Multi-modal molecule structure–text model for text-based retrieval and editing (Nat. Mach. Intell)

By *Dr. Jaewoong Choi*

(Dr. B. Lee group @KIST)

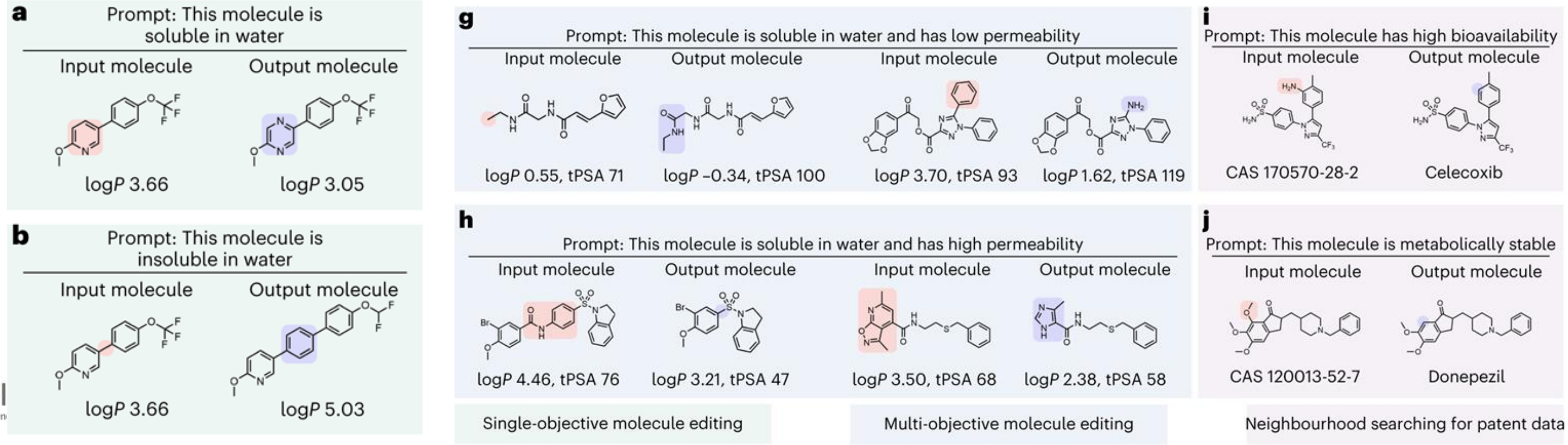
2024.01.11

Research purpose

❖ They present a multi-modal molecule structure–text model, MoleculeSTM,

✓ by jointly learning molecules' chemical structures and textual descriptions via a contrastive learning strategy.

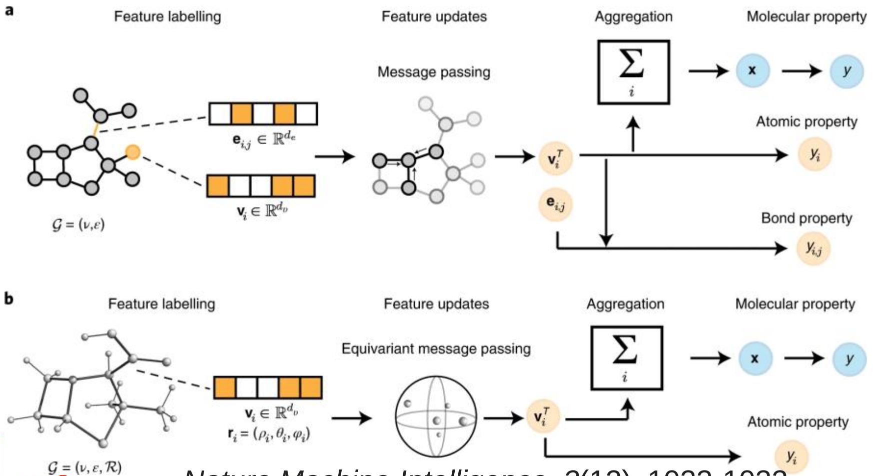
- Existing studies used chemical structures of molecules without textual knowledge, which enables to realize new drug design objectives, adapt to text-based instructions and predict complex biological activities.
- Solving three challenging tasks: Zero-shot structure-text retrieval, text-based molecule editing, + molecular property prediction.



Previous works

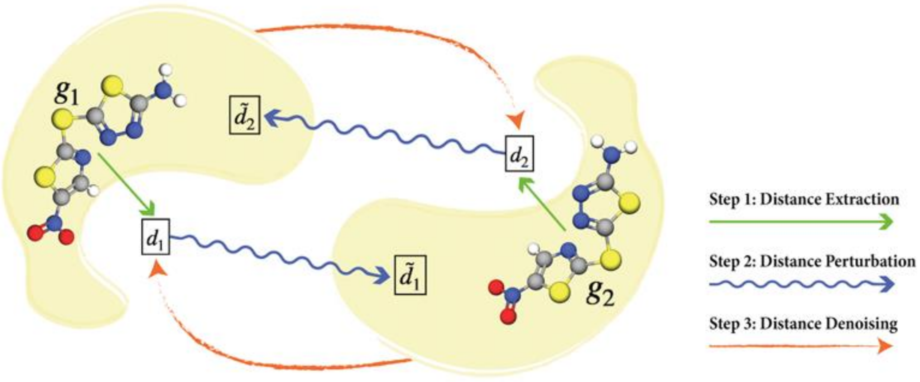
- ❖ Existing ML methods focus on modelling the chemical structure of molecules through 1 or 2 dimensional molecular graphs or 3-dimensional geometric structures in a supervised manner.
  - ✓ Supervised methods require expensive annotations on pre-determined label categories.
    - Unsupervised methods (pretrained models) are effective, but it is still an open challenge to generalize unseen categories and tasks without such labelled examples or fine-tuning → zero-shot setting X.

Supervised molecule learning



Nature Machine Intelligence, *3*(12), 1023-1032.

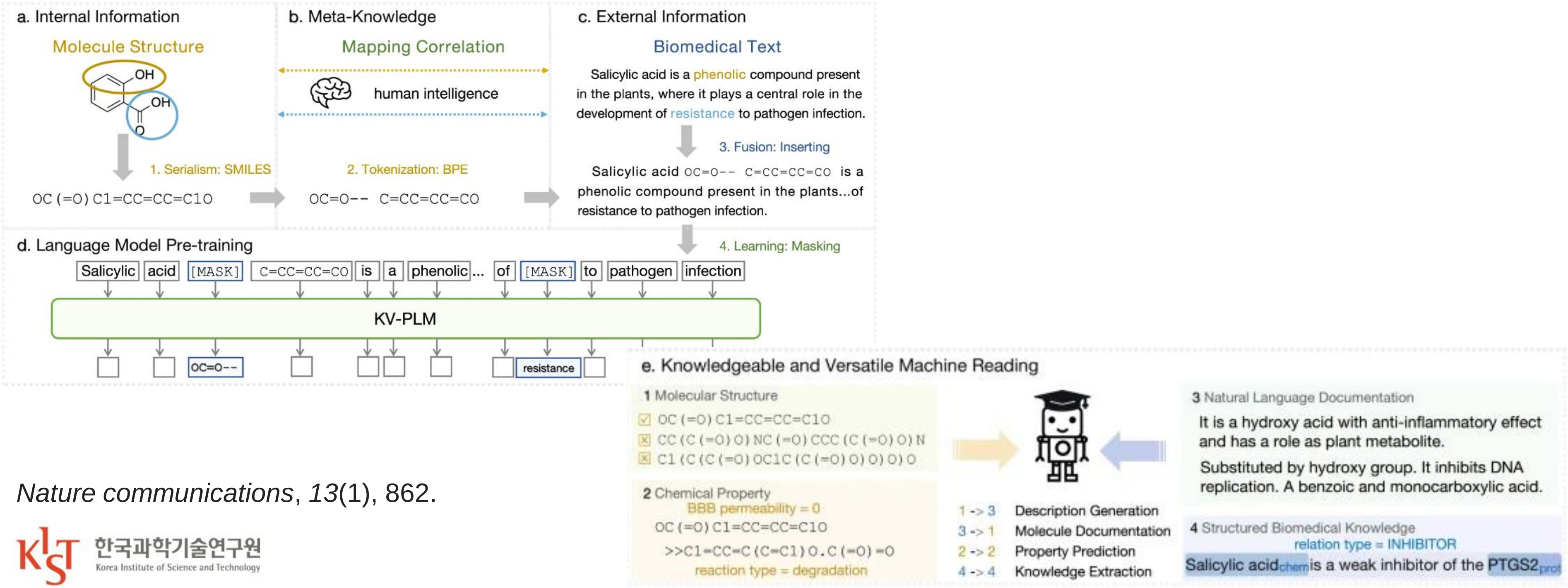
Unsupervised molecule learning


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

Molecular geometry pretraining with SE(3)-invariant denoising distance matching.

Previous works

- ❖ Previous work using textual knowledge for molecule representation.
  - ✓ only modelling with the and learns the one-dimensional chemical structures (SMILES) and textual descriptions on a small-scale dataset (*N* = 10,000).
    - It cannot adopt existing powerful pretrained models and the availability of aligned data is extremely limited.

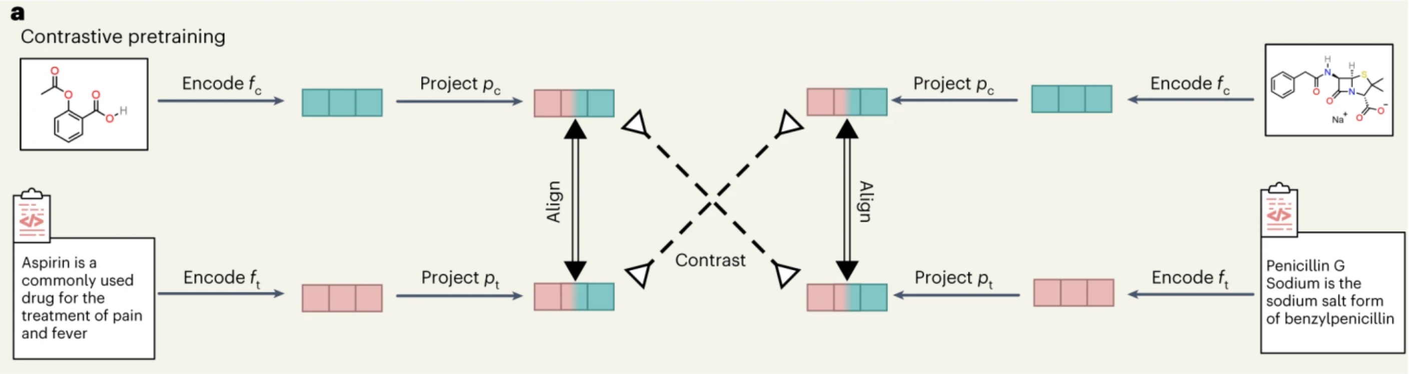

a. Internal Information
Molecule Structure
OH
OH
O
1. Serialism: SMILES
OC(=O)C1=CC=CC=C1O
b. Meta-Knowledge
Mapping Correlation
human intelligence
2. Tokenization: BPE
OC=O-- C=CC=CC=CO
c. External Information
Biomedical Text
Salicylic acid is a phenolic compound present in the plants, where it plays a central role in the development of resistance to pathogen infection.
3. Fusion: Inserting
Salicylic acid OC=O-- C=CC=CC=CO is a phenolic compound present in the plants...of resistance to pathogen infection.
4. Learning: Masking
d. Language Model Pre-training
Salicylic
acid
[MASK]
C=CC=CC=CO
is
a
phenolic...
of
[MASK]
to
pathogen
infection
KV-PLM
OC=O--
resistance
e. Knowledgeable and Versatile Machine Reading
1 Molecular Structure
☑ OC(=O)C1=CC=CC=C1O
☒ CC(C(=O)O)NC(=O)CCC(C(=O)O)N
☒ C1(C(C(=O)OC1C(C(=O)O)O)O)O
2 Chemical Property
BBB permeability = 0
OC(=O)C1=CC=CC=C1O
>>C1=CC=C(C=C1)O.C(=O)=O
reaction type = degradation
1 -> 3 Description Generation
3 -> 1 Molecule Documentation
2 -> 2 Property Prediction
4 -> 4 Knowledge Extraction
3 Natural Language Documentation
It is a hydroxy acid with anti-inflammatory effect and has a role as plant metabolite.
Substituted by hydroxy group. It inhibits DNA replication. A benzoic and monocarboxylic acid.
4 Structured Biomedical Knowledge
relation type = INHIBITOR
Salicylic acid chem is a weak inhibitor of the PTGS2 prot

Nature communications, *13*(1), 862.

한국과학기술연구원
Korea Institute of Science and Technology

Proposed approach

- MoleculeSTM, consisting of 2 branches, (1) the chemical structure branch and (2) the textual description branch.
  - ✓ With pretrained models for molecular structure and scientific language, MoleculeSTM bridges the two branches via a contrastive learning paradigm
  - ✓ They construct a structure–text dataset called PubChemSTM from PubChem.
    - Each chemical structure is paired with a textual description, illustrating the chemical and physical properties or high-level bioactivities accordingly.



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

Penicillin G Sodium is the sodium salt form of benzylpenicillin

Proposed approach

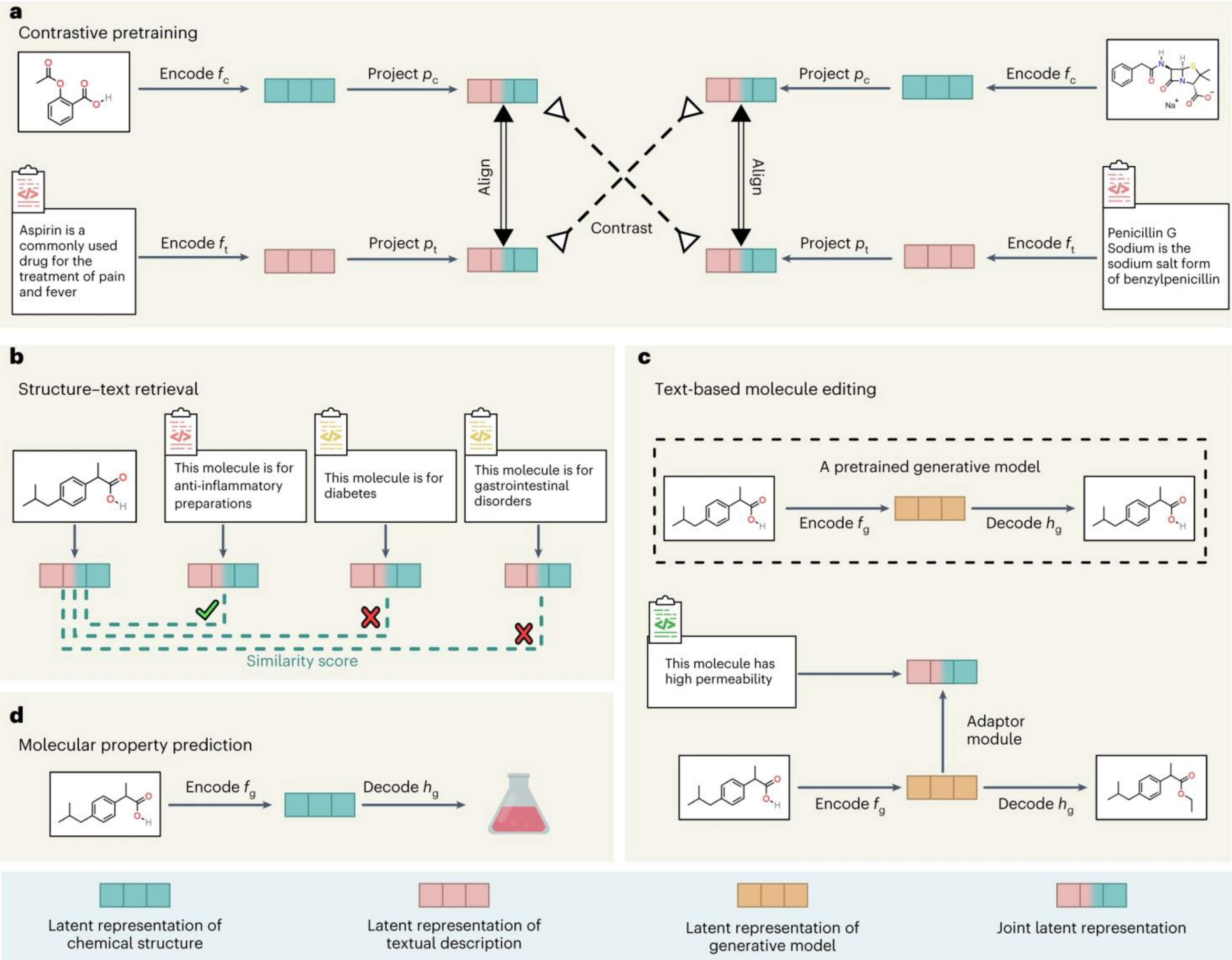
❖ MoleculeSTM is tested on the structure–text retrieval task and text-based molecule editing task in a zero-shot manner.
→ 6 zero-shot retrieval tasks (up to 50% higher accuracy) and 20 zero-shot text-based editing tasks (up to 40% higher hit ratio)

- ✓ (1) Open vocabulary: not limited to a fixed set of pre-defined molecule-related textual descriptions and can support exploring a wide range of biochemical concepts with the unbound vocabulary depicted by the natural language.
- ✓ (2) Compositionality: express a complex concept by decomposing it into several simple concepts, which is applied for the text-based multi-objective lead optimization task, rather than property-by-property filtering or optimizing a database.

‘molecule is soluble in water and has high permeability’

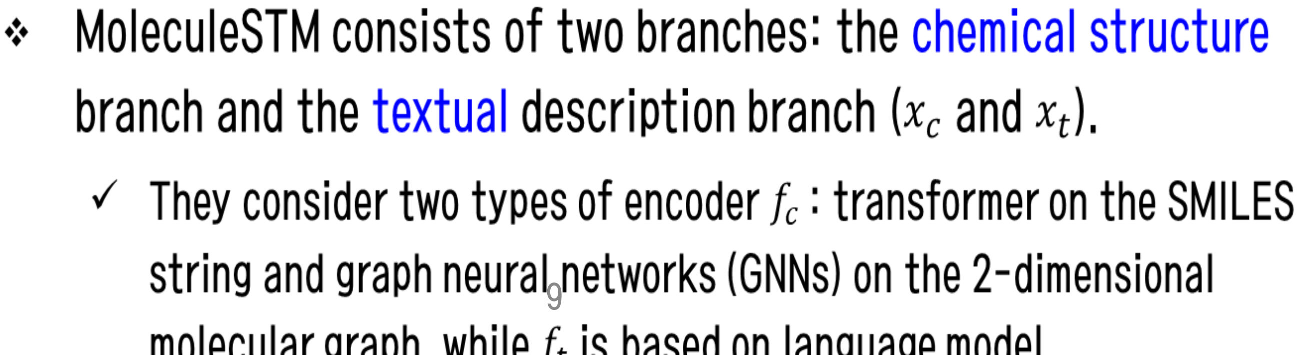
Proposed approach

- ❖ Pipeline of pretraining and downstream works of MoleculeSTM


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 benzylpenicillin
b
Structure–text retrieval
This molecule is for anti-inflammatory 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_g
Decode h_g
Latent representation of chemical structure
Latent representation of textual description
Latent representation of generative model
Joint latent representation

Proposed approach

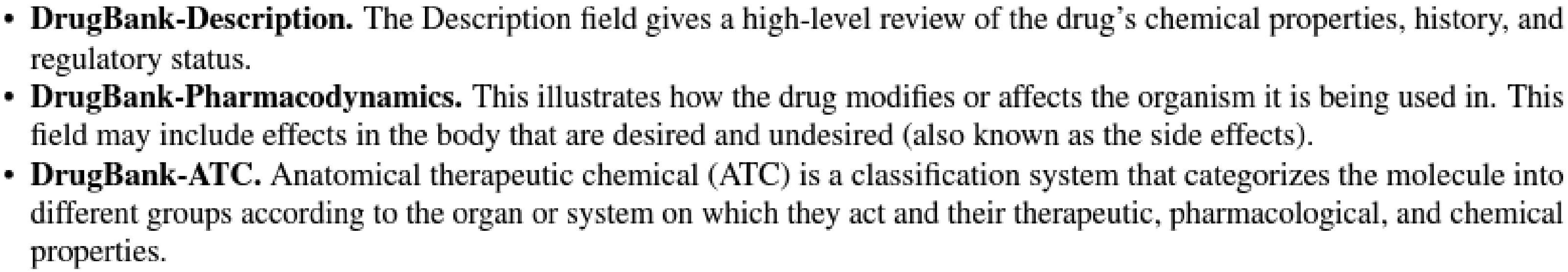
- MoleculeSTM consists of two branches: the chemical structure branch and the textual description branch (x_c and x_t).
  - ✓ They consider two types of encoder f_c : transformer on the SMILES string and graph neural networks (GNNs) on the 2-dimensional molecular graph, while f_t is based on language model.
  - ✓ Pretraining; MoleculeSTM aims to map the representations extracted from two branches to a joint space using two projectors (p_c and p_t) via contrastive learning.
    - From PubChem, they extract molecules with the textual description fields, leading to 281,000 chemical structure and text pairs.



- MoleculeSTM consists of two branches: the chemical structure branch and the textual description branch (x_c and x_t).
  - ✓ They consider two types of encoder f_c : transformer on the SMILES string and graph neural networks (GNNs) on the 2-dimensional molecular graph, while f_t is based on language model

MoleculeSTM pretraining - Dataset construction

❖ PubChem String field to construct a large-scale dataset called PubChemSTM, consisting of 250,000 molecules and 281,000 structure–text pairs.



- **DrugBank-Description.** The Description field gives a high-level review of the drug's chemical properties, history, and regulatory status.
- **DrugBank-Pharmacodynamics.** This illustrates how the drug modifies or affects the organism it is being used in. This field may include effects in the body that are desired and undesired (also known as the side effects).
- **DrugBank-ATC.** Anatomical therapeutic chemical (ATC) is a classification system that categorizes the molecule into different groups according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties.

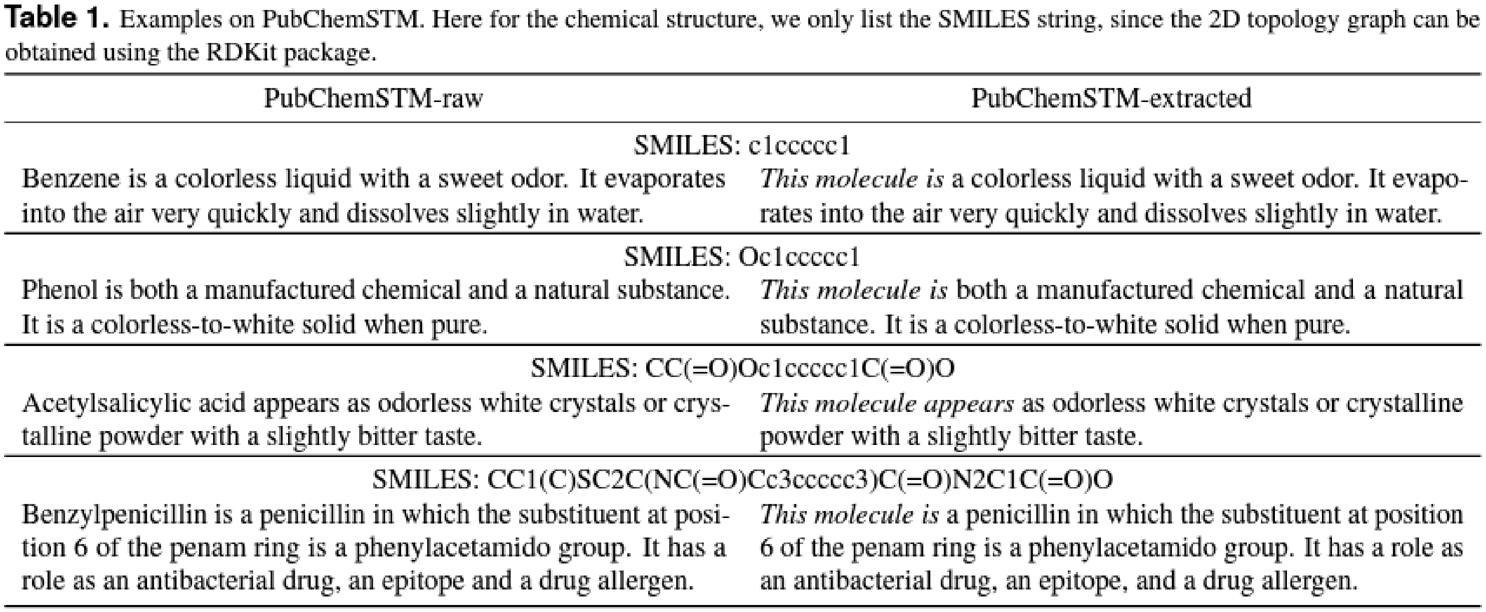


Table 1. Examples on PubChemSTM. Here for the chemical structure, we only list the SMILES string, since the 2D topology graph can be obtained using the RDKit package.

PubChemSTM-raw	PubChemSTM-extracted
SMILES: c1ccccc1	
Benzene is a colorless liquid with a sweet odor. It evaporates into the air very quickly and dissolves slightly in water.	*This molecule 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.	*This molecule 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.	*This molecule appears* as odorless white crystals or crystalline powder with a slightly bitter taste.
SMILES: CC1(C)SC2C(NC(=O)Cc3ccccc3)C(=O)N2C1C(=O)O	
Benzylpenicillin is a penicillin in which the substituent at position 6 of the penam ring is a phenylacetamido group. It has a role as an antibacterial drug, an epitope and a drug allergen.	*This molecule is* a penicillin in which the substituent at position 6 of the penam ring is a phenylacetamido group. It has a role as an antibacterial drug, an epitope, and a drug allergen.

MoleculeSTM pretraining – Chemical structure and textual branch

- f_c uses two types of chemical structure: the SMILES string views the molecule as a sequence, and the two-dimensional molecular graph takes the atoms and bonds as the nodes and edges.
  - ✓ For the SMILES string, we take the encoder from MegaMolBART, pretrained on 500 million molecules from the ZINC database.
  - ✓ For the molecular graph, we take a pretrained graph isomorphism network (GIN) using GraphMVP pretraining.
- f_t provides a high-level description of the molecule's functionality.
  - ✓ We further adapt the pretrained SciBERT, which was pretrained on the textual data from the chemical and biological domain.

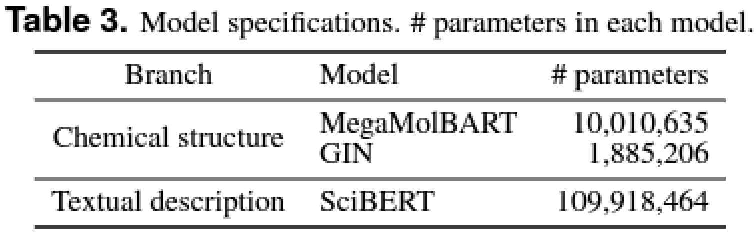


Table 3. Model specifications. # parameters in each model.

Branch	Model	# parameters
Chemical structure	MegaMolBART GIN	10,010,635 1,885,206
Textual description	SciBERT	109,918,464

MoleculeSTM pretraining - Contrastive pretraining

- Two contrastive learning strategies are tested,
  - ✓ ‘EBM-NCE’ and ‘InfoNCE’, which align the structure–text pairs for the same molecule and contrast the pairs for different molecules simultaneously.

$$\begin{aligned}\mathcal{L}_{\text{EBM-NCE}} &= -\frac{1}{2}\Big(\mathbb{E}_{\mathbf{x}_c,\mathbf{x}_t}\left[\log\sigma\left(E\left(\mathbf{x}_c,\mathbf{x}_t\right)\right)\right] + \mathbb{E}_{\mathbf{x}_c,\mathbf{x}_t'}\left[\log\left(1-\sigma(E(\mathbf{x}_c,\mathbf{x}_t'))\right]\Big) \\ &\qquad + \mathbb{E}_{\mathbf{x}_c,\mathbf{x}_t}\left[\log\sigma\left(E\left(\mathbf{x}_c,\mathbf{x}_t\right)\right)\right] + \mathbb{E}_{\mathbf{x}_c',\mathbf{x}_t}\left[\log\left(1-\sigma(E(\mathbf{x}_c',\mathbf{x}_t))\right]\Big),\end{aligned}$$

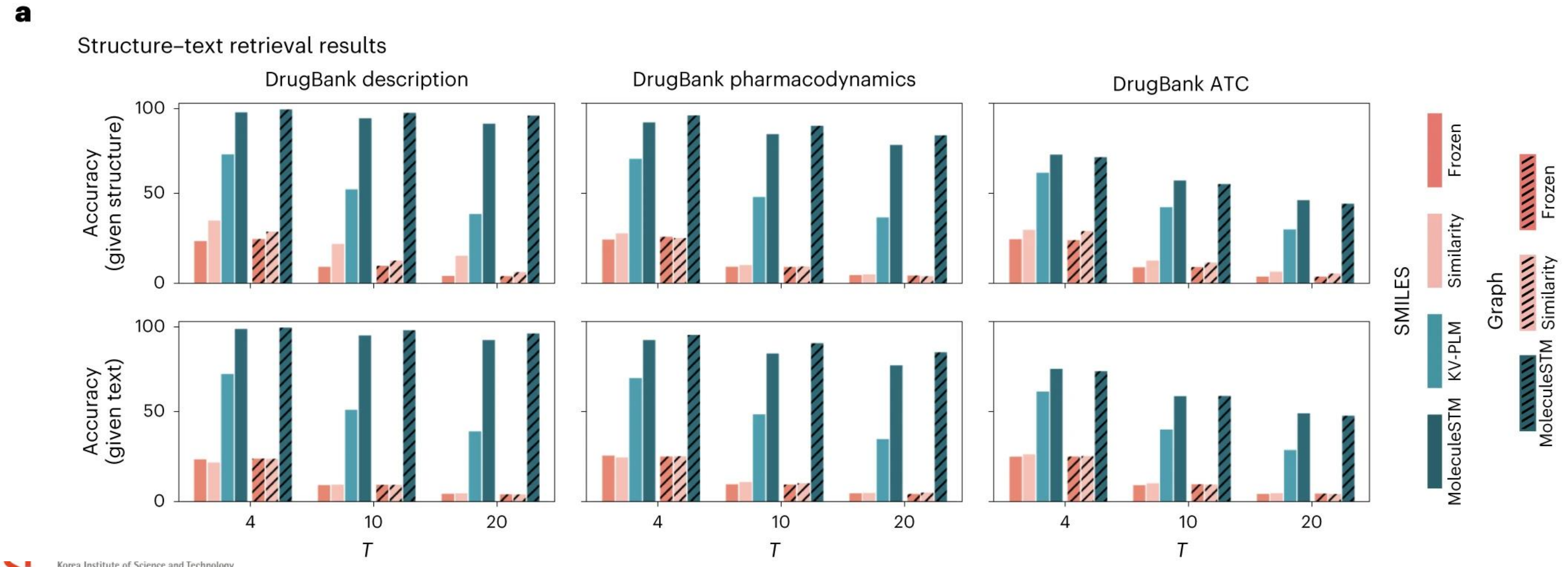
$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2}\mathbb{E}_{\mathbf{x}_c,\mathbf{x}_t}\left[\log\frac{\exp(E(\mathbf{x}_c,\mathbf{x}_t))}{\exp(E(\mathbf{x}_c,\mathbf{x}_t))+\sum_{\mathbf{x}_{t'}}\exp(E(\mathbf{x}_c,\mathbf{x}_{t'}))} + \log\frac{\exp(E(\mathbf{x}_c,\mathbf{x}_t))}{\exp(E(\mathbf{x}_c,\mathbf{x}_t))+\sum_{\mathbf{x}_{c'}}\exp(E(\mathbf{x}_{c'},\mathbf{x}_t))}\right], \tag{1}$$

- where σ is sigmoid function, x_c and x_t is structure and text of each molecule, $x_{c'}$ and $x_{t'}$ is a randomly selected negative sample from noise distribution. E() is the energy function with a flexible formulation, and we use the dot product on the jointly learned space, that is, E(x_c, x_t) = $\langle p_c \circ f_c(x_c),\ p_t \circ f_t(x_t)\rangle$, where ∘ is the function composition.

Zero-shot structure-text retrieval

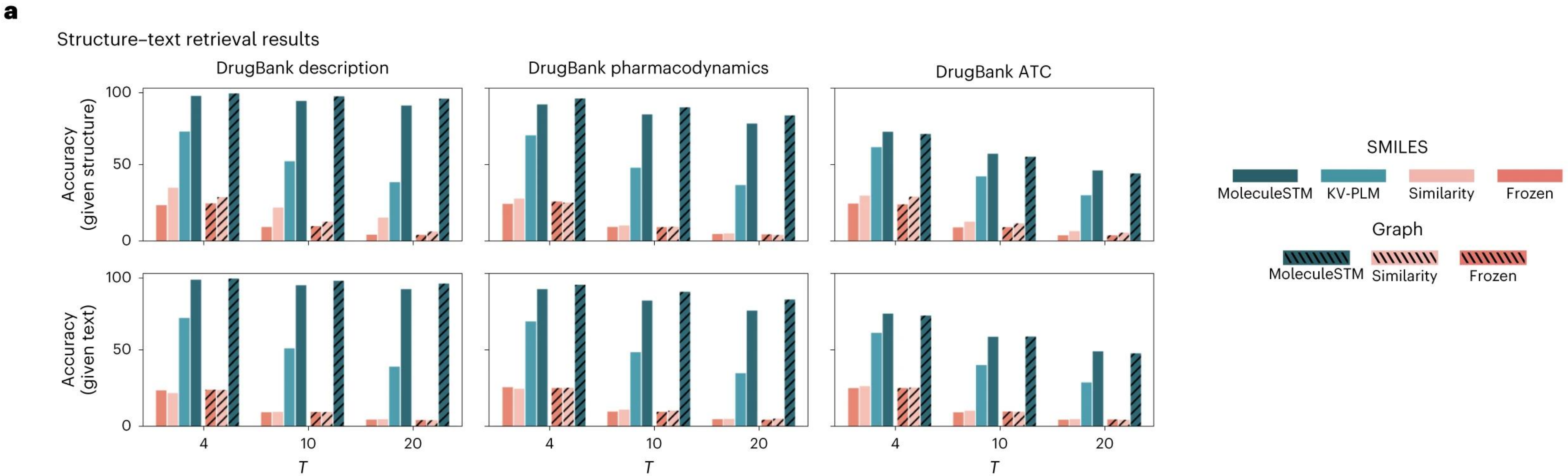
- Given a chemical structure and T textual descriptions, the retrieval task is to select the textual description with the highest similarity to the chemical structure (or vice versa) based on a score calculated on the joint representation space.

$$\text{Retrieval}(\mathbf{x}_c) = \arg\max_{\tilde{\mathbf{x}}_t} \{\langle p_c \circ f_c(\mathbf{x}_c), p_t \circ f_t(\tilde{\mathbf{x}}_t)\rangle | \tilde{\mathbf{x}}_t \in T \text{ textual descriptions}\},$$



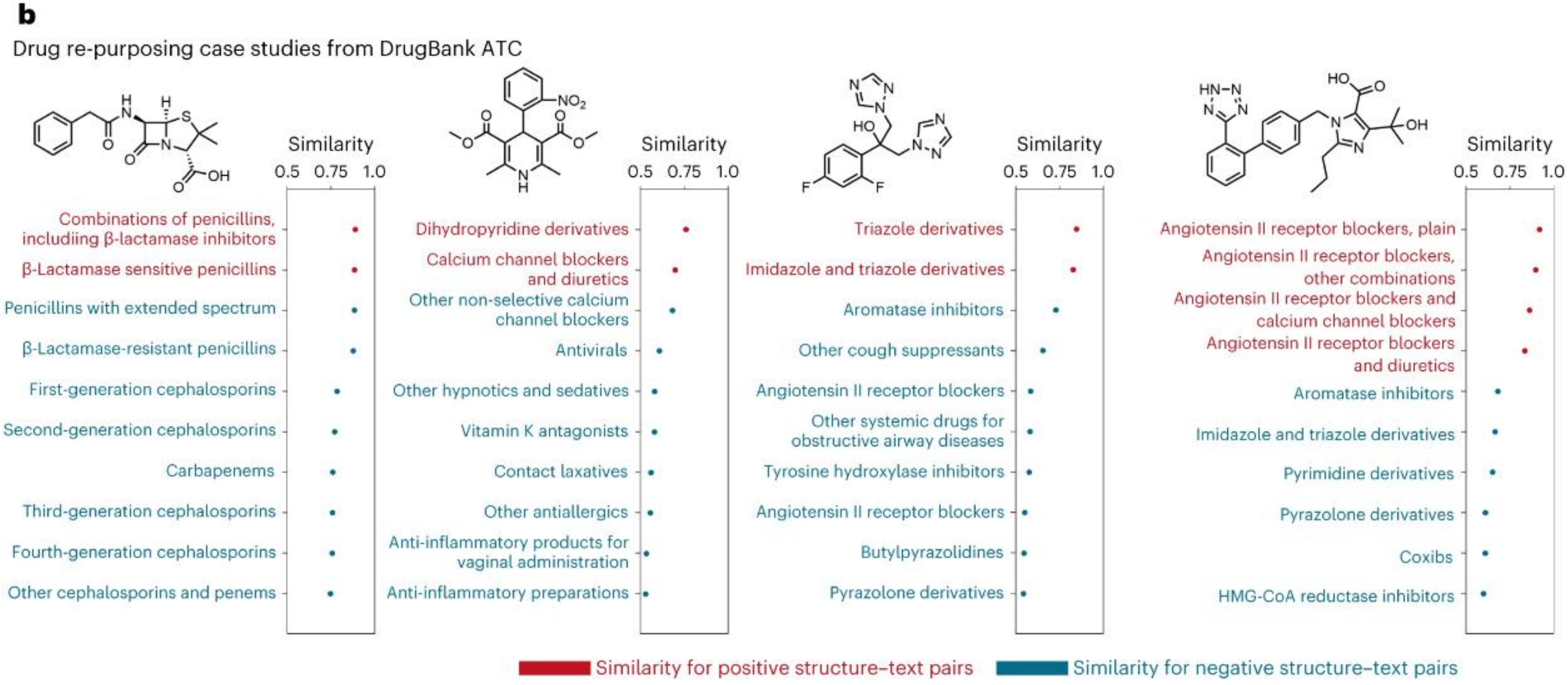
Zero-shot structure-text retrieval

- ❖ SMILES/Graph (MoleculeSTM, Similarity, Frozen) + KV-PLM
  - ✓ 'Frozen': pretrained encoders for the two branches + two randomly initialized projectors.
  - ✓ 'Similarity': use the similarity from a single branch;
    - when given a chemical structure, retrieve the most similar chemical structure from PubChemSTM, take the corresponding paired text representation as the proxy representation, and calculate the similarity between the proxy and T requested text representations.



Zero-shot structure-text retrieval

- Specifically, given the molecule's chemical structure, we take the 10 (out of 600) most similar ATC labels. It is observed that MoleculeSTM can retrieve the ground-truth ATC labels with high rankings.



Zero-shot structure-text retrieval

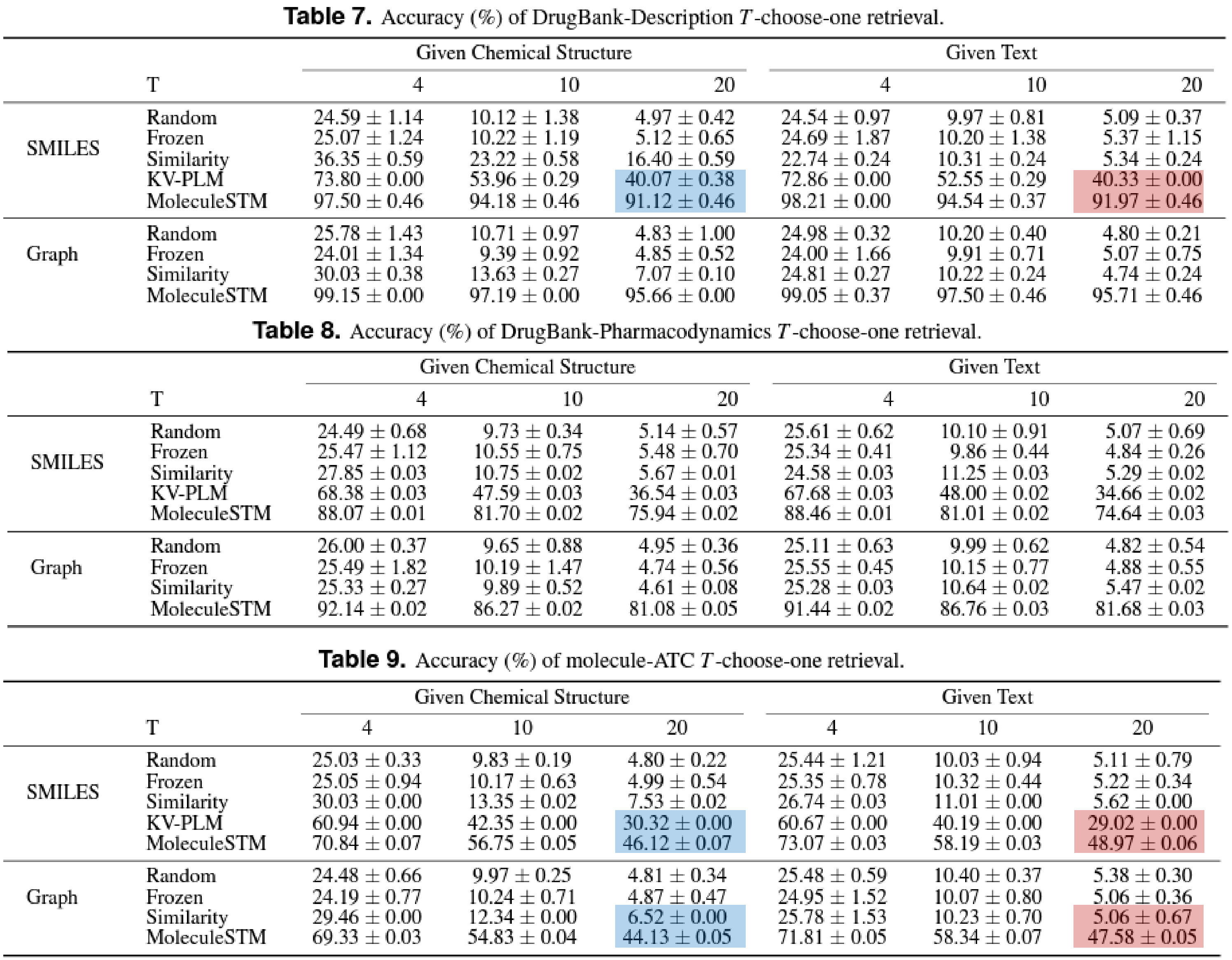


Table 7. Accuracy (%) of DrugBank-Description T-choose-one retrieval.

		Given Chemical Structure			Given Text		
	T	4	10	20	4	10	20
SMILES	Random	24.59 ± 1.14	10.12 ± 1.38	4.97 ± 0.42	24.54 ± 0.97	9.97 ± 0.81	5.09 ± 0.37
	Frozen	25.07 ± 1.24	10.22 ± 1.19	5.12 ± 0.65	24.69 ± 1.87	10.20 ± 1.38	5.37 ± 1.15
	Similarity	36.35 ± 0.59	23.22 ± 0.58	16.40 ± 0.59	22.74 ± 0.24	10.31 ± 0.24	5.34 ± 0.24
	KV-PLM	73.80 ± 0.00	53.96 ± 0.29	40.07 ± 0.38	72.86 ± 0.00	52.55 ± 0.29	40.33 ± 0.00
	MoleculeSTM	97.50 ± 0.46	94.18 ± 0.46	91.12 ± 0.46	98.21 ± 0.00	94.54 ± 0.37	91.97 ± 0.46
Graph	Random	25.78 ± 1.43	10.71 ± 0.97	4.83 ± 1.00	24.98 ± 0.32	10.20 ± 0.40	4.80 ± 0.21
	Frozen	24.01 ± 1.34	9.39 ± 0.92	4.85 ± 0.52	24.00 ± 1.66	9.91 ± 0.71	5.07 ± 0.75
	Similarity	30.03 ± 0.38	13.63 ± 0.27	7.07 ± 0.10	24.81 ± 0.27	10.22 ± 0.24	4.74 ± 0.24
	MoleculeSTM	99.15 ± 0.00	97.19 ± 0.00	95.66 ± 0.00	99.05 ± 0.37	97.50 ± 0.46	95.71 ± 0.46

Table 8. Accuracy (%) of DrugBank-Pharmacodynamics T-choose-one retrieval.

		Given Chemical Structure			Given Text		
	T	4	10	20	4	10	20
SMILES	Random	24.49 ± 0.68	9.73 ± 0.34	5.14 ± 0.57	25.61 ± 0.62	10.10 ± 0.91	5.07 ± 0.69
	Frozen	25.47 ± 1.12	10.55 ± 0.75	5.48 ± 0.70	25.34 ± 0.41	9.86 ± 0.44	4.84 ± 0.26
	Similarity	27.85 ± 0.03	10.75 ± 0.02	5.67 ± 0.01	24.58 ± 0.03	11.25 ± 0.03	5.29 ± 0.02
	KV-PLM	68.38 ± 0.03	47.59 ± 0.03	36.54 ± 0.03	67.68 ± 0.03	48.00 ± 0.02	34.66 ± 0.02
	MoleculeSTM	88.07 ± 0.01	81.70 ± 0.02	75.94 ± 0.02	88.46 ± 0.01	81.01 ± 0.02	74.64 ± 0.03
Graph	Random	26.00 ± 0.37	9.65 ± 0.88	4.95 ± 0.36	25.11 ± 0.63	9.99 ± 0.62	4.82 ± 0.54
	Frozen	25.49 ± 1.82	10.19 ± 1.47	4.74 ± 0.56	25.55 ± 0.45	10.15 ± 0.77	4.88 ± 0.55
	Similarity	25.33 ± 0.27	9.89 ± 0.52	4.61 ± 0.08	25.28 ± 0.03	10.64 ± 0.02	5.47 ± 0.02
	MoleculeSTM	92.14 ± 0.02	86.27 ± 0.02	81.08 ± 0.05	91.44 ± 0.02	86.76 ± 0.03	81.68 ± 0.03

Table 9. Accuracy (%) of molecule-ATC T-choose-one retrieval.

		Given Chemical Structure			Given Text		
	T	4	10	20	4	10	20
SMILES	Random	25.03 ± 0.33	9.83 ± 0.19	4.80 ± 0.22	25.44 ± 1.21	10.03 ± 0.94	5.11 ± 0.79
	Frozen	25.05 ± 0.94	10.17 ± 0.63	4.99 ± 0.54	25.35 ± 0.78	10.32 ± 0.44	5.22 ± 0.34
	Similarity	30.03 ± 0.00	13.35 ± 0.02	7.53 ± 0.02	26.74 ± 0.03	11.01 ± 0.00	5.62 ± 0.00
	KV-PLM	60.94 ± 0.00	42.35 ± 0.00	30.32 ± 0.00	60.67 ± 0.00	40.19 ± 0.00	29.02 ± 0.00
	MoleculeSTM	70.84 ± 0.07	56.75 ± 0.05	46.12 ± 0.07	73.07 ± 0.03	58.19 ± 0.03	48.97 ± 0.06
Graph	Random	24.48 ± 0.66	9.97 ± 0.25	4.81 ± 0.34	25.48 ± 0.59	10.40 ± 0.37	5.38 ± 0.30
	Frozen	24.19 ± 0.77	10.24 ± 0.71	4.87 ± 0.47	24.95 ± 1.52	10.07 ± 0.80	5.06 ± 0.36
	Similarity	29.46 ± 0.00	12.34 ± 0.00	6.52 ± 0.00	25.78 ± 1.53	10.23 ± 0.70	5.06 ± 0.67
	MoleculeSTM	69.33 ± 0.03	54.83 ± 0.04	44.13 ± 0.05	71.81 ± 0.05	58.34 ± 0.07	47.58 ± 0.05

Zero-shot text-based molecule editing

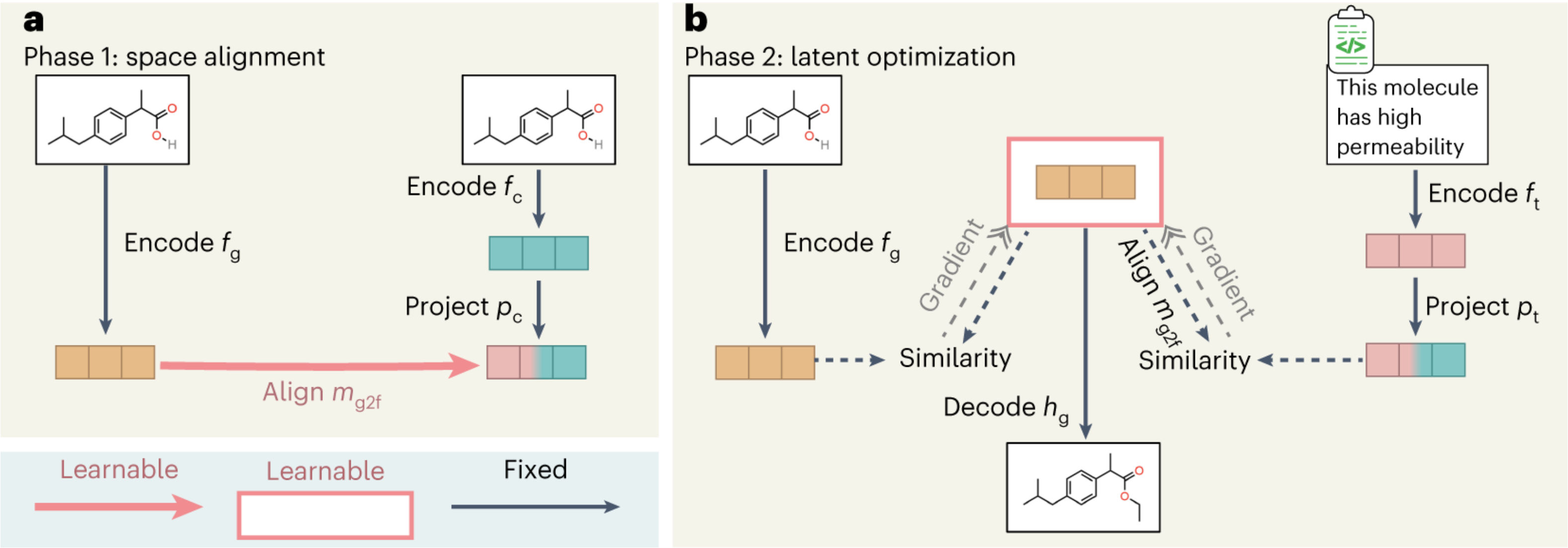
- ❖ For molecule editing, we randomly sample 200 molecules from ZINC and a text prompt as the inputs.
  - ✓ (1) Single-objective editing
    - using the single drug-related property for editing, such as ‘molecule with high solubility’ and ‘molecule more like a drug’.
  - ✓ (2) Multi-objective (compositionality) editing
    - applying multiple properties simultaneously, such as ‘molecule with high solubility and high permeability’.
  - ✓ (3) Binding-affinity-based editing
  - ✓ (4) Drug-relevance editing

Zero-shot text-based molecule editing

- For molecule editing, we randomly sample 200 molecules from ZINC and a text prompt as the inputs.
  - ✓ (3) Binding-affinity-based editing
    - assay description, where each assay corresponds to one binding-affinity task.
    - A concrete example is ChEMBL 1613777 with prompt as ‘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.’. ~ The output molecules should have higher binding-affinity scores.
  - ✓ (4) Drug-relevance editing
    - to make molecules structurally similar to certain common drugs, for example, ‘this molecule looks like penicillin’, expecting the output molecules to be more similar to the target drug than the input drug.

Zero-shot text-based molecule editing

- Given a fixed pretrained molecule generative model (encoder f_g and decoder h_g), the ML editing methods learn a semantically meaningful direction on the latent representation (or latent code) space.
  - ✓ The decoder h_g then generates output molecules with the desired properties by moving along the direction.



Zero-shot text-based molecule editing

- The first phase is space alignment,
  - ✓ where we train an adaptor module to align the representation space of the generative model to the joint representation space of MoleculeSTM.

$$\mathscr{L} = \| m_{\text{g2f}} \circ f_{\text{g}}(\mathbf{x}_{\text{c}}) - p_{\text{c}} \circ f_{\text{c}}(\mathbf{x}_{\text{c}}) \|^2$$

  - ✓ m_{g2f} is the adaptor module optimized to align the two latent spaces.
- The second phase is latent optimization,
  - ✓ where we directly learn the latent code using two similarity scores as the objective function.

$$w = \underset{w \in W}{\operatorname{argmin}} \left(-\mathscr{L}_{\text{cosine-sim}}\left(m_{\text{g2f}}(w), p_{\text{t}} \circ f_{\text{t}}(\mathbf{x}_{\text{t}})\right) + \lambda \cdot \mathscr{L}_{l_2}\left(w, f_{\text{g}}(\mathbf{x}_{\text{c,in}})\right)\right)$$

Latent code space

Text prompt

Input molecule

$$\mathbf{x}_{\text{c,out}} = h_{\text{g}}(w)$$

Zero-shot text-based molecule editing

- ❖ The evaluation metric is the satisfactory hit ratio.
  - ✓ Suppose we have an input molecule $x_{c,in}$ and a text prompt x_t, the editing algorithm will generate an output molecule $x_{c,out}$.
  - ✓ Then we use the hit ratio to measure if the output molecule can satisfy the conditions as indicated in the text prompt.

$$\mathrm{hit}(\mathbf{x}_{\mathrm{c,In}}, \mathbf{x}_{\mathrm{t}}) = \begin{cases} 1, & \exists \lambda,\ \text{s.t.}\ \mathbf{x}_{\mathrm{c,out}} = h_{\mathrm{g}}(\mathrm{w}; \lambda) \wedge \mathrm{satisfy}(\mathbf{x}_{\mathrm{c,In}}, \mathbf{x}_{\mathrm{c,out}}, \mathbf{x}_{\mathrm{t}}) \\ 0, & \text{otherwise} \end{cases},$$

$$\mathrm{hit}\,(t) = \frac{\sum_{i=1}^{N} \mathrm{hit}(\mathbf{x}^{i}_{\mathrm{c,In}}, \mathbf{x}_{\mathrm{t}})}{N},$$

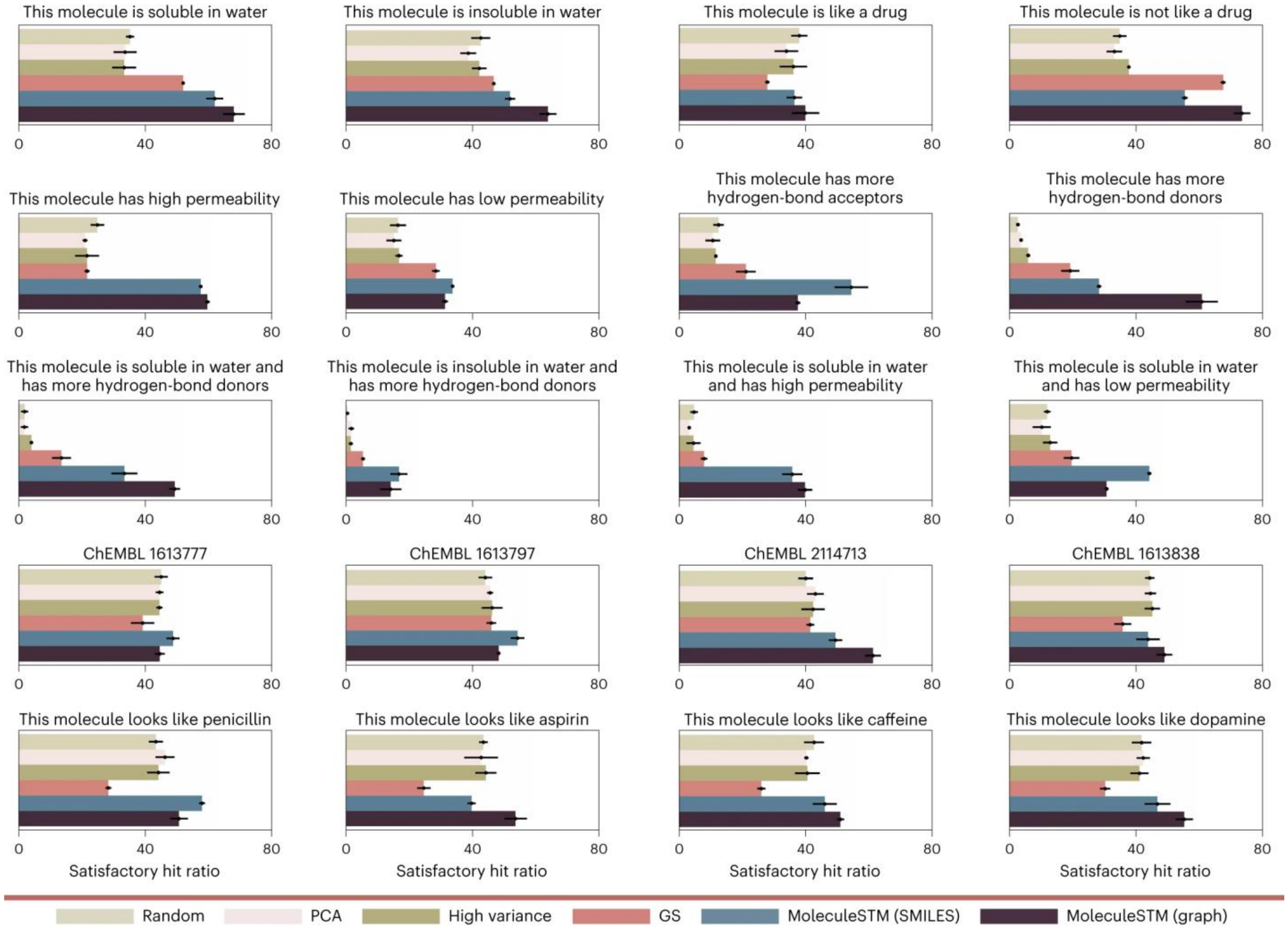
- For single-objective property-based editing, we use the logarithm of partition coefficient (logP), quantitative estimate of drug-likeness (QED) and topological polar surface area (tPSA) as the proxies to measure the molecule solubility, drug likeness and permeability.

Zero-shot text-based molecule editing

- ❖ Four baseline models are tested with MoleculeSTM
  - ✓ (1) Random: take a random noise as the perturbation to the representation of input molecules.
  - ✓ (2) Principal Component Analysis (PCA): take the eigenvectors as latent directions, where the eigenvectors are obtained after decomposing the latent representation of input molecules.
  - ✓ (3) High variance: take the latent representation dimension with the highest variance and apply the one-hot encoding on it as a semantic direction for editing.
  - ✓ (4) Genetic Search (GS): does a random search instead of a guided search by a reward function as no retrieval database is available in the zero-shot setting.

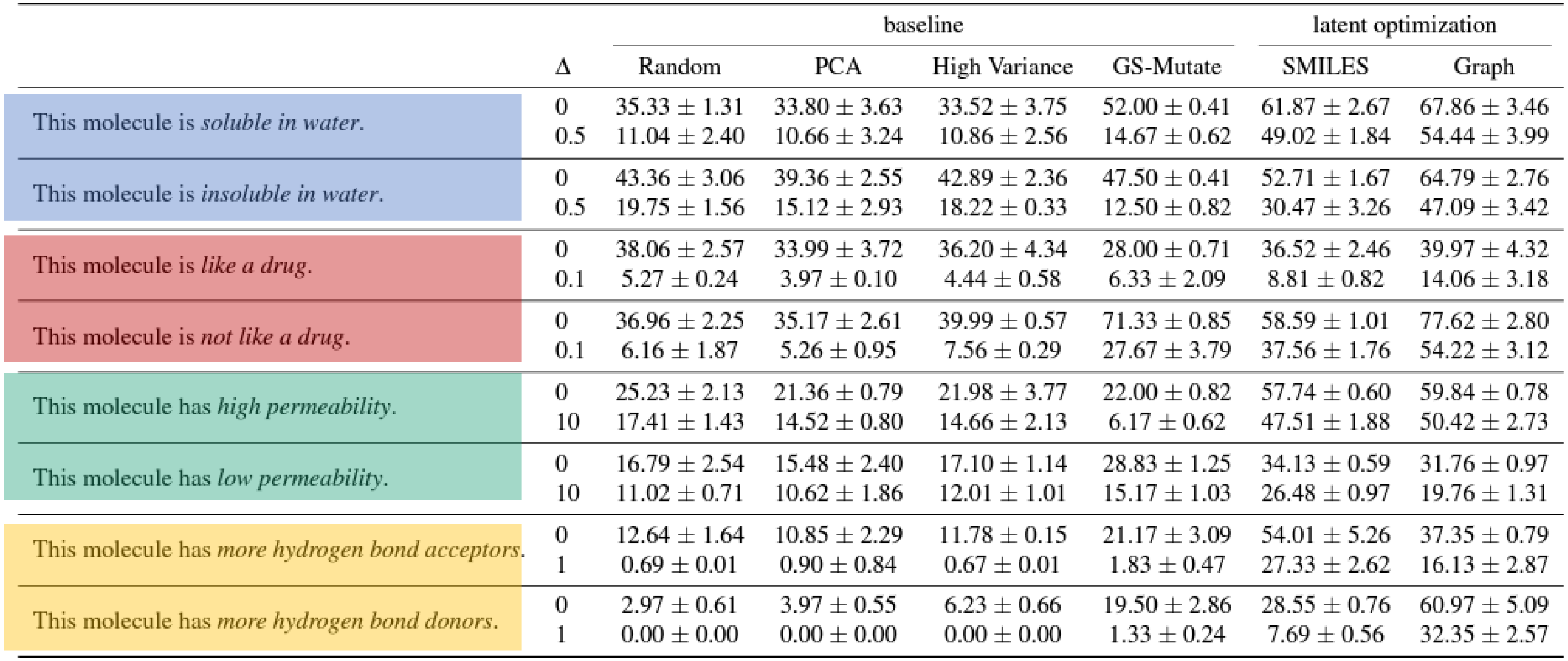
Zero-shot text-based molecule editing

- Satisfactory hit ratios (%) of four types of text-based editing task



Zero-shot text-based molecule editing

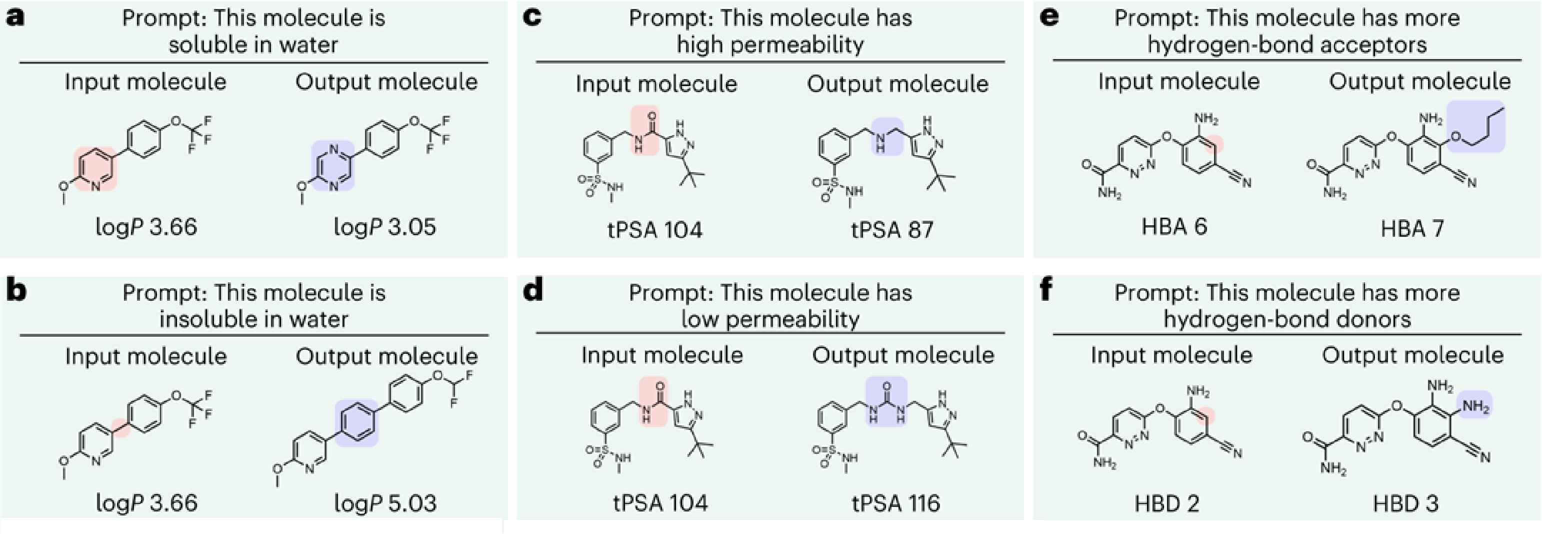
- ❖ Results on eight single-objective molecule editing
  - ✓ For ∆, it is the threshold that only difference above it can be viewed as a hit.
    - So the larger ∆ means a stricter editing criterion.



			baseline				latent optimization	
		∆	Random	PCA	High Variance	GS-Mutate	SMILES	Graph
LogP	This molecule is *soluble in water*.	0	35.33 ± 1.31	33.80 ± 3.63	33.52 ± 3.75	52.00 ± 0.41	61.87 ± 2.67	67.86 ± 3.46
		0.5	11.04 ± 2.40	10.66 ± 3.24	10.86 ± 2.56	14.67 ± 0.62	49.02 ± 1.84	54.44 ± 3.99
	This molecule is *insoluble in water*.	0	43.36 ± 3.06	39.36 ± 2.55	42.89 ± 2.36	47.50 ± 0.41	52.71 ± 1.67	64.79 ± 2.76
		0.5	19.75 ± 1.56	15.12 ± 2.93	18.22 ± 0.33	12.50 ± 0.82	30.47 ± 3.26	47.09 ± 3.42
QED	This molecule is *like a drug*.	0	38.06 ± 2.57	33.99 ± 3.72	36.20 ± 4.34	28.00 ± 0.71	36.52 ± 2.46	39.97 ± 4.32
		0.1	5.27 ± 0.24	3.97 ± 0.10	4.44 ± 0.58	6.33 ± 2.09	8.81 ± 0.82	14.06 ± 3.18
	This molecule is *not like a drug*.	0	36.96 ± 2.25	35.17 ± 2.61	39.99 ± 0.57	71.33 ± 0.85	58.59 ± 1.01	77.62 ± 2.80
		0.1	6.16 ± 1.87	5.26 ± 0.95	7.56 ± 0.29	27.67 ± 3.79	37.56 ± 1.76	54.22 ± 3.12
tPSA	This molecule has *high permeability*.	0	25.23 ± 2.13	21.36 ± 0.79	21.98 ± 3.77	22.00 ± 0.82	57.74 ± 0.60	59.84 ± 0.78
		10	17.41 ± 1.43	14.52 ± 0.80	14.66 ± 2.13	6.17 ± 0.62	47.51 ± 1.88	50.42 ± 2.73
	This molecule has *low permeability*.	0	16.79 ± 2.54	15.48 ± 2.40	17.10 ± 1.14	28.83 ± 1.25	34.13 ± 0.59	31.76 ± 0.97
		10	11.02 ± 0.71	10.62 ± 1.86	12.01 ± 1.01	15.17 ± 1.03	26.48 ± 0.97	19.76 ± 1.31
HBA/ HBD	This molecule has *more hydrogen bond acceptors*.	0	12.64 ± 1.64	10.85 ± 2.29	11.78 ± 0.15	21.17 ± 3.09	54.01 ± 5.26	37.35 ± 0.79
		1	0.69 ± 0.01	0.90 ± 0.84	0.67 ± 0.01	1.83 ± 0.47	27.33 ± 2.62	16.13 ± 2.87
	This molecule has *more hydrogen bond donors*.	0	2.97 ± 0.61	3.97 ± 0.55	6.23 ± 0.66	19.50 ± 2.86	28.55 ± 0.76	60.97 ± 5.09
		1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.33 ± 0.24	7.69 ± 0.56	32.35 ± 2.57

Zero-shot text-based molecule editing

- ❖ Visual analysis on single objective molecule editing.
  - ✓ The difference between input and output molecules using the single-objective property ~ the addition, removal and replacement of functional groups or cores of the molecules.


a
Prompt: This molecule is soluble in water
Input molecule
Output molecule
logP 3.66
logP 3.05
b
Prompt: This molecule is insoluble in water
Input molecule
Output molecule
logP 3.66
logP 5.03
c
Prompt: This molecule has high permeability
Input molecule
Output molecule
tPSA 104
tPSA 87
d
Prompt: This molecule has low permeability
Input molecule
Output molecule
tPSA 104
tPSA 116
e
Prompt: This molecule has more hydrogen-bond acceptors
Input molecule
Output molecule
HBA 6
HBA 7
f
Prompt: This molecule has more hydrogen-bond donors
Input molecule
Output molecule
HBD 2
HBD 3

↓

Replacement of pyridine to a pyrazine core improves the solubility, while insertion of a benzene linkage yields an insoluble molecule.

↓

changing an amide linkage to an alkyl amine and a urea results in higher and lower permeability of the edited molecules, respectively.

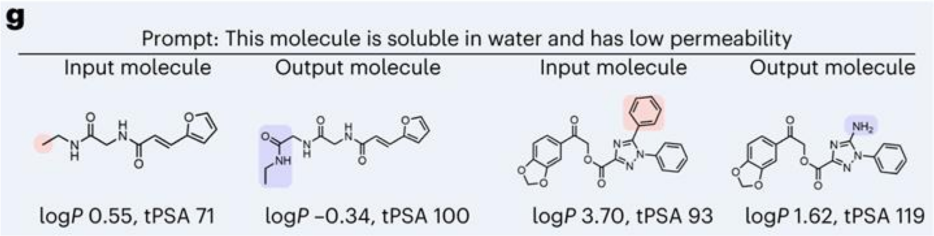
↓

a butyl addingether and a primary amine to the exact position of the molecule brings more hydrogen-bond acceptors and hydrogen-bond donors

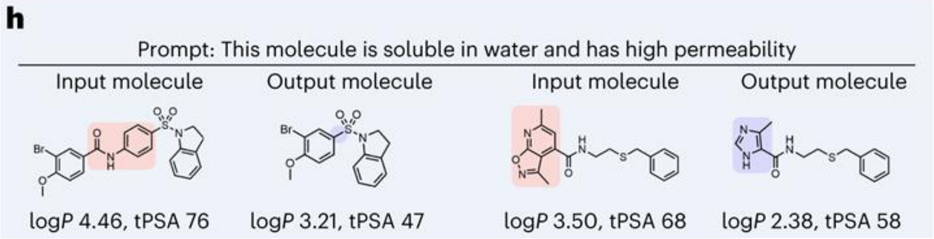
Zero-shot text-based molecule editing

❖ Visual analysis on multi-objective molecule editing

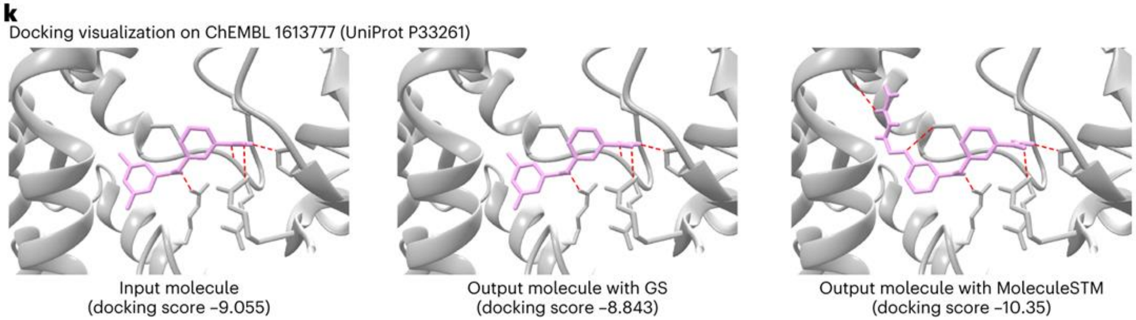
(g) Water solubility improvement and permeability reduction are consistent when introducing polar groups to the molecule and removing lipophilic hydrocarbons, such as an amide or primary amine replacing a methyl or phenyl.
↓



→(h) an amide and a benzene linkage are both removed in the left case, and a [1,2]oxazolo[5,4-b]pyridine substituent is replaced by a water-soluble imidazole with a smaller polar surface in the right case. ↓

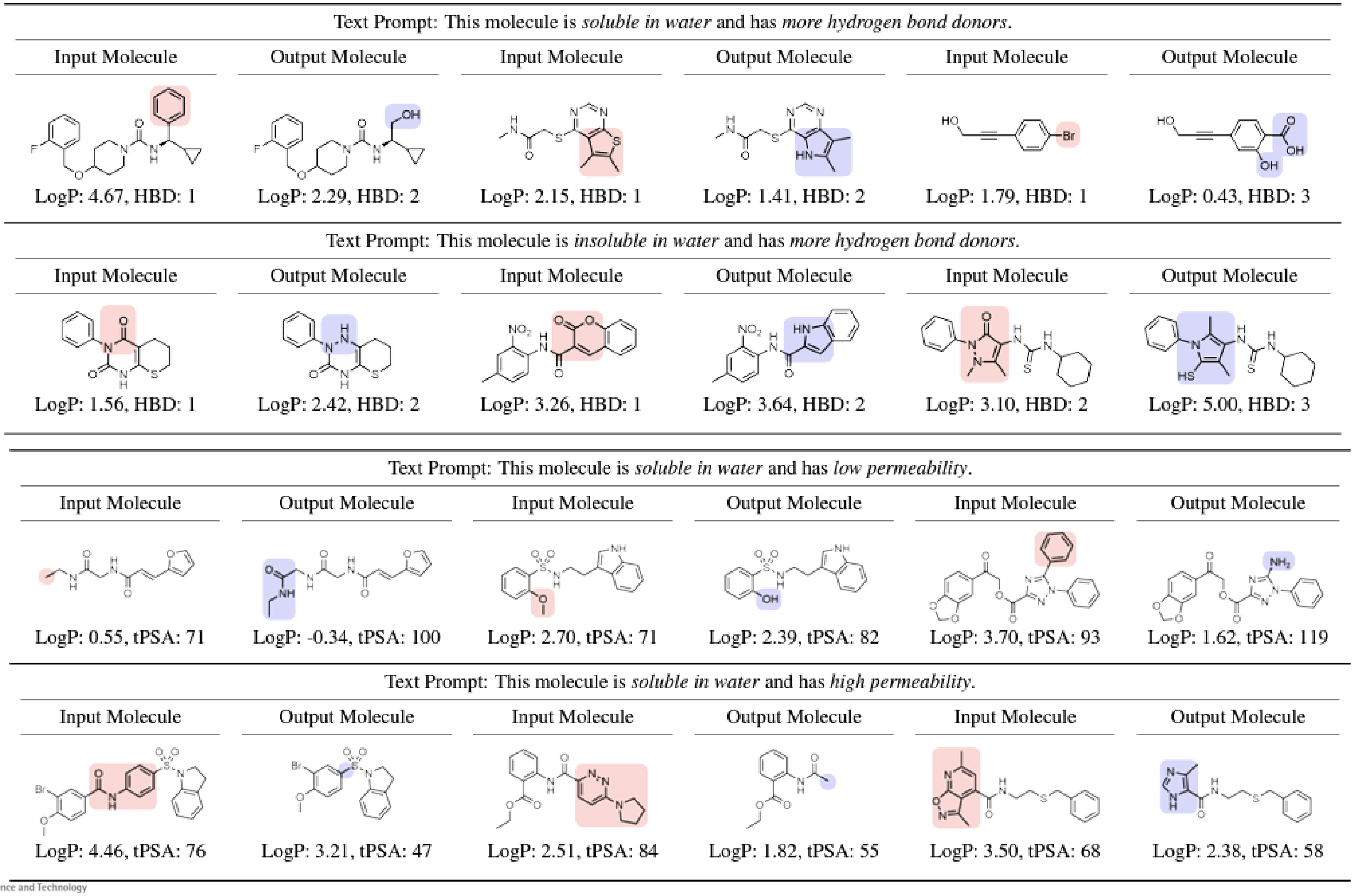


Binding-affinity-based editing.
The dashed red lines mark the potential bindings. →



Zero-shot text-based molecule editing

- ❖ Visualization of text-based editing on multi-objective (compositionality) properties



Text Prompt: This molecule is *soluble in water* and has *more hydrogen bond donors*.

Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
LogP: 4.67, HBD: 1	LogP: 2.29, HBD: 2	LogP: 2.15, HBD: 1	LogP: 1.41, HBD: 2	LogP: 1.79, HBD: 1	LogP: 0.43, HBD: 3

Text Prompt: This molecule is *insoluble in water* and has *more hydrogen bond donors*.

Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
LogP: 1.56, HBD: 1	LogP: 2.42, HBD: 2	LogP: 3.26, HBD: 1	LogP: 3.64, HBD: 2	LogP: 3.10, HBD: 2	LogP: 5.00, HBD: 3

Text Prompt: This molecule is *soluble in water* and has *low permeability*.

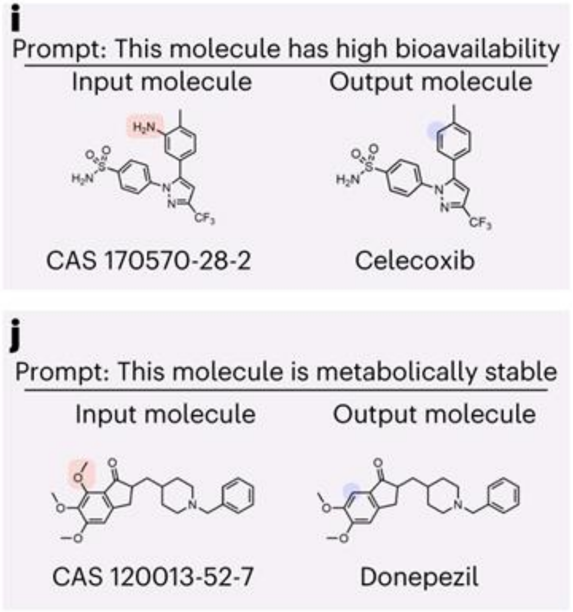
Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
LogP: 0.55, tPSA: 71	LogP: -0.34, tPSA: 100	LogP: 2.70, tPSA: 71	LogP: 2.39, tPSA: 82	LogP: 3.70, tPSA: 93	LogP: 1.62, tPSA: 119

Text Prompt: This molecule is *soluble in water* and has *high permeability*.

Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
LogP: 4.46, tPSA: 76	LogP: 3.21, tPSA: 47	LogP: 2.51, tPSA: 84	LogP: 1.82, tPSA: 55	LogP: 3.50, tPSA: 68	LogP: 2.38, tPSA: 58

Zero-shot text-based molecule editing

- Case studies on neighbourhood searching for patent drug molecules
  - ✓ Herein we demonstrate two examples of generating approved drugs from their patented analogues by addressing their property deficiencies based on text prompts.



Neighbourhood searching for patent data

→ Figure 5i generates celecoxib from its amino-substituted derivative, where the removal of the amino group yields a greater intestinal permeability of the molecule leading to higher bioavailability.

→ In Fig. 5j, the trimethoxy benzene moiety, an electron-rich arene known to undergo oxidative phase I metabolisms, is replaced by a dimethoxy arene in donepezil by calling for a metabolically stable molecule.

Molecular property prediction

- ❖ One advantage of MoleculeSTM is that the pretrained chemical structure representation shares information with the external domain knowledge, which can be beneficial for the property prediction tasks.
  - ✓ We consider two types of chemical structure, the SMILES string and the molecular graph.
    - For the SMILES string, the randomly initialized models and two pretrained language models (MegaMolBART and KV-PLM).
    - For the molecular graph, the random initialization, + pretraining (AttrMasking, ContextPred, InfoGraph, MolCLR and GraphMVP).

Molecular property prediction

- For modelling, we take the pretrained encoder f_c and add a prediction head h_c to predict a categorical-valued or scalar-valued molecular property such as binding affinity or toxicity.
  - ✓ Both f_c and h_c are optimized to fit the target property, that is, in a fine-tuning manner

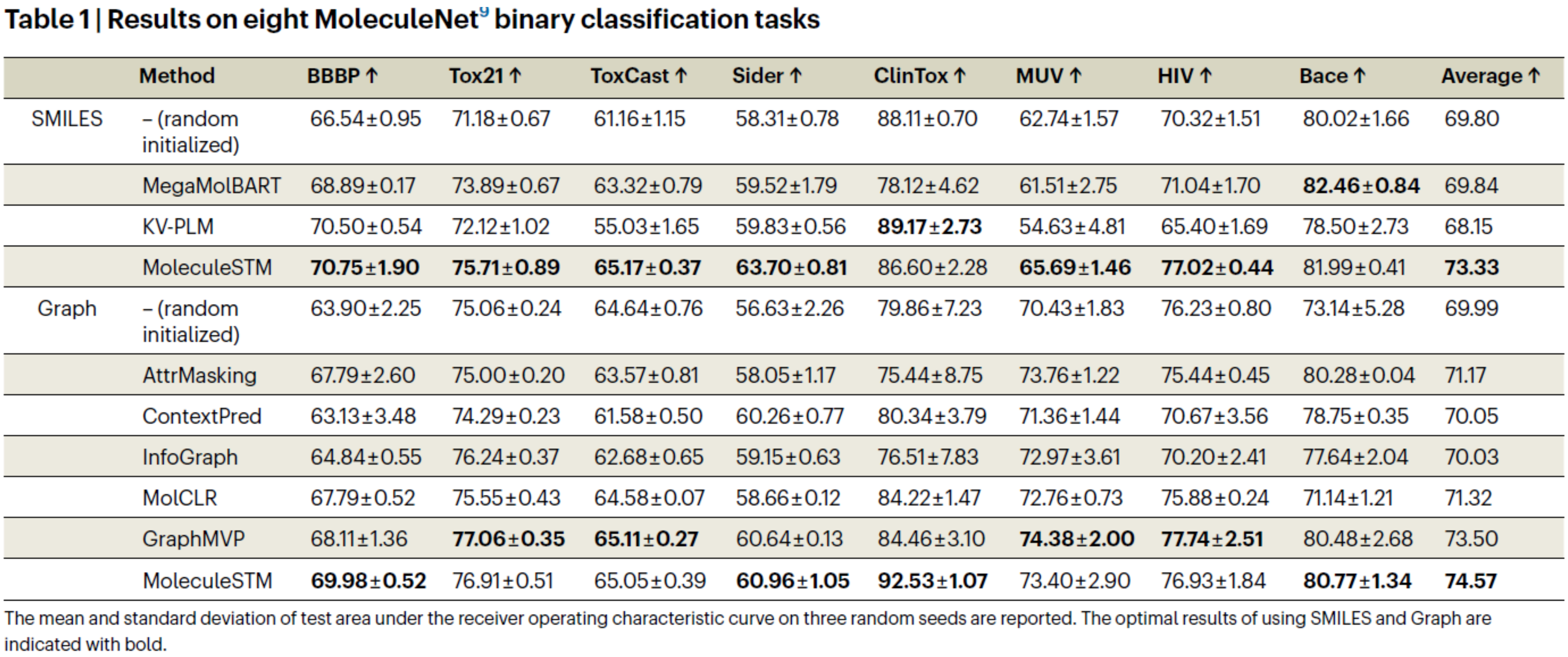


Table 1 | Results on eight MoleculeNet[9] binary classification tasks

	Method	BBBP ↑	Tox21 ↑	ToxCast ↑	Sider ↑	ClinTox ↑	MUV ↑	HIV ↑	Bace ↑	Average ↑
SMILES	– (random initialized)	66.54±0.95	71.18±0.67	61.16±1.15	58.31±0.78	88.11±0.70	62.74±1.57	70.32±1.51	80.02±1.66	69.80
	MegaMolBART	68.89±0.17	73.89±0.67	63.32±0.79	59.52±1.79	78.12±4.62	61.51±2.75	71.04±1.70	**82.46±0.84**	69.84
	KV-PLM	70.50±0.54	72.12±1.02	55.03±1.65	59.83±0.56	**89.17±2.73**	54.63±4.81	65.40±1.69	78.50±2.73	68.15
	MoleculeSTM	**70.75±1.90**	**75.71±0.89**	**65.17±0.37**	**63.70±0.81**	86.60±2.28	**65.69±1.46**	**77.02±0.44**	81.99±0.41	**73.33**
Graph	– (random initialized)	63.90±2.25	75.06±0.24	64.64±0.76	56.63±2.26	79.86±7.23	70.43±1.83	76.23±0.80	73.14±5.28	69.99
	AttrMasking	67.79±2.60	75.00±0.20	63.57±0.81	58.05±1.17	75.44±8.75	73.76±1.22	75.44±0.45	80.28±0.04	71.17
	ContextPred	63.13±3.48	74.29±0.23	61.58±0.50	60.26±0.77	80.34±3.79	71.36±1.44	70.67±3.56	78.75±0.35	70.05
	InfoGraph	64.84±0.55	76.24±0.37	62.68±0.65	59.15±0.63	76.51±7.83	72.97±3.61	70.20±2.41	77.64±2.04	70.03
	MolCLR	67.79±0.52	75.55±0.43	64.58±0.07	58.66±0.12	84.22±1.47	72.76±0.73	75.88±0.24	71.14±1.21	71.32
	GraphMVP	68.11±1.36	**77.06±0.35**	**65.11±0.27**	60.64±0.13	84.46±3.10	**74.38±2.00**	**77.74±2.51**	80.48±2.68	73.50
	MoleculeSTM	**69.98±0.52**	76.91±0.51	65.05±0.39	**60.96±1.05**	**92.53±1.07**	73.40±2.90	76.93±1.84	**80.77±1.34**	**74.57**

The mean and standard deviation of test area under the receiver operating characteristic curve on three random seeds are reported. The optimal results of using SMILES and Graph are indicated with bold.

Thank you

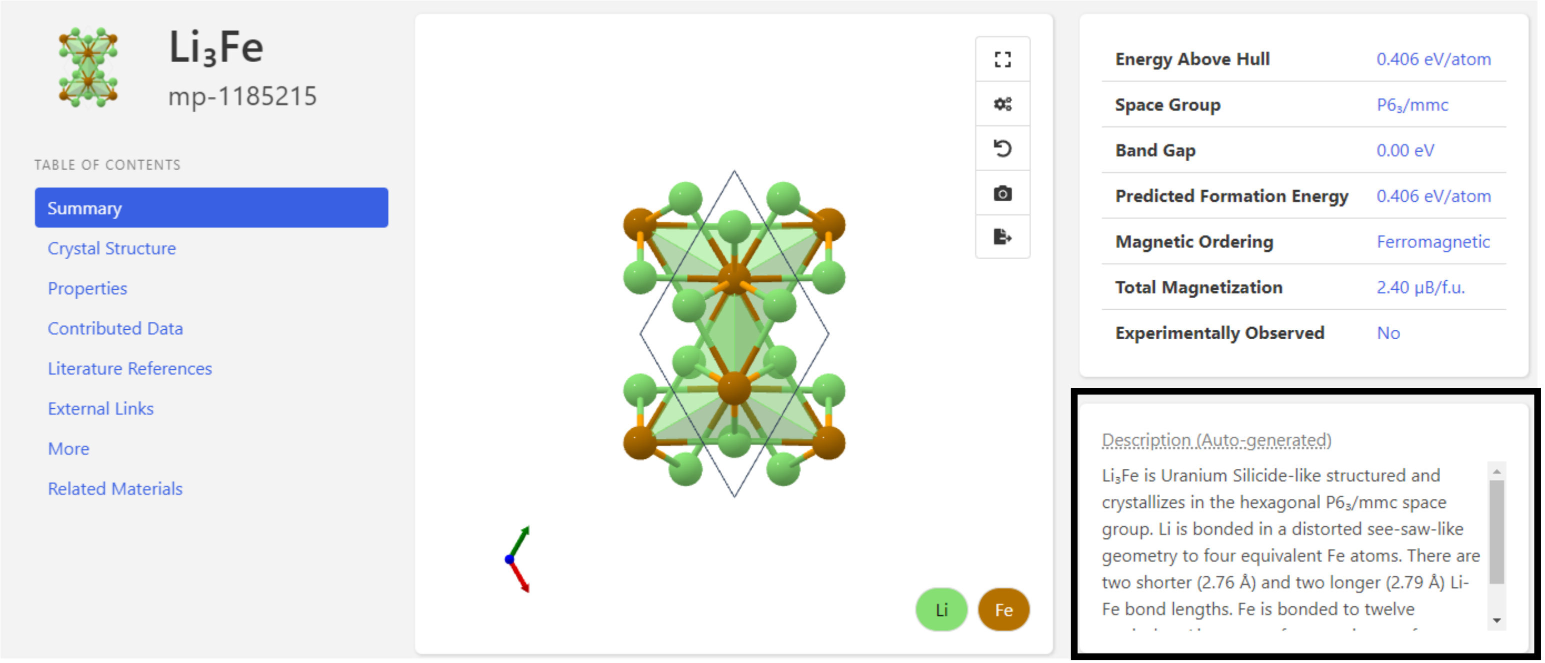
By *Dr.* ***Jaewoong Choi***

(Dr. B. Lee group @KIST)

2024.01.11

REF: text-to-property

- Robocrystallographer
  - ✓ The symmetry, local environment, and extended connectivity when generating a description.



Li_3Fe

mp-1185215

TABLE OF CONTENTS

Energy Above Hull	0.406 eV/atom
Space Group	$P6_3/mmc$
Band Gap	0.00 eV
Predicted Formation Energy	0.406 eV/atom
Magnetic Ordering	Ferromagnetic
Total Magnetization	2.40 μB/f.u.
Experimentally Observed	No

Description (Auto-generated)

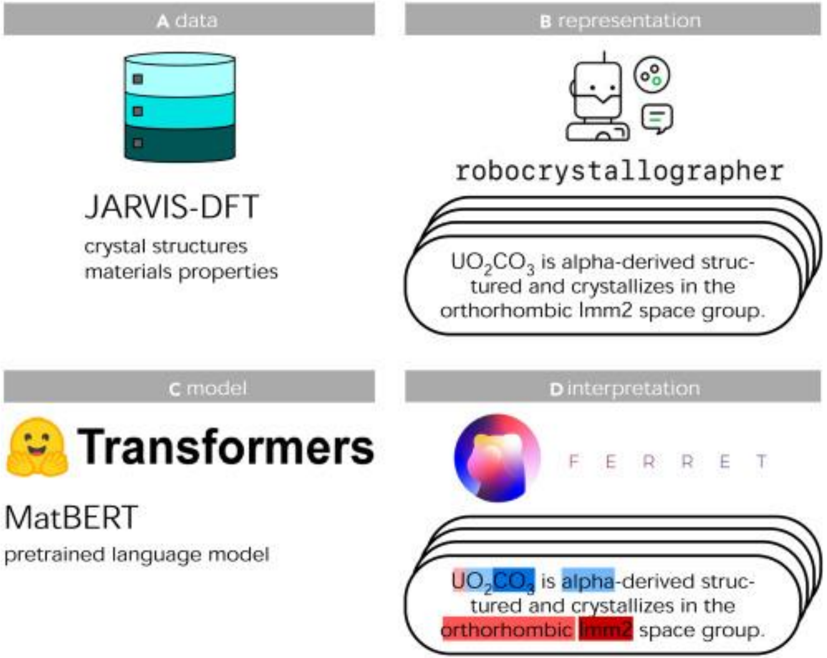
Li_3Fe is Uranium Silicide-like structured and crystallizes in the hexagonal $P6_3/mmc$ space group. Li is bonded in a distorted see-saw-like geometry to four equivalent Fe atoms. There are two shorter (2.76 Å) and two longer (2.79 Å) Li-Fe bond lengths. Fe is bonded to twelve

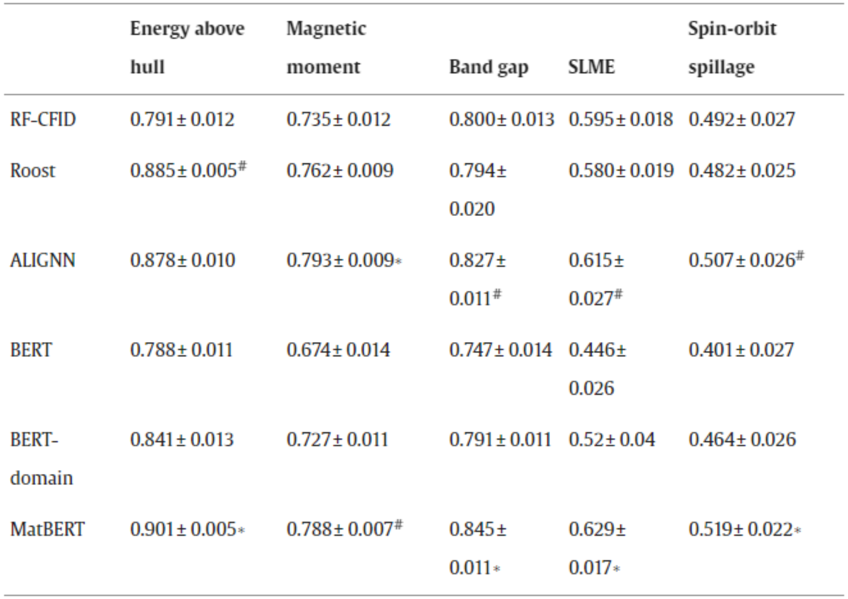
MRS Communications, *9*(3), 874-881.

Li_3Fe is Uranium Silicide-like structured and crystallizes in the hexagonal $P6_3/mmc$ space group. Li is bonded in a distorted see-saw-like geometry to four equivalent Fe atoms. There are two shorter (2.76 Å) and two longer (2.79 Å) Li-Fe bond lengths. Fe is bonded to twelve equivalent Li atoms to form a mixture of corner and face-sharing $FeLi_{12}$ cuboctahedra.

REF: text-to-property

- Paper #1: Text descriptions provide efficient materials representation for property prediction, overperforming graph neural networks
  - ✓ The authors conducted the classification on energy above hull, magnetic moment, band gap, SLME, spin-orbit spillage.


A data
JARVIS-DFT
crystal structures
materials properties
B representation
robocrystallographer
UO_2CO_3 is alpha-derived structured and crystallizes in the orthorhombic Imm2 space group.
C model
Transformers
MatBERT
pretrained language model
D interpretation
FERRET
UO_2CO_3 is alpha-derived structured and crystallizes in the orthorhombic Imm2 space group.

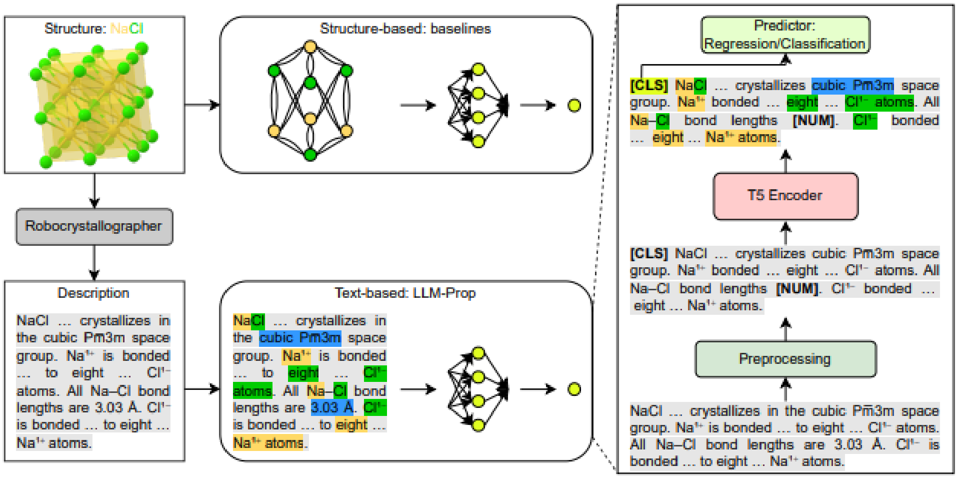


	Energy above hull	Magnetic moment	Band gap	SLME	Spin-orbit spillage
RF-CFID	0.791± 0.012	0.735± 0.012	0.800± 0.013	0.595± 0.018	0.492± 0.027
Roost	0.885± 0.005#	0.762± 0.009	0.794± 0.020	0.580± 0.019	0.482± 0.025
ALIGNN	0.878± 0.010	0.793± 0.009*	0.827± 0.011#	0.615± 0.027#	0.507± 0.026#
BERT	0.788± 0.011	0.674± 0.014	0.747± 0.014	0.446± 0.026	0.401± 0.027
BERT-domain	0.841± 0.013	0.727± 0.011	0.791± 0.011	0.52± 0.04	0.464± 0.026
MatBERT	0.901± 0.005*	0.788± 0.007#	0.845± 0.011*	0.629± 0.017*	0.519± 0.022*

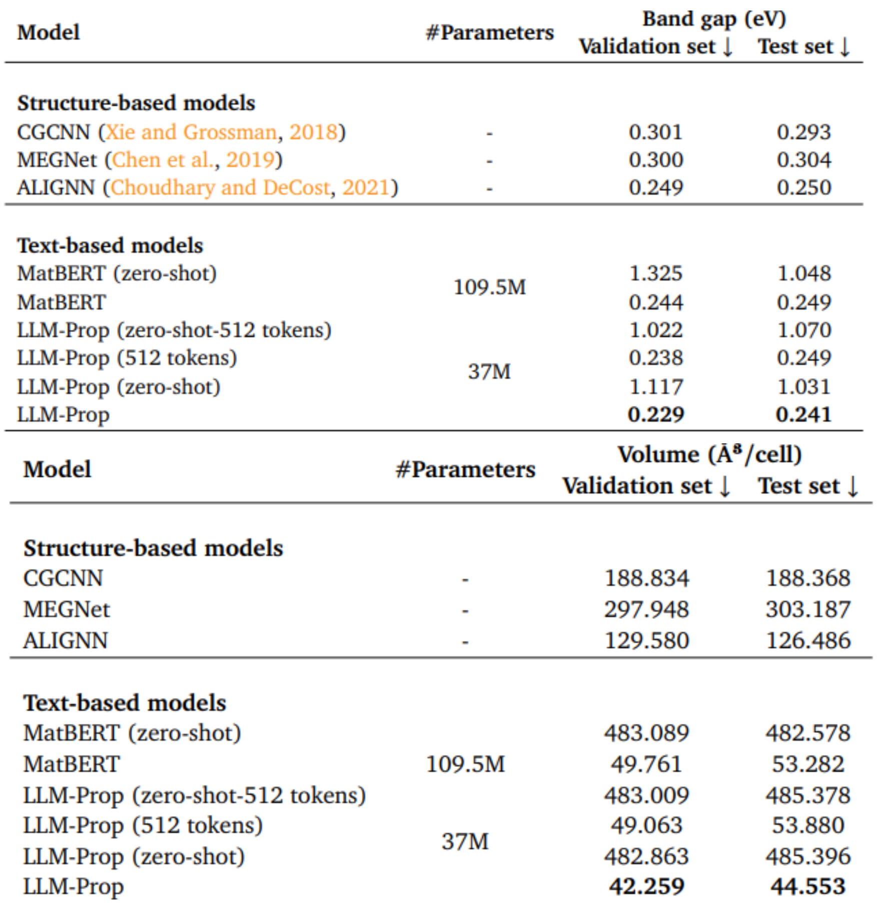
Patterns 4.10 (2023)

REF: text-to-property

- Paper #2: Recent study reported the use of LLMs outperforms GNN-based models in predicting bandgap and volume of 144,931 crystals


Structure: NaCl
Structure-based: baselines
Robocrystallographer
Description
NaCl … crystallizes in the cubic Pm̄3m space group. Na¹⁺ is bonded … to eight … Cl¹⁻ atoms. All Na–Cl bond lengths are 3.03 Å. Cl¹⁻ is bonded … to eight … Na¹⁺ atoms.
Text-based: LLM-Prop
NaCl … crystallizes in the cubic Pm̄3m space group. Na¹⁺ is bonded … to eight … Cl¹⁻ atoms. All Na–Cl bond lengths are 3.03 Å. Cl¹⁻ is bonded … to eight … Na¹⁺ atoms.
Predictor: Regression/Classification
[CLS] NaCl … crystallizes cubic Pm3m space group. Na¹⁺ bonded … eight … Cl¹⁻ atoms. All Na–Cl bond lengths [NUM]. Cl¹⁻ bonded … eight … Na¹⁺ atoms.
T5 Encoder
[CLS] NaCl … crystallizes cubic Pm3m space group. Na¹⁺ bonded … eight … Cl¹⁻ atoms. All Na–Cl bond lengths [NUM]. Cl¹⁻ bonded … eight … Na¹⁺ atoms.
Preprocessing
NaCl … crystallizes in the cubic Pm̄3m space group. Na¹⁺ is bonded … to eight … Cl¹⁻ atoms. All Na–Cl bond lengths are 3.03 Å. Cl¹⁻ is bonded … to eight … Na¹⁺ atoms.

"LLM-Prop: Predicting Physical And Electronic Properties Of Crystalline Solids From Their Text Descriptions"
arXiv:2310.14029v1



Model	#Parameters	Band gap (eV) Validation set ↓	Band gap (eV) Test set ↓
Structure-based models			
CGCNN (Xie and Grossman, 2018)	-	0.301	0.293
MEGNet (Chen et al., 2019)	-	0.300	0.304
ALIGNN (Choudhary and DeCost, 2021)	-	0.249	0.250
Text-based models			
MatBERT (zero-shot)	109.5M	1.325	1.048
MatBERT		0.244	0.249
LLM-Prop (zero-shot-512 tokens)	37M	1.022	1.070
LLM-Prop (512 tokens)		0.238	0.249
LLM-Prop (zero-shot)		1.117	1.031
LLM-Prop		**0.229**	**0.241**

Model	#Parameters	Volume ($Å^3$/cell) Validation set ↓	Volume ($Å^3$/cell) Test set ↓
Structure-based models			
CGCNN	-	188.834	188.368
MEGNet	-	297.948	303.187
ALIGNN	-	129.580	126.486
Text-based models			
MatBERT (zero-shot)	109.5M	483.089	482.578
MatBERT		49.761	53.282
LLM-Prop (zero-shot-512 tokens)	37M	483.009	485.378
LLM-Prop (512 tokens)		49.063	53.880
LLM-Prop (zero-shot)		482.863	485.396
LLM-Prop		**42.259**	**44.553**