

---

# Multi-modal molecule structure–text model for text-based retrieval and editing

---

In the format provided by the  
authors and unedited

---

## A Pretraining

### A.1 PubChemSTM Construction

We construct a chemical structure-text pair dataset called PubChemSTM, which is extracted from the PubChem database [1]. Below we explain the key steps of the dataset construction.

1. We use the [PUG View](#) (a REST-style web service) to download the textual descriptions of molecules. It has in total of 290 pages, and each page is downloaded in XML format. For reference, an example page (the first page) can be found [here](#). There is a "string" field in the XML data, and we treat it as the textual descriptions for molecules. After construction, we have 250K molecules (with unique PubChem ID) and 281K chemical structure-text pairs. Notice that each molecule can have multiple annotations from different resources.
  - Most of the molecule annotations start with the common name or the International Union of Pure and Applied Chemistry (IUPAC) name. We can either use the raw description (with a common name or IUPAC name) or replace it with the text template (e.g., "This molecule is ...").
  - Thus, we construct two versions of PubChemSTM datasets, PubChemSTM-raw and PubChemSTM-extracted, corresponding to using the raw annotation or replacing the molecule name with the text prompt, respectively. These two versions of PubChemSTM share the molecules, except for the molecule names.
2. We download the 326 SDF files from the PubChem [FTP service](#). Each SDF file contains the structural information (e.g., the SMILES string and molecular graph) for a batch of molecules.
3. We match the annotation and chemical structure for each molecule from the previous two steps using the PubChem ID, and most of the molecules from the first step contain the corresponding chemical structures from the SDF files. In specific, only 12 molecules failed to find the valid SMILES from SDF files, and we ignore these molecules.
4. Ultimately, following the above three steps will lead to a structure-text pair dataset with 281K pairs and 250K unique molecules. Note that the PubChem database [1] is updated online frequently, and the above numbers are collected in March 2022.

**Pre-processing Details** There is one field in the PubChem database called "name", which includes either the common name or the IUPAC name for each molecule. Notice that the tokenization on IUPAC is nontrivial. Thus we carry out two versions to test its effect, i.e., the PubChemSTM-raw and PubChemSTM-extracted. We find that there exist several patterns of textual descriptions in PubChemSTM-raw, which are further utilized to extract the cleaner version of molecule description as in PubChem-extract. A detailed illustration is given below:

- The most common pattern is that the molecule annotation starts with "XXX (name) is / are / was / were / appears / occurs / stands for / belongs to / exits ...". We manually extract this to obtain most of the molecule names and replace them with "This molecule ..." or "These molecules ...".
- **Extra word "Pure"**. Some molecule annotations start with "Pure xxx ..." and we remove the word "Pure".
- **Typos**. For example, the "Mercurycombines ..." should be "Mercury combines ...".

**Dataset Examples** We provide four examples of the PubChemSTM-raw and PubChemSTM-extracted in Table 1.

**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.	<i>This molecule is</i> 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.	<i>This molecule is</i> 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.	<i>This molecule appears</i> 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.	<i>This molecule is</i> 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.

**Reproducibility** Because the PubChem database [1] has been updated online frequently, so we provide all the pre-processed datasets used in this work for reproducibility. In addition, the source codes for the above steps are also provided for future usage.

**Comparison** As mentioned, we adopt a pretrained SciBERT model [2] and continue training on PubChemSTM. SciBERT is a BERT model specifically trained for scientific discovery. It randomly samples 1.14M papers from Semantic Scholar [3], where around 18% papers are from the computer science domain and 82% papers are from the broad biomedical domain. Its corpus has 3.17B tokens and the vocabulary size is 31K. Besides, SciBERT was trained on the full paper, not just the abstract. One potential issue is the vocabulary shift from the Semantic Scholar to PubChemSTM. Although we adapt the pretrained checkpoints from SciBERT (together with its vocabulary) in this work, we still want to carefully examine the vocabulary for the textual data.

**Table 2.** The vocabulary comparison.

Data Source	Tokenization Method	size of vocabulary	overlap with SciBERT
Semantic Scholar (used in SciBERT)	SciBERT tokenizer	31,090	-
PubChemSTM-raw	white space	315,704	7,635
	spaCy	114,976	719
	SciBERT tokenizer	18,320	18,320
PubChemSTM-extract	white space	100,877	7,562
	spaCy	27,519	691
	SciBERT tokenizer	17,442	17,442

In Table 2, we list the vocabulary size of PubChemSTM-raw and PubChemSTM-extract with three tokenization methods: using white space, spaCy [4], and the SciBERT tokenizer. We can observe that the difference between PubChemSTM-raw and PubChemSTM-extract using the SciBERT tokenizer is quite small, compared to the ones using white space and spaCy. Thus, we want to claim that vocabulary is also an important factor, and the SciBERT tokenizer has shown quite a stable tokenization effect. In the future, more comprehensive tokenization and vocabulary are required to push forwards this research line, *i.e.*, to enable the large language model for drug discovery. But it is beyond the scope of this paper and requires efforts from the entire community.

## A.2 Architecture Details

We have two branches, the chemical structure branch  $f_c$  and the textual description branch  $f_t$ .

**Chemical structure branch  $f_c$**  This work considers two types of chemical structures: the SMILES string views the molecule as a sequence and the 2D molecular graph takes the atoms and bonds as the nodes and edges, respectively. Then based on the chemical structures, we apply a deep learning encoder  $f_c$  to get a latent vector as molecule representation. Specifically, for the SMILES string, we take the encoder from MegaMolBART [5], which is pretrained on 500M molecules from ZINC database [6]. For the molecular graph, we take a pretrained graph isomorphism network (GIN) [7] using GraphMVP pretraining [8]. GraphMVP is doing a multi-view pretraining between the 2D topologies and 3D geometries on 250K conformations from GEOM dataset [9]. Thus, though we are not explicitly utilizing the 3D geometries, the state-of-the-art pretrained GIN models can implicitly encode such information.

**Textual description branch  $f_t$**  The textual description branch provides a high-level description of the molecule’s functionality. We can view this branch as domain knowledge to strengthen the molecule representation. Such domain knowledge is in the form of natural language, and we use the BERT model [10] as the text encoder  $f_t$ . We further adapt the pretrained SciBERT [2], 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	10,010,635
	GIN	1,885,206
Textual description	SciBERT	109,918,464

## A.3 Pretraining Details

**Pretraining Objective** For the MoleculeSTM pretraining, we apply contrastive learning. More concretely, we choose one of the EBM-NCE [8] and InfoNCE [11]. Both are essentially doing the same thing, yet EBM-NCE has been found to be more

effective for graph-data [8, 12]. The objective for EBM-NCE is:

$$\mathcal{L} = -\frac{1}{2} \left( \mathbb{E}_{\mathbf{x}_c, \mathbf{x}_t} [\log \sigma(E(\mathbf{x}_c, \mathbf{x}_t))] + \mathbb{E}_{\mathbf{x}_c, \mathbf{x}_t'} [\log(1 - \sigma(E(\mathbf{x}_c, \mathbf{x}_t')))] \right) - \frac{1}{2} \left( \mathbb{E}_{\mathbf{x}_c, \mathbf{x}_t} [\log \sigma(E(\mathbf{x}_c, \mathbf{x}_t))] + \mathbb{E}_{\mathbf{x}_c', \mathbf{x}_t} [\log(1 - \sigma(E(\mathbf{x}_c', \mathbf{x}_t)))] \right), \quad (1)$$

where  $\mathbf{x}_c$  and  $\mathbf{x}_t$  form the structure-text pair for each molecule, and  $\mathbf{x}_{c'}$  and  $\mathbf{x}_{t'}$  are the negative samples randomly sampled from the noise distribution, which we use the empirical data distribution.  $E(\cdot)$  is the energy function with a flexible formulation, and we use the dot product on the jointly learned space, *i.e.*,  $E(\mathbf{x}_c, \mathbf{x}_t) = \langle p_c \circ f_c(\mathbf{x}_c), p_t \circ f_t(\mathbf{x}_t) \rangle$ . Similarly, we have the objective for InfoNCE as:

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

**Hyperparameters** We list the key hyperparameters used for MoleculeSTM pretraining with the SMILES string and 2D molecular graph as inputs, respectively.

**Table 4.** Hyperparameter specifications for MoleculeSTM pretraining.

Input	Hyperparameter	Value
SMILES string	epochs	{32}
	learning rate for text branch	{1e-4}
	learning rate for chemical structure branch	{1e-5, 3e-5}
	objective function	{EBM-NCE, InfoNCE}
2D molecular graph	epochs	{32}
	learning rate for text branch	{1e-4}
	learning rate for chemical structure branch	{1e-5, 3e-5}
	objective function	{EBM-NCE, InfoNCE}

**Running time** We list the running time of MoleculeSTM with the SMILES string and 2D molecular graph as inputs, respectively.

**Table 5.** Running time for MoleculeSTM pretraining.

Input	Running Time
SMILES string	44min / epoch
2D molecular graph	42min / epoch

## B Design Principles for Downstream Tasks

In this section, we discuss the key principles when designing downstream tasks.

**Applicable Evaluation** One of the biggest differences between the foundation model in the vision-language domain and our MoleculeSTM can be reflected in the evaluation. Most of the vision and language tasks can be viewed as art problems, *i.e.*, there does not exist a standard and exact solution that is applicable for evaluation. For instance, we can detect if the image is "a horse riding an astronaut" or "a panda making latte art" [13], but only visually not computationally, which prevents large-scale evaluation. This is not the case for drug discovery, because it is a scientific task, where the results (*e.g.*, properties of the output molecules in the editing task) can be evaluated exactly, either in vitro or in silico. Following this, the physical experiments are usually expensive and long-lasting, so in this work, we want to focus on tasks that are computationally feasible for evaluation.

**Fuzzy Matching** Specifically for the molecule editing task, the text prompts should follow the "fuzzy matching" criterion because there could exist multiple output molecules. This is in contradiction with "exact matching", where the output molecules are deterministic. For example, for the functional group change, we can feed in the prompts like "change the third nitrogen in the ring to oxygen". This prompt is very explicit with an exact solution, and there exist rule-based chemistry tools in handling this problem perfectly. Thus, text-based editing cannot show its benefits in this track. Instead, text-based editing can provide more benefits in the fuzzy matching setting by wandering around the semantically meaningful directions in the latent space. This also reflects the *open vocabulary* attribute of the language model that we have been focusing on.

## C Downstream: Zero-shot Structure-text Retrieval

### C.1 Dataset Construction

The DrugBank database [14] has many fields that can be interesting to explore drug discovery tasks. Here we extract three fields of each small molecule drug for the zero-shot retrieval task: the Description field, the Pharmacodynamics field, and the anatomical therapeutic chemical (ATC) field, as detailed below:

- **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.

We list the key steps in dataset construction as follows:

1. We download the full DrugBank database (in XML format) and small chemical structure files (in SDF format) from the [website](#).
2. We parse the XML file, and extract the data with three fields: Description, Pharmacodynamics, and ATC.
3. We do the mapping from the extracted files to chemical structures in SDF files. For DrugBank-Description and DrugBank-Pharmacodynamics datasets, we exclude the molecules that have shown up in PubChemSTM, filtered with the canonical SMILES. Meanwhile, for DrugBank-ATC, we exclude the molecules satisfying the following two criteria simultaneously:
  - **Chemical structure filtering** If the molecule with the same canonical SMILES has shown up in the PubChemSTM;
  - **Textual data filtering** We first need to define a similarity between two textual data as in Equation (3), where  $\text{text}_{\text{DrugBank}}$  and  $\text{text}_{\text{PubChemSTM}}$  are the textual data for the same molecule from DrugBank and PubChemSTM, respectively,  $\text{len}()$  is the length of textual data, and  $\text{Levenshtein}()$  is the Levenshtein distance between two textual data. Thus, the second condition is: if the similarity between the DrugBank text and the PubChemSTM text is above a certain threshold (*e.g.*, 0.6).

Another detail is that, for DrugBank-ATC, there exist multiple ATC fields ( $\text{text}_{\text{DrugBank}}$ ) for each small molecule. In PubChemSTM, there also exist multiple textual descriptions ( $\text{text}_{\text{PubChemSTM}}$ ) for each molecule. Thus during the textual data filtering step, for each shared molecule between DrugBank and PubChemSTM, we calculate the similarity for all the  $\text{text}_{\text{DrugBank}}-\text{text}_{\text{PubChemSTM}}$  pairs, and exclude the molecule if there exists one pair with similarity above the threshold 0.6.

4. Some basic dataset statistics can be found in Table 6. Notice that ATC has many levels, and we are using level 5 for retrieval in this work.

$$\text{sim}(\text{text}_{\text{DrugBank}}, \text{text}_{\text{PubChemSTM}}) = 1 - \frac{\text{Levenshtein}(\text{text}_{\text{DrugBank}}, \text{text}_{\text{PubChemSTM}})}{\text{len}(\text{text}_{\text{DrugBank}})}. \quad (3)$$

**Table 6.** Statistics on three fields in DrugBank. The filtering steps have been illustrated above.

Field	# structure-text pairs molecule not in PubChemSTM	# structure-text pairs molecule shared in PubChemSTM but text similarity below 0.6	total
DrugBank-Description	1,154	–	1,154
DrugBank-Pharmacodynamics	1,005	–	1,005
DrugBank-ATC	1,507	1,500	3,007

## C.2 Experiments

For experiments, we introduce three baselines in the main body. As a proof-of-concept, we carry out another baseline called Random. For Random, both encoders ( $f_c$  and  $f_i$ ) are randomly initialized. The zero-shot retrieval results on three datasets are shown in Tables 7 to 9.

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

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

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

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

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

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

### C.3 Ablation Study: Fixed Pretrained Encoders

In the main body, we conduct pretraining by adopting pretrained single-modality checkpoints, *i.e.*, the GraphMVP and MegaMolBART for  $f_c$ , and SciBERT for  $f_t$ . Then for MoleculeSTM pretraining, we use contrastive learning and update all the model parameters. Here we take an ablation study by only optimizing the projection layers to the joint space of the two branches ( $p_c, p_t$ ) while keeping the two encoders ( $f_c, f_t$ ) fixed. The results on the three datasets are shown in Tables 10 to 12.

**Table 10.** Accuracy (%) of DrugBank-Description  $T$ -choose-one retrieval.

T		Given Chemical Structure			Given Text		
		4	10	20	4	10	20
SMILES	Random	24.59 $\pm$ 1.14	10.12 $\pm$ 1.38	4.97 $\pm$ 0.42	24.54 $\pm$ 0.97	9.97 $\pm$ 0.81	5.09 $\pm$ 0.37
	Frozen	25.07 $\pm$ 1.24	10.22 $\pm$ 1.19	5.12 $\pm$ 0.65	24.69 $\pm$ 1.87	10.20 $\pm$ 1.38	5.37 $\pm$ 1.15
	Similarity	36.35 $\pm$ 0.59	23.22 $\pm$ 0.58	16.40 $\pm$ 0.59	22.74 $\pm$ 0.24	10.31 $\pm$ 0.24	5.34 $\pm$ 0.24
	MoleculeSTM	47.64 $\pm$ 0.40	29.21 $\pm$ 0.47	19.69 $\pm$ 0.47	52.60 $\pm$ 0.46	32.24 $\pm$ 0.37	21.45 $\pm$ 0.37
Graph	Random	25.78 $\pm$ 1.43	10.71 $\pm$ 0.97	4.83 $\pm$ 1.00	24.98 $\pm$ 0.32	10.20 $\pm$ 0.40	4.80 $\pm$ 0.21
	Frozen	24.01 $\pm$ 1.34	9.39 $\pm$ 0.92	4.85 $\pm$ 0.52	24.00 $\pm$ 1.66	9.91 $\pm$ 0.71	5.07 $\pm$ 0.75
	Similarity	30.03 $\pm$ 0.38	13.63 $\pm$ 0.27	7.07 $\pm$ 0.10	24.81 $\pm$ 0.27	10.22 $\pm$ 0.24	4.74 $\pm$ 0.24
	MoleculeSTM	51.28 $\pm$ 0.00	31.99 $\pm$ 0.41	20.71 $\pm$ 0.47	55.27 $\pm$ 0.00	33.08 $\pm$ 0.00	21.77 $\pm$ 0.00

**Table 11.** Accuracy (%) of DrugBank-Pharmacodynamics  $T$ -choose-one retrieval.

T		Given Chemical Structure			Given Text		
		4	10	20	4	10	20
SMILES	Random	24.49 $\pm$ 0.68	9.73 $\pm$ 0.34	5.14 $\pm$ 0.57	25.61 $\pm$ 0.62	10.10 $\pm$ 0.91	5.07 $\pm$ 0.69
	Frozen	25.47 $\pm$ 1.12	10.55 $\pm$ 0.75	5.48 $\pm$ 0.70	25.34 $\pm$ 0.41	9.86 $\pm$ 0.44	4.84 $\pm$ 0.26
	Similarity	27.85 $\pm$ 0.03	10.75 $\pm$ 0.02	5.67 $\pm$ 0.01	24.58 $\pm$ 0.03	11.25 $\pm$ 0.03	5.29 $\pm$ 0.02
	MoleculeSTM	46.43 $\pm$ 0.00	27.42 $\pm$ 0.47	18.24 $\pm$ 0.47	52.53 $\pm$ 0.41	30.53 $\pm$ 0.00	19.98 $\pm$ 0.00
Graph	Random	26.00 $\pm$ 0.37	9.65 $\pm$ 0.88	4.95 $\pm$ 0.36	25.11 $\pm$ 0.63	9.99 $\pm$ 0.62	4.82 $\pm$ 0.54
	Frozen	25.49 $\pm$ 1.82	10.19 $\pm$ 1.47	4.74 $\pm$ 0.56	25.55 $\pm$ 0.45	10.15 $\pm$ 0.77	4.88 $\pm$ 0.55
	Similarity	25.33 $\pm$ 0.27	9.89 $\pm$ 0.52	4.61 $\pm$ 0.08	25.28 $\pm$ 0.03	10.64 $\pm$ 0.02	5.47 $\pm$ 0.02
	MoleculeSTM	46.29 $\pm$ 0.03	27.18 $\pm$ 0.02	17.73 $\pm$ 0.02	50.95 $\pm$ 0.04	31.65 $\pm$ 0.03	23.00 $\pm$ 0.03

**Table 12.** Accuracy (%) of DrugBank-ATC  $T$ -choose-one retrieval.

T		Given Chemical Structure			Given Text		
		4	10	20	4	10	20
SMILES	Random	25.03 $\pm$ 0.33	9.83 $\pm$ 0.19	4.80 $\pm$ 0.22	25.44 $\pm$ 1.21	10.03 $\pm$ 0.94	5.11 $\pm$ 0.79
	Frozen	25.05 $\pm$ 0.94	10.17 $\pm$ 0.63	4.99 $\pm$ 0.54	25.35 $\pm$ 0.78	10.32 $\pm$ 0.44	5.22 $\pm$ 0.34
	Similarity	30.03 $\pm$ 0.00	13.35 $\pm$ 0.02	7.53 $\pm$ 0.02	26.74 $\pm$ 0.03	11.01 $\pm$ 0.00	5.62 $\pm$ 0.00
	MoleculeSTM	43.41 $\pm$ 0.12	25.66 $\pm$ 0.06	15.69 $\pm$ 0.06	48.75 $\pm$ 0.11	29.44 $\pm$ 0.06	19.75 $\pm$ 0.03
Graph	Random	24.48 $\pm$ 0.66	9.97 $\pm$ 0.25	4.81 $\pm$ 0.34	25.48 $\pm$ 0.59	10.40 $\pm$ 0.37	5.38 $\pm$ 0.30
	Frozen	24.19 $\pm$ 0.77	10.24 $\pm$ 0.71	4.87 $\pm$ 0.47	24.95 $\pm$ 1.52	10.07 $\pm$ 0.80	5.06 $\pm$ 0.36
	Similarity	29.46 $\pm$ 0.00	12.34 $\pm$ 0.00	6.52 $\pm$ 0.00	25.78 $\pm$ 1.53	10.23 $\pm$ 0.70	5.06 $\pm$ 0.67
	MoleculeSTM	42.53 $\pm$ 0.07	24.34 $\pm$ 0.00	14.78 $\pm$ 0.03	48.91 $\pm$ 0.03	28.77 $\pm$ 0.07	19.28 $\pm$ 0.07



## D Downstream: Zero-shot Text-based Molecule Editing

Molecule editing or controllable molecule generation refers to changing the structures of the molecules based on a given and pretrained molecule generative model. In this work, with the help of a large language model in MoleculeSTM, we are able to do the zero-shot text-based molecule editing. First, we would like to list two key challenges, comparing the editing task between the vision domain and molecule domain, as follows:

- **Backbone generative model.** For domains in vision, the image controllable generation can be quite feasible based on StyleGAN [15], a well-disentangled backbone model. However, it is nontrivial for deep molecule generative models. A recent work GraphCG [16] has explored the disentanglement property of the graph-based controllable molecule generation methods, and the conclusion is that, even though the backbone generative models are not perfectly disentangled, there still exist methods for controllable generation on highly structured data like molecular graphs or point clouds. Meanwhile, developing a novel disentangled molecule generative model is out of the scope of this work, since the editing solution by MoleculeSTM is model-agnostic, and can be easily generalized to future models.
- **Evaluation.** Image controllable generation is an art problem, *i.e.*, it is subjective and can have multiple (or even infinitely many) answers. On the contrary, controllable molecule generation is a science problem, *i.e.*, it is objective and has only a few answers. This has been discussed in Appendix B.

### D.1 Experiment Set-up

**Implementation Details** Because most of the modules are fixed, we only need to learn the adaptor module and the optimized latent code  $w$ . The two key hyperparameters are the learning rate  $\{1e-2, 1e-3\}$  and  $\lambda \in \{1e1, 1e0, 1e-1, 1e-2, 1e-3\}$ . As a fair comparison, for baselines, we take the form of  $w = w_{in} + \alpha \cdot D$ , where  $D$  is obtained using random, PCA and variance and  $\lambda \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$ . For GS, we repeat the random sampling five times of each input molecule.

Next, we will conduct the zero-shot text-based molecule editing on four types of editing tasks, as well as three case study, as discussed below:

- Single-objective molecule editing in Appendix D.2 (eight tasks).
- Multi-objective molecule editing in Appendix D.3 (six tasks).
- Binding-affinity-based molecule editing in Appendix D.4 (six tasks).
- Drug relevance editing in Appendix D.5 (four tasks).
- Neighborhood searching for patent drug molecules in Appendix D.6 (three case studies).

Due to the page limit, we only show four multi-objective and four binding-affinity-based editing tasks in the main body. Here we show more comprehensive results.

We want to mention that for single- and multi-objective editing, we randomly select 200 molecules from ZINC as the input molecules. None of these 200 input molecules appears in PubChemSTM. Furthermore, the random selection process ensures that the property distributions of these 200 molecules remain consistent with the entire dataset. Illustrated below (Figures 1 and 2) are three examples of molecular properties: LogP (measuring water solubility), tPSA (measuring permeability) and molecular weight.

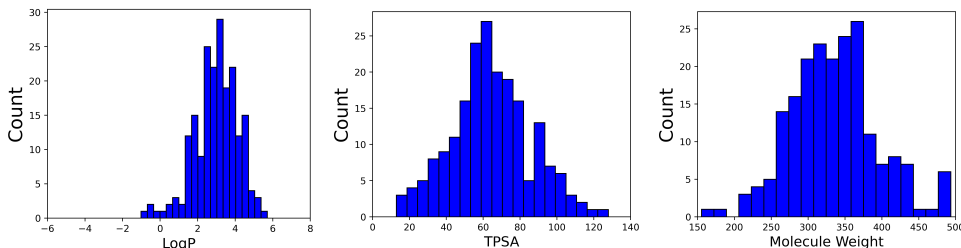


Figure 1. Three property distributions on 200 randomly sampled molecules for editing.

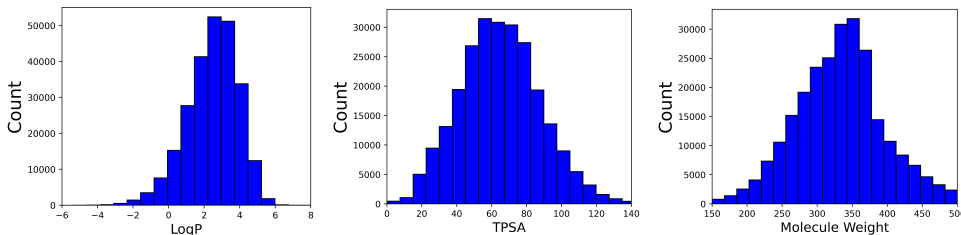


Figure 2. Three property distributions on 250K molecules from ZINC250K.

## D.2 Single-objective Molecule Editing

We first consider eight single-objective properties for molecule editing. As shown in the Methods section, the definitions of the satisfaction function and threshold  $\Delta$  are based on each task specifically, as:

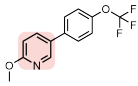
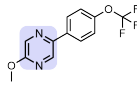
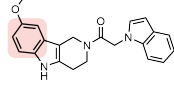
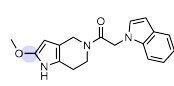
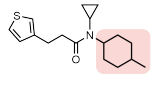
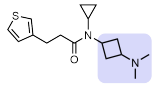
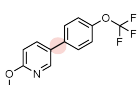
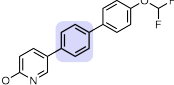
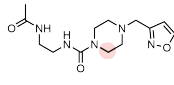
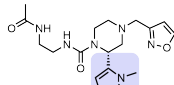
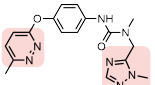
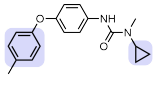
- We use LogP to evaluate the solubility and insolubility. We take 0 and 0.5 as the different thresholds.
- We use QED to evaluate the drug-likeness. We take 0 and 0.1 as the different thresholds.
- We use tPSA to evaluate the high and low permeability. We take 0 and 10 as the different thresholds.
- For the hydrogen bond acceptor (HBA) and hydrogen bond donor (HBD), we can directly count them in the molecules, and we use 0 and 1 as the different thresholds.

For  $\Delta$ , it is the threshold that only difference above it can be viewed as a hit. So the larger  $\Delta$  means a stricter editing criterion. Below we show both the quantitative and qualitative results on eight single-objective property molecule editing results.

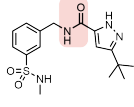
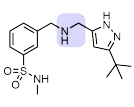
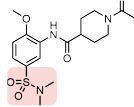
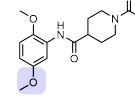
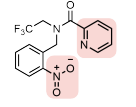
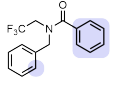
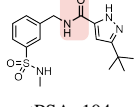
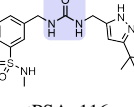
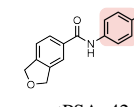
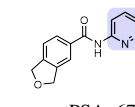
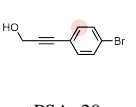
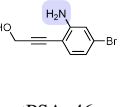
**Table 13.** Results on eight single-objective molecule editing. The inputs are 200 molecules randomly sampled from ZINC, and the evaluation is the hit ratio of the property change. The latent optimization is text-based molecule editing with MoleculeSTM, with the SMILES string and the molecular graph, respectively.

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

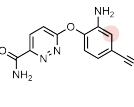
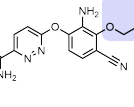
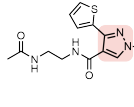
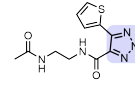
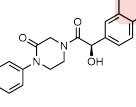
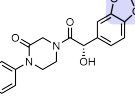
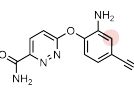
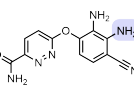
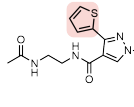
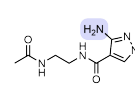
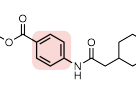
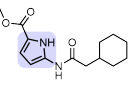
**Table 14.** Visualization of text-based editing on solubility, measured by the logarithm of the octanol-water partition coefficient (LogP) of the molecules. Generally, molecules with smaller LogP are more soluble in water. For generating molecules soluble in water, we can add polar components (e.g., oxygens and nitrogens), remove hydrophobic moieties (e.g., benzene and cyclohexane), or replace hydrophobic groups with polar functionalities in the input molecule. For generating molecules insoluble in water, we can make opposite modifications to the input molecule. The pink and blue regions highlight the modified structure in the input and output molecules, respectively.

Text Prompt: This molecule is <i>soluble in water</i> .					
Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
					
LogP: 3.66	LogP: 3.05	LogP: 3.72	LogP: 2.56	LogP: 4.25	LogP: 2.76
Text Prompt: This molecule is <i>insoluble in water</i> .					
Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
					
LogP: 3.66	LogP: 5.03	LogP: -0.36	LogP: 0.72	LogP: 2.37	LogP: 4.41

**Table 15.** Visualization of text-based editing on permeability, measured by the topological polar surface area (tPSA) of the molecules. Generally, molecules with smaller tPSA are more permeable. For generating molecules with high permeability, we can remove functional groups or heterocycles with high polarity from the input molecule, such as amides, sulfonamides, ureas, nitro groups, and nitrogen-containing arenes. For generating molecules with low permeability, we can make opposite modifications to the input molecule. The pink and blue regions highlight the modified structure in the input and output molecules, respectively.

Text Prompt: This molecule has <i>high permeability</i> .					
Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
 tPSA: 104	 tPSA: 87	 tPSA: 96	 tPSA: 68	 tPSA: 76	 tPSA: 20
Text Prompt: This molecule has <i>low permeability</i> .					
Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
 tPSA: 104	 tPSA: 116	 tPSA: 42	 tPSA: 67	 tPSA: 20	 tPSA: 46

**Table 16.** Visualization of text-based editing on hydrogen bond acceptors (HBA) and hydrogen bond donors (HBD). For generating molecules with more HBA, we can add heteroatoms to the input molecule such as oxygen, nitrogen, and sulfur, or replace existing groups with heteroatom-containing structural motifs. For generating molecules with more HBD, we can add heteroatoms that bear attached hydrogens, such as functional groups like amines, and heterocycles like pyrroles. The pink and blue regions highlight the modified structure in the input and output molecules, respectively.

Text Prompt: This molecule has <i>more hydrogen bond acceptors</i> .					
Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
 HBA: 6	 HBA: 7	 HBA: 5	 HBA: 6	 HBA: 3	 HBA: 5
Text Prompt: This molecule has <i>more hydrogen bond donors</i> .					
Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
 HBD: 2	 HBD: 3	 HBD: 2	 HBD: 3	 HBD: 1	 HBD: 2

### D.3 Multi-objective Molecule Editing

We then consider six multi-objective properties for molecule editing. As shown in the Methods section, the definitions of the satisfaction function and threshold  $\Delta$  are based on each task specifically. First, for each single-objective, we follow the evaluation metric in Appendix D.2, including the solubility, permeability, and the number of HBA and HBD. Then for the multi-objective evaluation, we consider two cases:

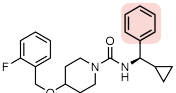
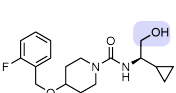
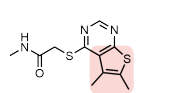
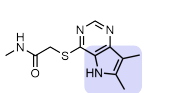
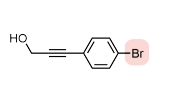
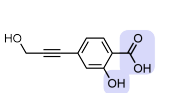
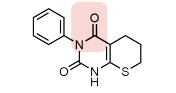
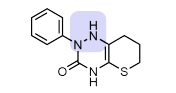
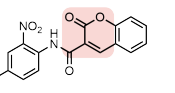
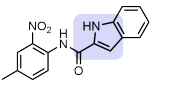
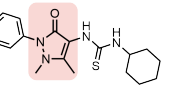
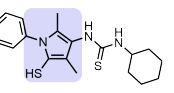
- The **simple** case with the loose thresholds, such as threshold 0 and 0 for solubility and permeability simultaneously.
- The **challenging** case with strict thresholds, such as threshold 0.5 and 1 for solubility and HBA/HBD simultaneously and threshold 0.5 and 10 for solubility and permeability simultaneously.

Then a successful hit needs to satisfy both conditions simultaneously. Below we show both the quantitative and qualitative results on six multi-objective property molecule editing results.

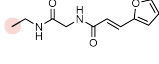
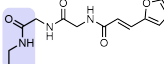
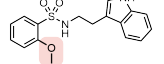
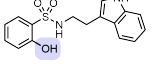
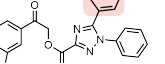
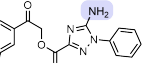
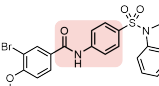
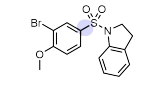
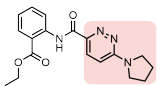
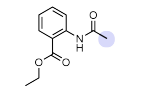
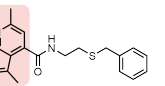
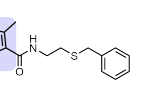
**Table 17.** Results on six multi-objective molecule editing. The inputs are 200 molecules randomly sampled from ZINC, and the evaluation is the hit ratio of the property change. The latent optimization is text-based molecule editing with MoleculeSTM, with the SMILES string and the molecular graph, respectively.

	$\Delta$	baseline				latent optimization	
		Random	PCA	High Variance	GS-Mutate	SMILES	Graph
This molecule is <i>soluble in water</i> and has <i>more hydrogen bond acceptors</i> .	0 – 0	9.88 $\pm$ 1.03	8.64 $\pm$ 2.06	9.09 $\pm$ 1.25	14.00 $\pm$ 2.48	27.87 $\pm$ 3.86	27.43 $\pm$ 3.41
	0.5 – 1	0.23 $\pm$ 0.33	0.45 $\pm$ 0.64	0.22 $\pm$ 0.31	0.67 $\pm$ 0.62	8.80 $\pm$ 0.04	11.10 $\pm$ 1.80
This molecule is <i>insoluble in water</i> and has <i>more hydrogen bond acceptors</i> .	0 – 0	2.99 $\pm$ 0.38	2.00 $\pm$ 0.58	2.45 $\pm$ 0.67	7.17 $\pm$ 0.85	8.55 $\pm$ 2.75	8.21 $\pm$ 0.81
	0.5 – 1	0.45 $\pm$ 0.32	0.00 $\pm$ 0.00	0.22 $\pm$ 0.31	0.17 $\pm$ 0.24	2.93 $\pm$ 0.30	0.00 $\pm$ 0.00
This molecule is <i>soluble in water</i> and has <i>more hydrogen bond donors</i> .	0 – 0	2.28 $\pm$ 1.15	2.23 $\pm$ 1.16	4.44 $\pm$ 0.58	13.83 $\pm$ 2.95	33.51 $\pm$ 4.08	49.23 $\pm$ 1.71
	0.5 – 1	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	9.98 $\pm$ 1.03	23.94 $\pm$ 1.09
This molecule is <i>insoluble in water</i> and has <i>more hydrogen bond donors</i> .	0 – 0	0.69 $\pm$ 0.58	1.96 $\pm$ 0.87	1.79 $\pm$ 0.66	5.67 $\pm$ 0.62	17.03 $\pm$ 2.75	14.42 $\pm$ 3.43
	0.5 – 1	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	2.59 $\pm$ 1.14	3.84 $\pm$ 0.71
This molecule is <i>soluble in water</i> and has <i>high permeability</i> .	0 – 0	5.06 $\pm$ 1.21	3.53 $\pm$ 0.38	4.88 $\pm$ 2.21	8.17 $\pm$ 1.03	35.69 $\pm$ 3.19	39.74 $\pm$ 2.26
	0.5 – 10	1.16 $\pm$ 0.68	0.67 $\pm$ 0.55	0.66 $\pm$ 0.54	0.00 $\pm$ 0.00	19.15 $\pm$ 0.73	22.66 $\pm$ 1.90
This molecule is <i>soluble in water</i> and has <i>low permeability</i> .	0 – 0	12.17 $\pm$ 1.05	10.43 $\pm$ 2.88	13.08 $\pm$ 2.28	19.83 $\pm$ 2.46	44.35 $\pm$ 0.68	30.87 $\pm$ 0.62
	0.5 – 10	6.20 $\pm$ 0.64	6.23 $\pm$ 2.31	6.67 $\pm$ 0.53	4.83 $\pm$ 0.85	28.67 $\pm$ 2.22	20.06 $\pm$ 1.26

**Table 18.** Visualization of text-based editing on multi-objective (compositionality) properties: solubility and hydrogen bond donors (HBD), measured by LogP and number of HBD of the molecules. Molecules with more HBD are likely also soluble in water, such as replacing hydrophobic groups (benzene, thiophene, bromide, etc.) with polar groups or rings containing hydrogen-attached heteroatoms (alcohol, azaindole, carboxylic acid, etc.) in the input molecules. Nevertheless, we can add HBD to the input molecule while reducing its solubility, such as replacing high-polarity structural motifs (amide, lactone, etc.) with less hydrophilic HBD (indole, thiol, etc.) in the input molecules. The pink and blue regions highlight the modified structure in the input and output molecules, respectively.

Text Prompt: This molecule is <i>soluble in water</i> and has <i>more hydrogen bond donors</i> .					
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 <i>insoluble in water</i> and has <i>more hydrogen bond donors</i> .					
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

**Table 19.** Visualization of text-based editing on multi-objective (compositionality) properties: solubility and permeability, measured by LogP and tPSA of the molecules. Molecules with low permeability are likely also soluble in water, such as adding polar functional groups (*e.g.*, amide, amine) and removing hydrocarbons (*e.g.*, methyl, phenyl) with regard to the input molecules. Nevertheless, we can increase both the solubility and permeability of the molecule, such as removing hydrocarbons and polar moieties simultaneously or reducing the size of the heterocycles (*e.g.*, [1,2]oxazolo[5,4-*b*]pyridine to imidazole) in the input molecules. The pink and blue regions highlight the modified structure in the input and output molecules, respectively.

Text Prompt: This molecule is <i>soluble in water</i> and has <i>low permeability</i> .					
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 <i>soluble in water</i> and has <i>high permeability</i> .					
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

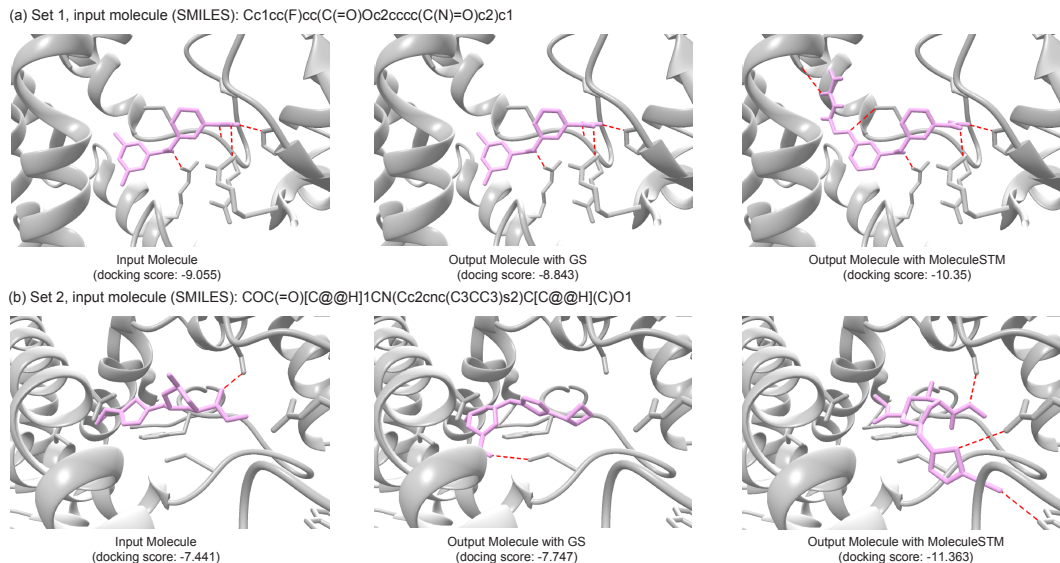
#### D.4 Binding-affinity-based Molecule Editing

We further apply text-based editing on the binding affinity assays. In specific, we take six binding affinity tasks from ChEMBL [17]. Each assay has a textual description, as listed in Table 20.

**Table 20.** ChEMBL assay descriptions.

ChEMBL ID	Assay Description
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.
1613797	This molecule is tested positive in an assay for Anthrax Lethal, which acts as a protease that cleaves the N-terminal of most dual specificity mitogen-activated protein kinase kinases.
2114713	This molecule is tested positive in an assay for Activators of ClpP, which cleaves peptides in various proteins in a process that requires ATP hydrolysis and has a limited peptidase activity in the absence of ATP-binding subunits.
1613838	This molecule is tested positive in an assay for activators involved in the transport of proteins between the endosomes and the trans Golgi network.
1614236	This molecule is an inhibitor of a protein that prevents the establishment of the cellular antiviral state by inhibiting ubiquitination that triggers antiviral transduction signal and inhibits post-transcriptional processing of cellular pre-mRNA.
1613903	This molecule is tested positive in the high throughput screening assay to identify inhibitors of the SARS coronavirus 3C-like Protease, which cleaves the C-terminus of replicase polypeptide at 11 sites.

For evaluation, we follow the Methods section. Recall that each binding affinity assay can correspond to molecules with positive and negative labels. Thus, we can train a classifier on these data points, and the satisfy criteria here is if the output molecules can have higher confidence than the input molecule, where the confidence is predicted using the classifier for each task. The pipeline can be found in Figure 3.



**Figure 3.** Pipeline for binding-affinity-based molecule editing. The input molecules are randomly sampled from ZINC, and the text prompt is the assay description. For evaluation, the small molecules for each assay are used to train a binary classifier, and two types of models (random forest and logistic regression) are considered.

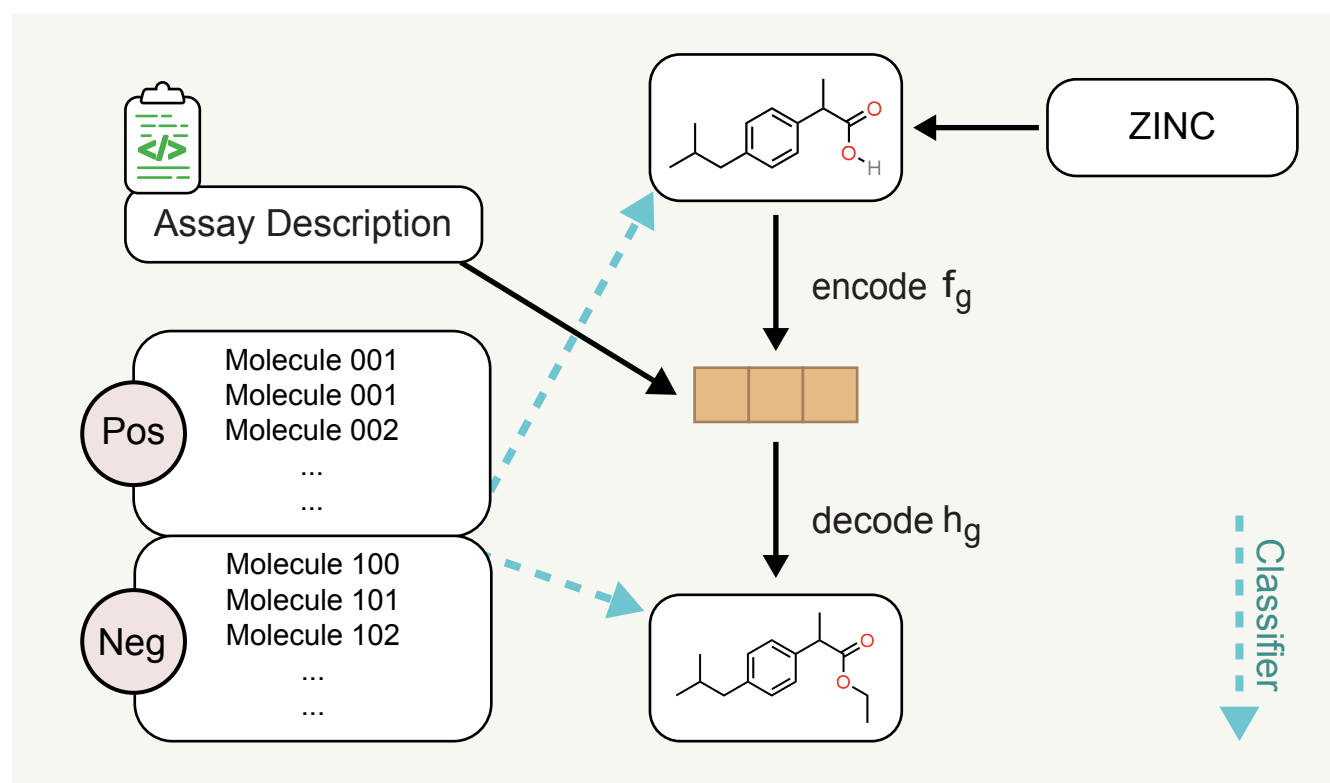
The hit ratio results are shown in Table 21. Notice that to better prove the validity of our results, we train two classifiers for each assay: random forest (RF) and logistic regression (LR), with the fingerprint as featurization. the  $\Delta$  is 0.

**Table 21.** Results on six ChEMBL assay editing. Each ChEMBL assay is a binary task and we train a classifier to obtain the confidence score of each molecule (input and output molecules). The inputs are 200 molecules randomly sampled from ZINC, and the evaluation is the hit ratio of the confidence change. The latent optimization is the text-based molecule editing with MoleculeSTM, with the SMILES string and the molecular graph, respectively.

ChEMBL ID		baseline				latent optimization	
		Random	PCA	High Variance	GS-Mutate	SMILES	Graph
1613777	RF	44.99 $\pm$ 2.08	44.49 $\pm$ 1.22	44.45 $\pm$ 1.01	39.17 $\pm$ 3.66	48.70 $\pm$ 2.06	44.53 $\pm$ 1.60
	LR	47.34 $\pm$ 5.53	49.13 $\pm$ 0.86	49.69 $\pm$ 6.75	51.50 $\pm$ 2.86	54.09 $\pm$ 1.94	50.55 $\pm$ 3.14
1613797	RF	44.76 $\pm$ 2.18	46.25 $\pm$ 0.97	46.92 $\pm$ 3.34	46.67 $\pm$ 1.55	55.03 $\pm$ 2.23	49.03 $\pm$ 0.03
	LR	48.40 $\pm$ 3.71	49.92 $\pm$ 4.31	48.67 $\pm$ 1.64	49.17 $\pm$ 3.01	57.98 $\pm$ 3.34	54.95 $\pm$ 3.74
2114713	RF	39.87 $\pm$ 2.32	42.91 $\pm$ 2.64	42.19 $\pm$ 3.68	41.33 $\pm$ 1.25	49.20 $\pm$ 2.11	60.93 $\pm$ 2.53
	LR	51.39 $\pm$ 1.15	52.62 $\pm$ 1.64	52.24 $\pm$ 1.07	50.50 $\pm$ 1.47	56.93 $\pm$ 3.67	58.77 $\pm$ 2.41
1613838	RF	44.49 $\pm$ 1.48	44.71 $\pm$ 1.80	45.30 $\pm$ 2.47	36.00 $\pm$ 2.68	43.94 $\pm$ 3.75	49.13 $\pm$ 2.52
	LR	50.22 $\pm$ 4.23	49.73 $\pm$ 2.33	44.69 $\pm$ 2.41	41.33 $\pm$ 3.17	47.50 $\pm$ 2.28	56.13 $\pm$ 1.50
1614236	RF	41.33 $\pm$ 3.59	42.28 $\pm$ 1.91	42.85 $\pm$ 2.88	45.33 $\pm$ 1.65	57.90 $\pm$ 2.39	35.71 $\pm$ 4.19
	LR	46.57 $\pm$ 0.51	49.34 $\pm$ 1.80	50.62 $\pm$ 3.86	56.00 $\pm$ 1.08	65.78 $\pm$ 5.67	46.36 $\pm$ 2.53
1613903	RF	44.28 $\pm$ 0.77	43.83 $\pm$ 2.65	42.00 $\pm$ 3.19	46.17 $\pm$ 0.85	56.82 $\pm$ 3.96	58.70 $\pm$ 1.43
	LR	53.94 $\pm$ 3.30	48.63 $\pm$ 4.49	56.19 $\pm$ 2.51	56.33 $\pm$ 0.94	58.31 $\pm$ 2.98	64.64 $\pm$ 5.23

Then we add docking for visualization in Figure 4. We choose the ChEMBL 1613777 with the available PDB structure. In specific, we first extract the output molecules using MoleculeSTM with confidence (RF and LR) higher than the ones generated with baselines. Then we run the molecular docking software for the results. The details of docking settings are listed below.

- We use Merck molecular force field (MMFF) [18] provided in RDKit [19] to embed (generate) 3D conformers for each molecule. The dielectric constant is set to be 80 and the maximum iteration of optimization is 1000 for MMFF, and the up-to-5 conformers from each molecule are used for further analysis.
- For the binding target, we consider assay P450 (CYP) 2C19 [20] (ChEMBL id: 1613777) and select the corresponding crystal structure available in the Protein Data Bank (PDB) (PDB id: 4GQS). Further, we take chain A for docking running. Later for the binding, the binding pockets are aligned with the original ligand in the crystal structure of PDB complexes: the center is set to (-81.48, 16.55, -41.6), and the box is (20.0, 23.0, 25.0).
- Then we take a preprocessing step to complement the hydrogen atoms and add partial charges. We utilize meeko v0.3.3 for small molecules and AutoDock Flexible Receptor (ADFR) suite v1.2 for proteins.
- For docking, we use AutoDock Vina v1.2.3 [21]. Each molecule conformer is docked with *exhaustiveness* being 32, and the pose with the best (lowest) docking score is picked and used for visualization. For visualization, we use UCSF Chimera.



**Figure 4.** Two sets of docking visualization for binding-affinity-based molecule editing. The text prompt is from 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.”). For visualization, the input molecule and output molecules with GS and MoleculeSTM are displayed. It is observed that MoleculeSTM can generate molecules with the lowest docking scores (with the most Hydrogen bonds, and marked in red dashed lines). In set 1 (a), the output molecules are sharing the same molecule scaffold. In set 2 (b), the motif of the output molecule using MoleculeSTM also changes.



## D.5 Drug Relevance Editing

As a proof-of-concept, we further take four editing tasks on common drug editing. The text prompts used here are to make the input molecules look like an existing drug, *e.g.*, “This molecule looks like *Penicillin*.” Following the Methods section, the satisfy function used is the Tanimoto similarity, and the threshold  $\Delta$  takes the value of 0 and 0.05.

**Table 22.** Results on four common drug molecule editing. The inputs are 200 molecules randomly sampled from ZINC, and the evaluation is the hit ratio on the increase of the Tanimoto similarity with the common drug. The latent optimization is text-based molecule editing with MoleculeSTM, with the SMILES string and the molecular graph, respectively.

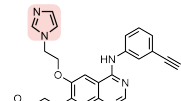
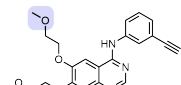
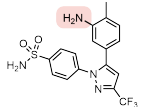
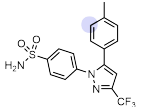
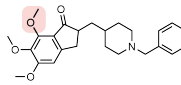
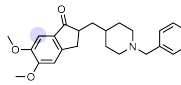
	$\Delta$	baseline				latent optimization	
		Random	PCA	High Variance	GS-Mutate	SMILES	Graph
This molecule <i>looks like Penicillin</i> .	0	43.61 $\pm$ 2.23	46.51 $\pm$ 3.02	44.42 $\pm$ 3.56	28.67 $\pm$ 0.94	58.13 $\pm$ 0.97	50.91 $\pm$ 2.80
	0.05	0.69 $\pm$ 0.55	0.23 $\pm$ 0.32	0.89 $\pm$ 0.30	0.67 $\pm$ 0.62	11.01 $\pm$ 0.58	3.64 $\pm$ 0.57
This molecule <i>looks like Aspirin</i> .	0	43.82 $\pm$ 1.41	43.12 $\pm$ 5.35	44.63 $\pm$ 3.33	25.00 $\pm$ 2.16	40.13 $\pm$ 1.33	54.05 $\pm$ 3.58
	0.05	2.99 $\pm$ 0.38	3.08 $\pm$ 0.82	2.45 $\pm$ 0.33	0.33 $\pm$ 0.47	4.28 $\pm$ 1.22	10.84 $\pm$ 1.26
This molecule <i>looks like Caffeine</i> .	0	42.71 $\pm$ 3.16	40.33 $\pm$ 0.71	40.64 $\pm$ 3.89	26.17 $\pm$ 1.31	46.08 $\pm$ 3.81	51.01 $\pm$ 1.22
	0.05	0.69 $\pm$ 0.01	0.23 $\pm$ 0.32	0.44 $\pm$ 0.31	0.33 $\pm$ 0.24	1.61 $\pm$ 0.67	0.61 $\pm$ 0.01
This molecule <i>looks like Dopamine</i> .	0	42.00 $\pm$ 3.08	42.50 $\pm$ 2.12	41.33 $\pm$ 2.86	30.50 $\pm$ 1.63	47.00 $\pm$ 4.11	55.50 $\pm$ 2.73
	0.05	0.00 $\pm$ 0.00	0.44 $\pm$ 0.31	0.22 $\pm$ 0.31	0.83 $\pm$ 0.24	2.30 $\pm$ 0.44	6.24 $\pm$ 0.56

## D.6 Case Studies on Neighborhood Searching for Patent Drug Molecules

To demonstrate the utility of text-based molecule editing, we show three case studies of generating approved drugs from their analogs. Lead optimization is a critical phase of drug discovery in which closely related compounds are made based on the lead molecule, aiming to improve its efficacy and DMPK (drug metabolism and pharmacokinetics) properties and ultimately identifying a drug candidate [22]. A text prompt calling for greater drug-like properties will thus be informative towards improving on deficiencies in the lead molecule and accelerating drug discovery research.

In specific here, the input molecules are the patented analogs of each approved drug molecule, and the input text prompt is single-objective, like the ones in Appendix D.2. The goal here is to check if the approved drugs can be successfully generated as the output molecules, with the structural changes consistent with the property improvement reflected in the text prompt. For example, in Table 23 (a), Erlotinib is successfully generated from an analog by replacing an imidazole substituent to a methoxy group [23]. This change reflects a tPSA drop from 83 to 75, consistent with the text prompt indicating a higher permeability. Table 23 (b) generates Celecoxib from its amino-substituted derivative [24], where the removal of the amino group yields a greater intestinal permeability of the molecule leading to higher bioavailability. Bioavailability is the fraction of a drug molecule that reaches the systemic circulation, a key factor for oral drug absorption [25]. Finally, Table 23 (c) illustrates how potential metabolic liabilities in a molecule can be addressed via text-based editing. A text calling for a metabolically stable molecule successfully turns a trimethoxy arene to a dimethoxy arene in Donepezil [26], where the former represents an electron-rich aromatic compound known to undergo oxidative phase I metabolisms [27].

**Table 23.** Visualization on three single-objective molecule editing on drug analogs that generates approved drugs based on the text prompt. The pink and blue regions highlight the modified structure in the input and output molecules, respectively.

(a) Prompt: This molecule has <i>high permeability</i> .		(b) Prompt: This molecule has <i>high bioavailability</i> .		(c) Prompt: This molecule is <i>metabolically stable</i> .	
Input Molecule	Output Molecule	Input Molecule	Output Molecule	Input Molecule	Output Molecule
					
CAS: 183320-43-6	Tarevea (Erlotinib)	CAS: 170570-28-2	Celebrex (Celecoxib)	CAS: 120013-52-7	Aricept (Donepezil)

## E Downstream: Molecular Property Prediction

In this section, we review two main categories of datasets used for molecular property prediction downstream tasks from MoleculeNet and molecule benchmarking works [28, 29].

**Molecular Property: Pharmacology** The Blood-Brain Barrier Penetration (BBBP) [30] dataset measures whether a molecule will penetrate the central nervous system. All three toxicity-related datasets, Tox21 [31], ToxCast [28], and ClinTox [32] are related to the toxicity of molecular compounds. The Side Effect Resource (SIDER) [33] dataset stores the adverse drug reactions on a marketed drug database.

**Molecular Property: Biophysics** Maximum Unbiased Validation (MUV) [34] is another sub-database from PCBA, and is obtained by applying a refined nearest neighbor analysis. HIV is from the Drug Therapeutics Program (DTP) AIDS Antiviral Screen [35], and it aims at predicting the inhibition of HIV replication. BACE measures the binding results for a set of inhibitors of  $\beta$ -secretase 1 (BACE-1) and is gathered in MoleculeNet [28].

**Table 24.** Summary for the molecule chemical datasets.

Dataset	Task	# Tasks	# Molecules
BBBP	Classification	1	2,039
Tox21	Classification	12	7,831
ToxCast	Classification	617	8,576
Sider	Classification	27	1,427
ClinTox	Classification	2	1,478
MUV	Classification	17	93,087
HIV	Classification	1	41,127
Bace	Classification	1	1,513

For data splitting, we adopt the scaffold splitting [28]. Scaffold measures the skeleton structure of molecules, and scaffold splitting means we will put the molecules with more common scaffolds into training, and the rest into validation and test, so as to mimic the out-of-distribution (OOD) setting. The OOD setting is more common in real scenarios and thus is preferred to test the pretrained molecule representation power.

**Implementation Details** For the SMILES string, we use MegaMolBART [5] as the backbone Transformer model. For the molecular graph, we use the same backbone GIN model, and we use rich features (as used for the regression tasks in GraphMVP [8]). We list the main hyperparameters below.

**Table 25.** Hyperparameter specifications for molecular property prediction.

	Hyperparameter	Value
Pretraining Baseline	epochs	{100}
	learning rate	{1e-3}
	weight decay	{0}
Downstream	epochs	{100}
	learning rate	{1e-3, 5e-4}
	weight decay	{0}

**Choice of backbone models.** We want to clarify that the MoleculeSTM is agnostic to the backbone encoders for each modality, *e.g.*, the molecule representation model.

- For the backbone model, we use the GIN model as the fixed 2D GNN backbone encoder. In other words, the performance of MoleculeSTM is limited by the 2D backbone model.
- In the molecule pretraining research line, (*e.g.*, AttrMask [36], MolCLR [37], GraphMVP [8], MoleculeSDE [38]), all of these works adopt GIN as the 2D backbone model, serving as a control to test the effectiveness of various pretraining algorithms. This is a similar case for our proposed MoleculeSTM.
- In the future, we would like to explore more advanced GNN models on molecules.

## Supplementary References

- [1] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. “PubChem in 2021: new data content and improved web interfaces”. In: *Nucleic acids research* 49.D1 (2021), pp. D1388–D1395.
- [2] Iz Beltagy, Kyle Lo, and Arman Cohan. “SciBERT: Pretrained Language Model for Scientific Text”. In: *EMNLP*. 2019. eprint: [arXiv:1903.10676](https://arxiv.org/abs/1903.10676).
- [3] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. “Construction of the literature graph in semantic scholar”. In: *arXiv preprint arXiv:1805.02262* (2018).
- [4] M Honnibal, I Montani, and S Van Landeghem. “Boyd”. In: *A. spaCy: industrial-strength natural language processing in Python* (2020).
- [5] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. “Chemformer: a pre-trained transformer for computational chemistry”. In: *Machine Learning: Science and Technology* 3.1 (2022), p. 015022.
- [6] Teague Sterling and John J Irwin. “ZINC 15—ligand discovery for everyone”. In: *Journal of chemical information and modeling* 55.11 (2015), pp. 2324–2337.
- [7] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. “How powerful are graph neural networks?” In: *arXiv preprint arXiv:1810.00826* (2018).
- [8] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. “Pre-training Molecular Graph Representation with 3D Geometry”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=xQUelpOKPam>.
- [9] Simon Axelrod and Rafael Gomez-Bombarelli. “GEOM, energy-annotated molecular conformations for property prediction and molecular generation”. In: *Scientific Data* 9.1 (2022), pp. 1–14.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [11] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding”. In: *arXiv preprint arXiv:1807.03748* (2018).
- [12] Kaveh Hassani and Amir Hosein Khasahmadi. “Contrastive multi-view representation learning on graphs”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4116–4126.
- [13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”. In: *arXiv preprint arXiv:2205.11487* (2022).
- [14] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. “DrugBank 5.0: a major update to the DrugBank database for 2018”. In: *Nucleic acids research* 46.D1 (2018), pp. D1074–D1082.
- [15] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.
- [16] Shengchao Liu, Chengpeng Wang, Weili Nie, Hanchen Wang, Jiarui Lu, Bolei Zhou, and Jian Tang. “GraphCG: Unsupervised Discovery of Steerable Factors in Graphs”. In: *NeurIPS 2022 Workshop: New Frontiers in Graph Learning*. 2022. URL: [https://openreview.net/forum?id=BhR44NzeK\\_1](https://openreview.net/forum?id=BhR44NzeK_1).
- [17] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. “ChEMBL: towards direct deposition of bioassay data”. In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp. D930–D940. ISSN: 0305-1048. DOI: [10.1093/nar/gky1075](https://doi.org/10.1093/nar/gky1075). eprint: <https://academic.oup.com/nar/article-pdf/47/D1/D930/27437436/gky1075.pdf>. URL: <https://doi.org/10.1093/nar/gky1075>.
- [18] Thomas A Halgren. “Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94”. In: *Journal of computational chemistry* 17.5-6 (1996), pp. 490–519.
- [19] Greg Landrum et al. *RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling*. 2013.
- [20] R Leila Reynald, Stefaan Sansen, C David Stout, and Eric F Johnson. “Structural Characterization of Human Cytochrome P450 2C19: ACTIVE SITE DIFFERENCES BETWEEN P450s 2C8, 2C9, AND 2C19”. In: *Journal of Biological Chemistry* 287.53 (2012), pp. 44581–44591.
- [21] Oleg Trott and Arthur J Olson. “AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading”. In: *Journal of computational chemistry* 31.2 (2010), pp. 455–461.

- [22] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. “Principles of early drug discovery”. In: *British journal of pharmacology* 162.6 (2011), pp. 1239–1249.
- [23] Rodney Caughren Schnur and Lee Daniel Arnold. *Alkynyl and azido-substituted 4-anilinoquinazolines*. US Patent 5,747,498. May 1998.
- [24] John J Talley, Thomas D Penning, Paul W Collins, Donald J Rogier Jr, James W Malecha, Julie M Miyashiro, Stephen R Bertenshaw, Ish K Khanna, Matthew J Graneto, Roland S Rogers, et al. *Substituted pyrazolyl benzenesulfonamides for the treatment of inflammation*. US Patent 5,760,068. June 1998.
- [25] David Dahlgren and Hans Lennernäs. “Intestinal Permeability and Drug Absorption: Predictive Experimental, Computational and In Vivo Approaches”. In: *Pharmaceutics* 11.8 (2019). ISSN: 1999-4923. DOI: [10.3390/pharmaceutics11080411](https://doi.org/10.3390/pharmaceutics11080411). URL: <https://www.mdpi.com/1999-4923/11/8/411>.
- [26] Hachiro Sugimoto, Youichi Iimura, Yoshiharu Yamanishi, and Kiyomi Yamatsu. “Synthesis and Structure-Activity Relationships of Acetylcholinesterase Inhibitors: 1-Benzyl-4-[(5,6-dimethoxy-1-oxoindan-2-yl)methyl]piperidine Hydrochloride and Related Compounds”. In: *Journal of Medicinal Chemistry* 38.24 (1995). PMID: 7490731, pp. 4821–4829. DOI: [10.1021/jm00024a009](https://doi.org/10.1021/jm00024a009). eprint: <https://doi.org/10.1021/jm00024a009>. URL: <https://doi.org/10.1021/jm00024a009>.
- [27] Gordon Guroff, Jean Renson, Sidney Udenfriend, John W Daly, Donald M Jerina, and Bernhard Witkop. “Hydroxylation-Induced Migration: The NIH Shift: Recent experiments reveal an unexpected and general result of enzymatic hydroxylation of aromatic compounds.” In: *Science* 157.3796 (1967), pp. 1524–1530.
- [28] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. “MoleculeNet: a benchmark for molecular machine learning”. In: *Chemical science* 9.2 (2018), pp. 513–530.
- [29] Hanchen Wang, Jean Kaddour, Shengchao Liu, Jian Tang, Matt Kusner, Joan Lasenby, and Qi Liu. *Evaluating Self-Supervised Learning for Molecular Graph Embeddings*. 2022. URL: <https://openreview.net/forum?id=ctX2eXYIW3>.
- [30] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. “A Bayesian approach to in silico blood-brain barrier penetration modeling”. In: *Journal of chemical information and modeling* 52.6 (2012), pp. 1686–1697.
- [31] Tox21 Data Challenge. “Tox21 Data Challenge 2014”. In: <https://tripod.nih.gov/tox21/challenge/> (2014).
- [32] Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. “A data-driven approach to predicting successes and failures of clinical trials”. In: *Cell chemical biology* 23.10 (2016), pp. 1294–1301.
- [33] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. “The SIDER database of drugs and side effects”. In: *Nucleic acids research* 44.D1 (2015), pp. D1075–D1079.
- [34] Sebastian G. Rohrer and Knut Baumann. “Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data”. In: *Journal of Chemical Information and Modeling* 49.2 (2009). PMID: 19161251, pp. 169–184. DOI: [10.1021/ci8002649](https://doi.org/10.1021/ci8002649). eprint: <https://doi.org/10.1021/ci8002649>. URL: <https://doi.org/10.1021/ci8002649>.
- [35] Daniel Zaharevitz. *Aids antiviral screen data*. 2015.
- [36] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. “Strategies for pre-training graph neural networks”. In: *International Conference on Learning Representations, ICLR*. 2020.
- [37] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. “Molclr: Molecular contrastive learning of representations via graph neural networks”. In: *arXiv preprint arXiv:2102.10056* (2021).
- [38] Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. “A group symmetric stochastic differential equation model for molecule multi-modal pretraining”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 21497–21526.