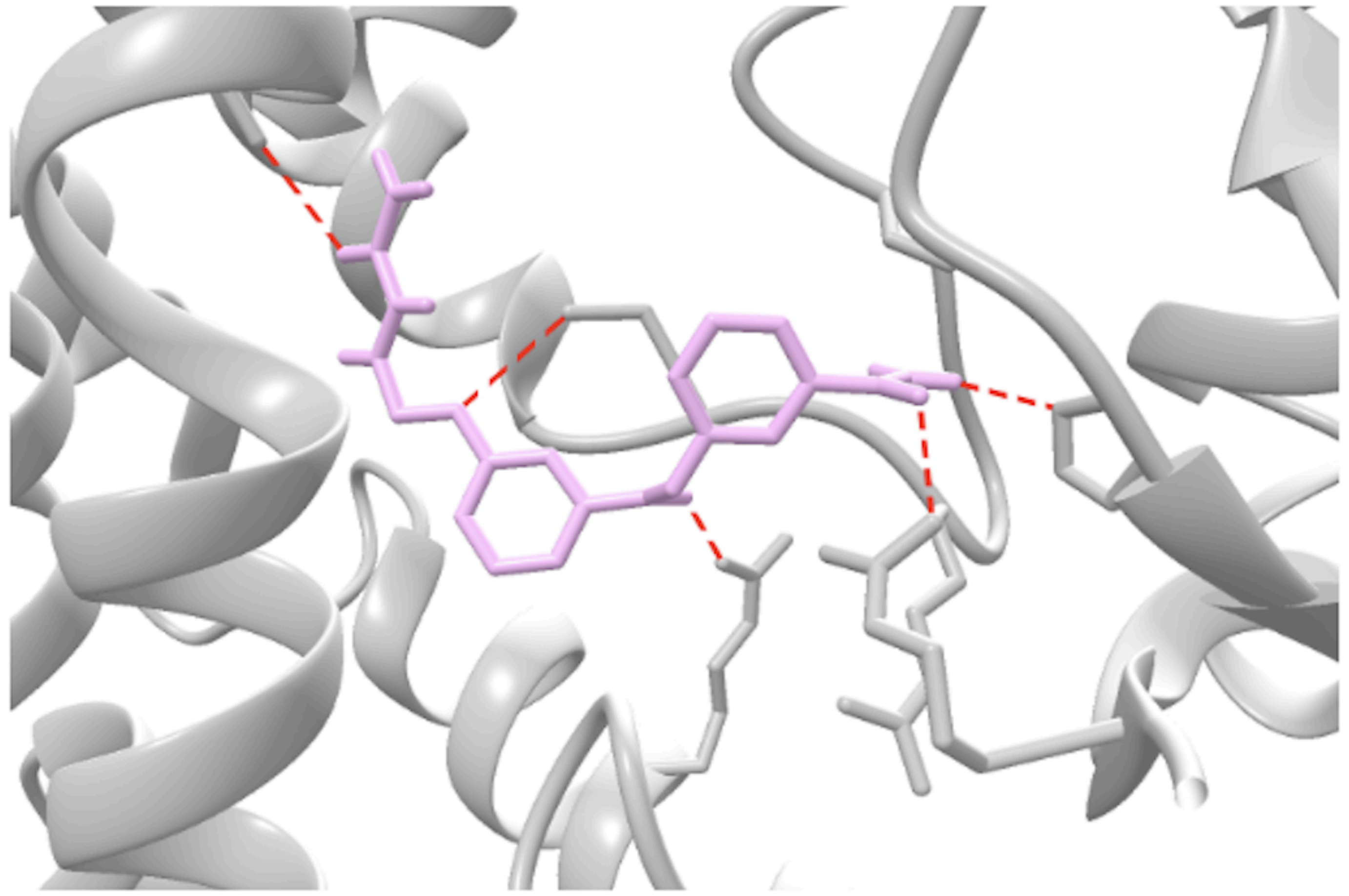
Text prompt, ChEMBL 1613777:

“This molecule is tested positive in an assay that are inhibitors and substrates of an enzyme protein. It uses molecular oxygen inserting one oxygen atom into a substrate, and reducing the second into a water molecule.”

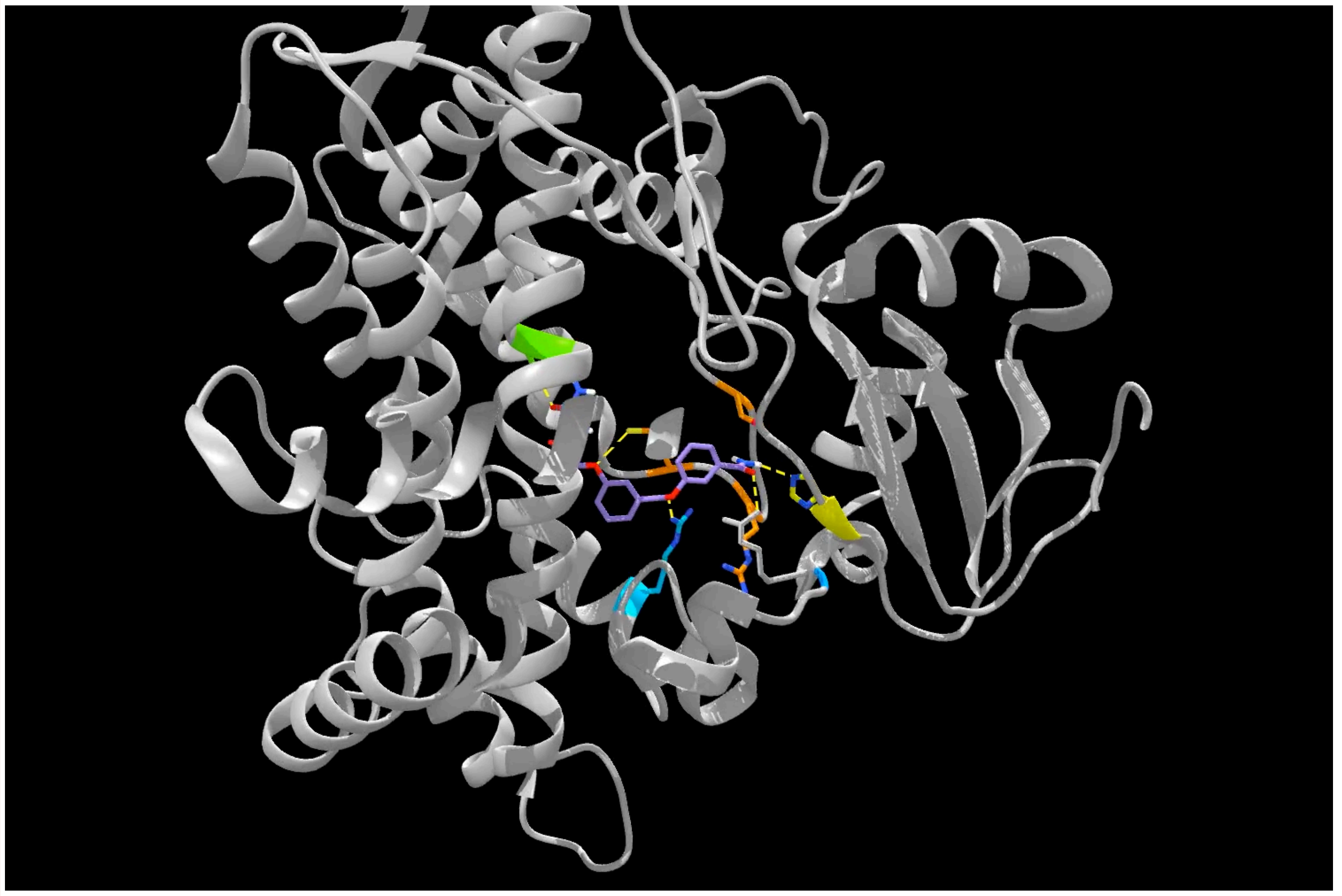
(a) Set 1, input molecule (SMILES): Cc1cc(F)cc(C(=O)Oc2cccc(C(N)=O)c2)c1



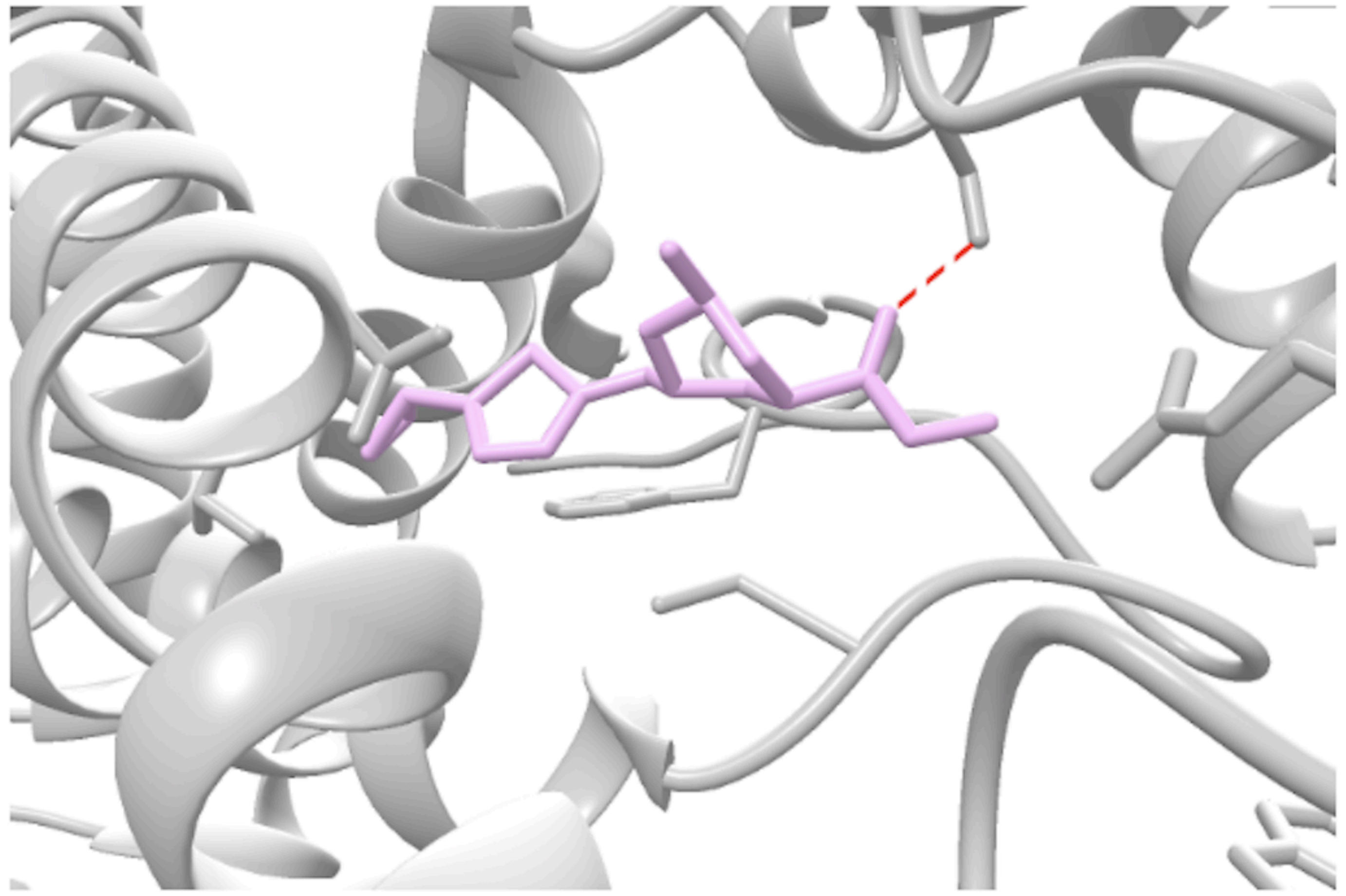
Input Molecule
(docking score: -9.055)



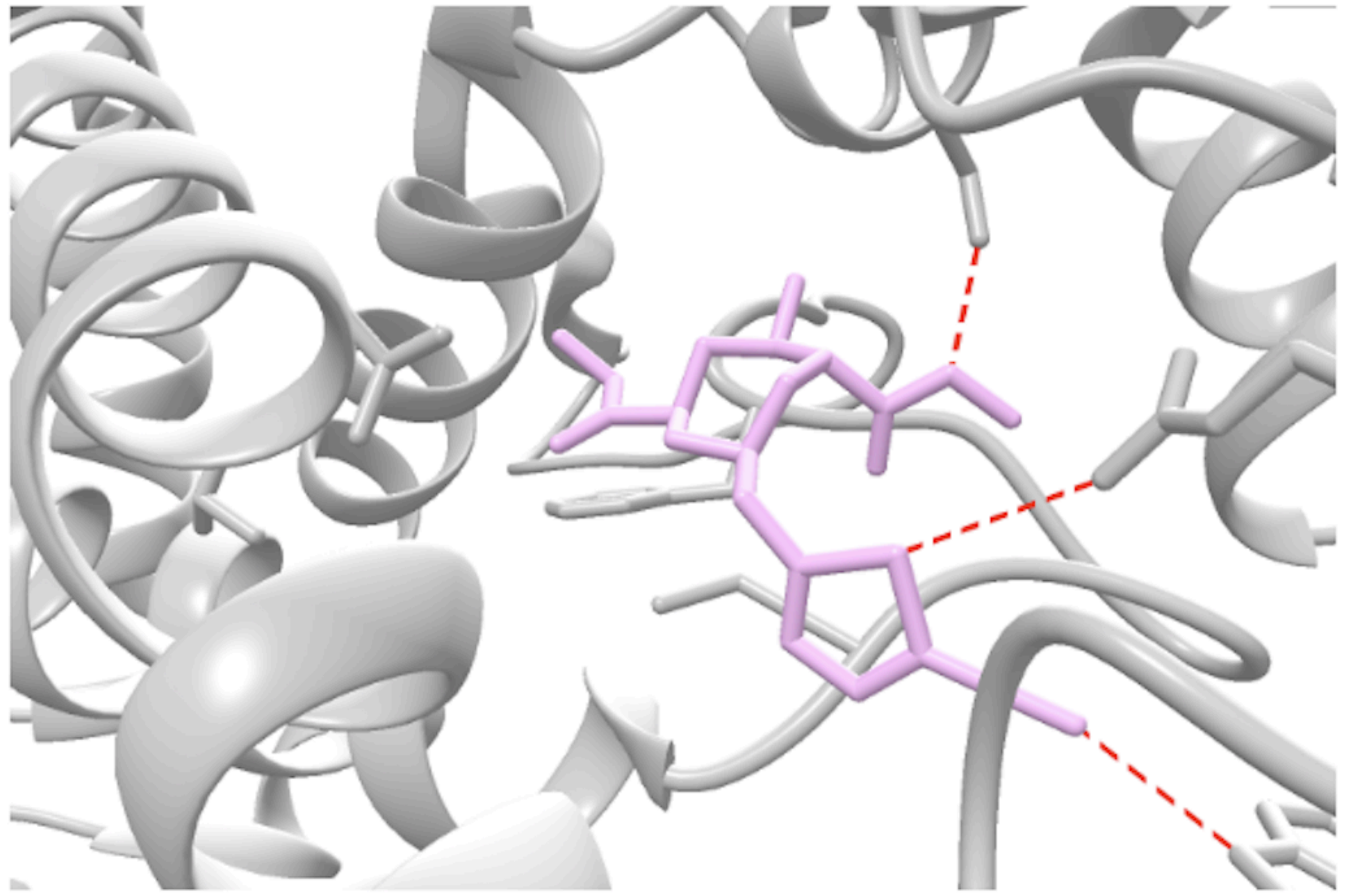
Output Molecule with MoleculeSTM
(docking score: -10.35)



(b) Set 2, input molecule (SMILES): COC(=O)[C@@H]1CN(Cc2cnc(C3CC3)s2)C[C@@H](C)O1

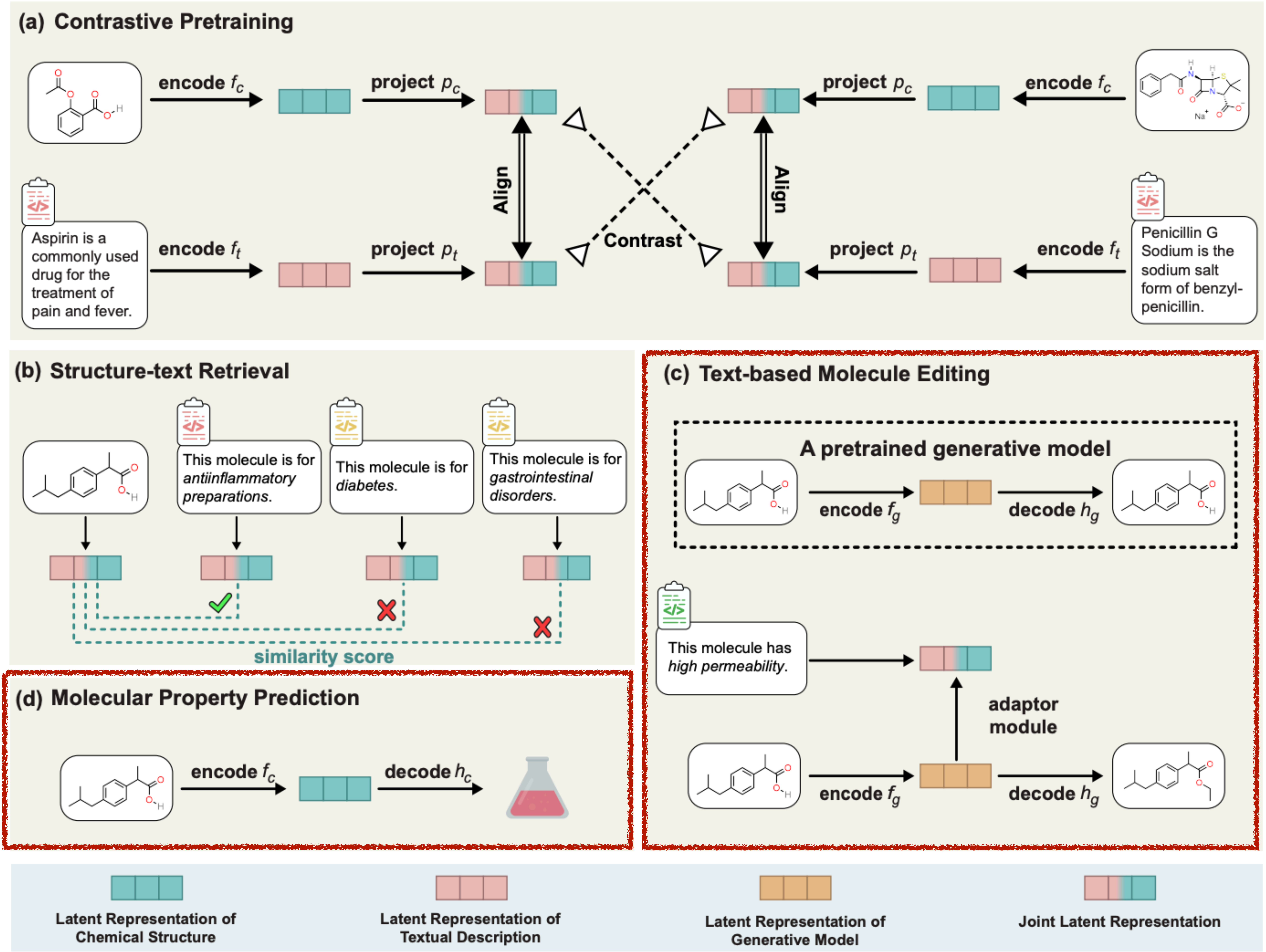


Input Molecule
(docking score: -7.441)



Output Molecule with MoleculeSTM
(docking score: -11.363)

Pipeline


(a) Contrastive Pretraining
encode f_c
project p_c
Align
Contrast
encode f_t
project p_t
(b) Structure-text Retrieval
similarity score
(c) Text-based Molecule Editing
A pretrained generative model
encode f_g
decode h_g
adaptor module
(d) Molecular Property Prediction
encode f_c
decode h_c
Latent Representation of Chemical Structure
Latent Representation of Textual Description
Latent Representation of Generative Model
Joint Latent Representation

Aspirin is a commonly used drug for the treatment of pain and fever.

Penicillin G Sodium is the sodium salt form of benzyl-penicillin.

This molecule is for *antiinflammatory preparations*.

This molecule is for *diabetes*.

This molecule is for *gastrointestinal disorders*.

This molecule has *high permeability*.

Follow-up [in submission]

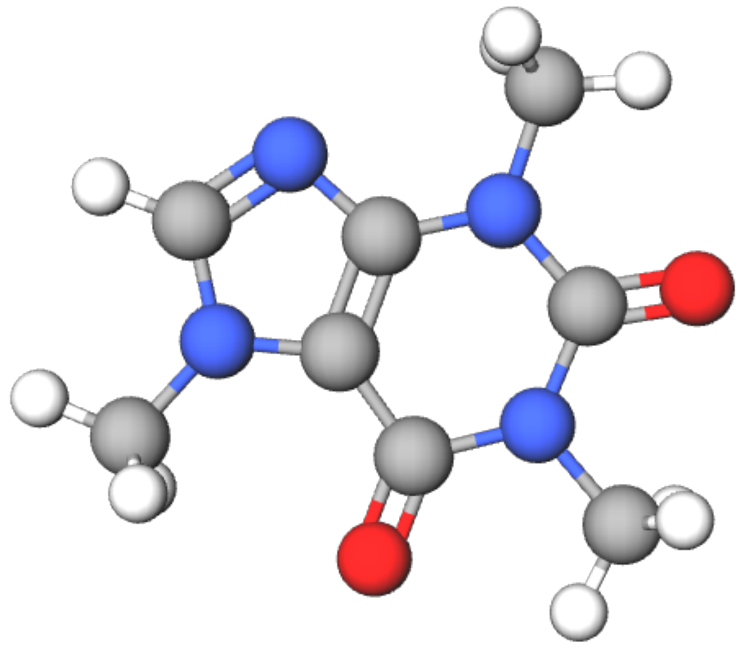
Bridging the gap between protein sequence/structure and natural language.

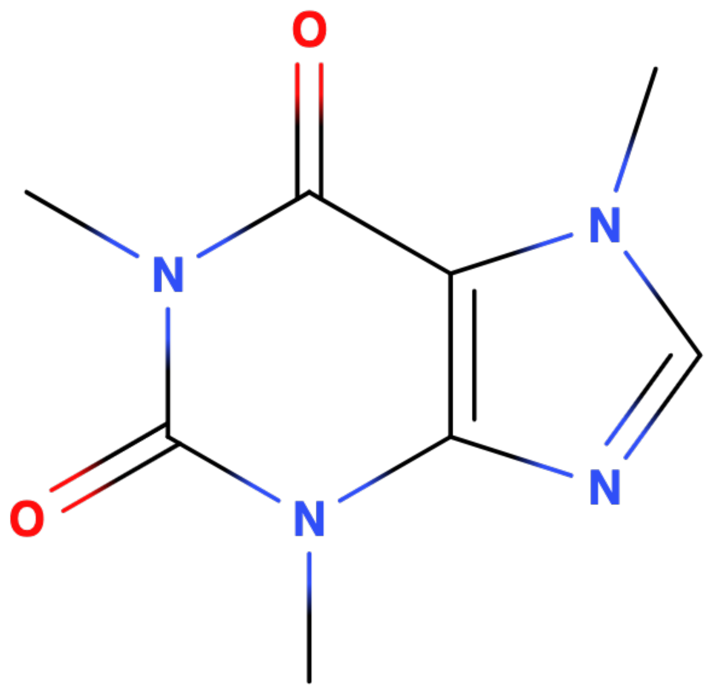
(7) SGNN-EBM (Multi-task Learning)
AISTATS’22

Biological Knowledge Graph (KG)
GNN

String (SMILES, SELFIES)
CNN, RNN, LM

(3) MoleculeSTM (Foundation Model)
In submission





Biomedical Text

Salicylic acid is a phenolic compound present in the plants, where it plays a central role in the development of resistance to pathogen infection.

(1) GraphMVP (2D-3D SSL)
ICLR’22

3D Molecular Graph
3D GNN

2D Molecular Graph
2D GNN

Textual Description
RNN, LM

(2) GeoSSL (3D SSL)
ICLR’23

(5) ProteinDT (Foundation Model)
In Submission

(6) Geometric Benchmark
ongoing

(4) GraphCG (molecule editing)
In submission

Thank you!

Q&A