Translation between Molecules and Natural Language

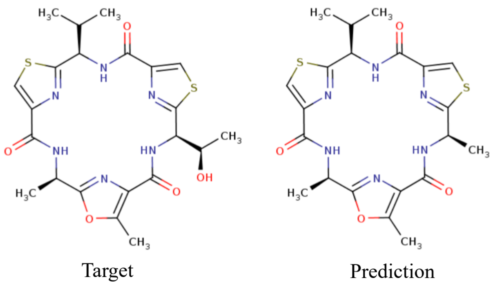
Carl Edwards[1]*, Tuan Lai[1,2]*, Kevin Ros[1], Garrett Honke[2], Kyunghyun Cho[3,4], Heng Ji[1]
[1]University of Illinois Urbana-Champaign
[2]X, the Moonshot Factory
[3]New York University, [4] Genentech
{cne2, tuanml2, kjros2, hengji}@illinois.edu
ghonk@google.com, kyunghyun.cho@nyu.edu

Abstract

We present **MolT5** – a self-supervised learning framework for pretraining models on a vast amount of unlabeled natural language text and molecule strings. **MolT5** allows for new, useful, and challenging analogs of traditional vision-language tasks, such as molecule captioning and text-based de novo molecule generation (altogether: translation between molecules and language), which we explore for the first time. Since **MolT5** pretrains models on single-modal data, it helps overcome the chemistry domain shortcoming of data scarcity. Furthermore, we consider several metrics, including a new cross-modal embedding-based metric, to evaluate the tasks of molecule captioning and text-based molecule generation. Our results show that **MolT5**-based models are able to generate outputs, both molecules and captions, which in many cases are high quality[1].

1 Introduction

Imagine a future where a doctor can write a few sentences describing a specialized drug for treating a patient and then receive the exact structure of the desired drug. Although this seems like science fiction now, with progress in integrating natural language and molecules, it might well be possible in the future. Historically, drug creation has commonly been done by humans who design and build individual molecules. In fact, bringing a new drug to market can cost over a billion dollars and take over ten years (Gaudelet et al., 2021). Recently, there has been considerable interest in using new deep learning tools to facilitate in silico drug design– a field often called cheminformatics (Rifaioglu et al., 2018). Yet, many of these experiments still focus on molecules and their low-level properties such as logP (the octanol-water partition coefficient) (Bagal et al., 2021). In the future, we foresee a need for a higher-level control over molecule design, which can easily be facilitated by natural language.



The molecule is an eighteen-membered homodetic cyclic peptide which is isolated from Oscillatoria sp. and exhibits antimalarial activity against the W2 chloroquine-resistant strain of the malarial parasite, Plasmodium falciparum. It has a role as a metabolite and an antimalarial. It is a homodetic cyclic peptide, a member of 1,3-oxazoles, a member of 1,3-thiazoles and a macrocycle.

Figure 1: An example output from our model for the molecule generation task. The left is the ground truth, and the right is a molecule generated from the given natural language caption.

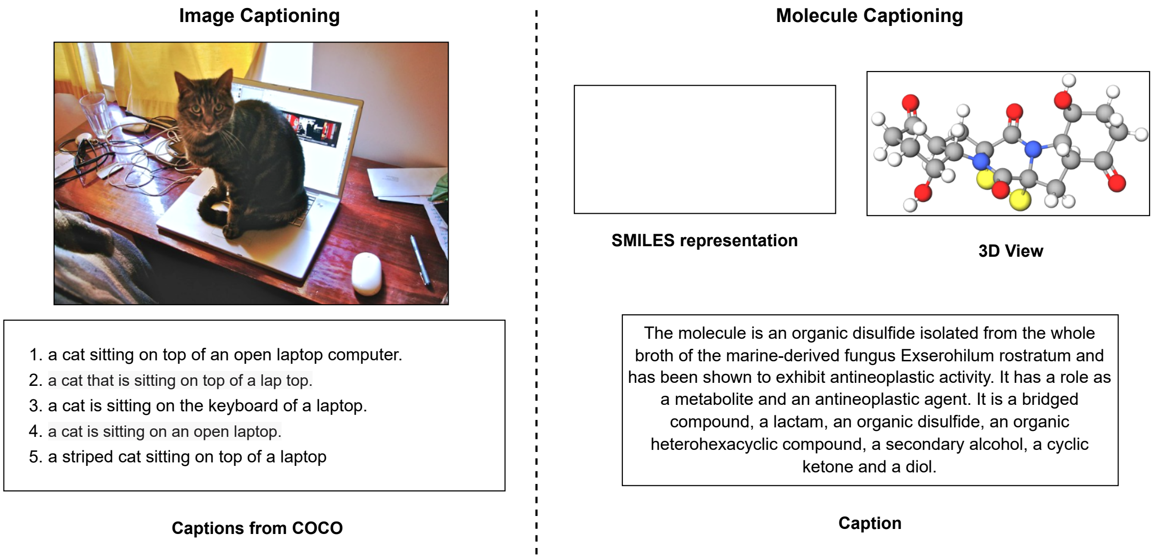
In this work, we pursue an ambitious goal of translating between molecules and language by proposing two new tasks: molecule captioning and text-guided de novo molecule generation. In molecule captioning, we take a molecule (e.g., as a SMILES string) and generate a caption that describes it (Figure 2). In text-guided molecule generation, the task is to create a molecule that matches a given natural language description (Figure 1). These new tasks would help to accelerate research in multiple scientific domains by enabling chemistry domain experts to generate new molecules and better understand them using natural language.

While our proposed molecule-language tasks share some similarities with vision-language tasks, they have several inherent difficulties that separate them from existing vision-language analogs: 1) creating annotations for molecules requires significant domain expertise, 2) thus, it is significantly more difficult to acquire large numbers of molecule-

* indicates equal contributions.

[1] All resources are publicly available at github.com/blender-nlp/MolT5

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 375–413
December 7-11, 2022 ©2022 Association for Computational Linguistics



1. a cat sitting on top of an open laptop computer.
2. a cat that is sitting on top of a lap top.
3. a cat is sitting on the keyboard of a laptop.
4. a cat is sitting on an open laptop.
5. a striped cat sitting on top of a laptop

The molecule is an organic disulfide isolated from the whole broth of the marine-derived fungus Exserohilum rostratum and has been shown to exhibit antineoplastic activity. It has a role as a metabolite and an antineoplastic agent. It is a bridged compound, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary alcohol, a cyclic ketone and a diol.

Figure 2: An example of both the image captioning task (Chen et al., 2015) and molecule captioning. Molecule captioning is considerably more difficult because of the increased linguistic variety in possible captions.

description pairs, 3) the same molecule can have many functions and thus be described in very different ways, which causes 4) existing evaluation measures based on reference descriptions, such as BLEU, to fail to adequately evaluate these tasks.

To address the issue of data scarcity (i.e., difficulties 1 and 2), we propose a new self-supervised learning framework named MolT5 (**Mol**ecular **T5**) that is inspired by the recent progress in pretraining multilingual models (Devlin et al., 2019; Liu et al., 2020). MolT5 first pretrains a model on a vast amount of unlabeled natural language text and molecule strings using a simple denoising objective. After that, the pretrained model is finetuned on limited gold standard annotations. Furthermore, to adequately evaluate models for molecule captioning or generation, we consider various kinds of metrics and also adopt a new metric based on Text2Mol (Edwards et al., 2021). We repurpose this retrieval model for assessing the similarity between the ground truth molecule/description and the generated description/molecule, respectively.

To the best of our knowledge, there is no work yet on molecule captioning or text-guided molecule generation. The closest existing work to molecule captioning falls within the scope of image captioning (Vinyals et al., 2015). However, molecule captioning is arguably much more challenging due to the increased linguistic variety in possible captions (Figure 2). A molecule could be described with an IUPAC name, with one of many different synthetic routes from known precursor molecules, in terms of the properties (e.g. carcinogenic or lipophilic), with the applications of the molecule (e.g. a dye, an antipneumonic, or an antifungal), or in terms of its functional groups (e.g. "substituted by hydroxy groups at positions 5 and 7 and a methyl group at position 8"), among other methods.

In summary, our main contributions are:

1. We propose two new tasks: 1) molecule captioning, where a description is generated for a given molecule, and 2) text-based de novo molecule generation, where a molecule is generated to match a given text description.
2. We consider multiple evaluation metrics for these new tasks, and we adopt a new cross-modal retrieval similarity metric based on Text2Mol (Edwards et al., 2021).
3. We propose **MolT5**: a self-supervised learning framework for jointly training a model on molecule string representations and natural language text, which can then be finetuned on a cross-modal task.

2 Tasks

With the ambitious goal of bi-directional translation between molecules and language, we propose two new novel tasks: molecule captioning (Section 2.1) and text-based molecule generation (Section 2.2).

2.1 Molecule Captioning

For any given molecule, the goal of molecule captioning is to describe the molecule and what it does. An example is shown in Figure 2. Molecules are often represented as SMILES strings (Weininger, 1988; Weininger et al., 1989), a linearization of the molecular graph which can be interpreted as a language for molecules. Thus, this task can be considered an exotic translation task, and sequence to sequence models serve as excellent baselines.

2.2 Text-Based de Novo Molecule Generation

The goal of the de novo molecule generation task is to train a model which can generate a variety of possible new molecules. Existing work tends to focus on evaluating the model coverage of the chemical space (Polykovskiy et al., 2020). Instead, we propose generating molecules based on a natural language description of the desired molecule– this is essentially swapping the input and output for the captioning task. An example of this task is shown in Figure 1. Recent work, such as DALL·E (Ramesh et al., 2021, 2022), which generates images from text, has shown the ability to seamlessly integrate multiple properties, such as chairs and avocados, in an image. This points towards similar applications in the molecule generation domain via the usage of natural language.

3 Evaluation Metrics

3.1 Text2Mol Metric

Since we are considering new cross-modal tasks between molecules and text, we also introduce a new cross-modal evaluation metric. This is based on Text2Mol (Edwards et al., 2021), which aims to train a retrieval model to rank molecules given their text descriptions. Since the ranking function uses cosine similarity between embeddings, a trained model can be repurposed for evaluating the similarity between the ground truth molecule/description and the generated description/molecule (respectively). To this end, we first train a base multi-layer perceptron (MLP) model from Text2Mol. This model is then used to generate similarities of the candidate molecule-description pairs, which can be compared to the average similarity of the ground truth molecule-description pairs. We also note that negative molecule-description pairs have an average similarity of roughly zero.

3.2 Evaluating Molecule Captioning

Traditionally, captioning tasks have been evaluated by natural language generation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). Unlike captioning tasks such as COCO (Chen et al., 2015), which has several captions per image, in our task we only have one reference caption. This makes these metrics less effective, especially because there are many non-overlapping ways to describe a molecule. Nevertheless, for comparison, we still report these scores (e.g., aggregated sentence-level METEOR scores).

3.3 Evaluating Text-Based de Novo Molecule Generation

Considerable interest has grown in applying deep generative models to de novo molecule generation. Because of this, a number of metrics have been proposed, such as novelty and scaffold similarity (Polykovskiy et al., 2020). However, many of these metrics do not apply to our problem– we want our generated molecule to match the input text instead of being generally diverse. Instead, we consider metrics which measure the distance of the generated molecule to either the ground truth molecule or the ground truth description, such as our proposed Text2Mol-based metric.

We employ three fingerprint metrics: MACCS FTS, RDK FTS, and Morgan FTS, where FTS stands for fingerprint Tanimoto similarity (Tanimoto, 1958). MACCS (Durant et al., 2002), RDK (Schneider et al., 2015), and Morgan (Rogers and Hahn, 2010) are each fingerprinting methods for molecules. The fingerprints of two molecules are compared using Tanimoto similarity (also known as Jaccard index), and the average similarity over the evaluation dataset is reported. See (Campos and Ji, 2021) for more details. We also report exact SMILES string matches, Levenshtein distance (Miller et al., 2009), and SMILES BLEU scores.

Preuer et al. (2018) propose Fréchet ChemNet Distance (FCD), which is inspired by the Fréchet Inception Distance (FID) (Heusel et al., 2017). FCD is based on the penultimate layer of a network called "ChemNet", which was trained to predict the activity of drug molecules. Thus, FCD takes into account chemical and biological information about molecules in order to compare them. This allows molecules to be compared based on the latent information required to predict useful properties rather than a string-based metric.

In the case of models which use SMILES strings, generated molecules can be syntactically invalid. Therefore, we also report validity as the percent of molecules which can be processed by RDKIT

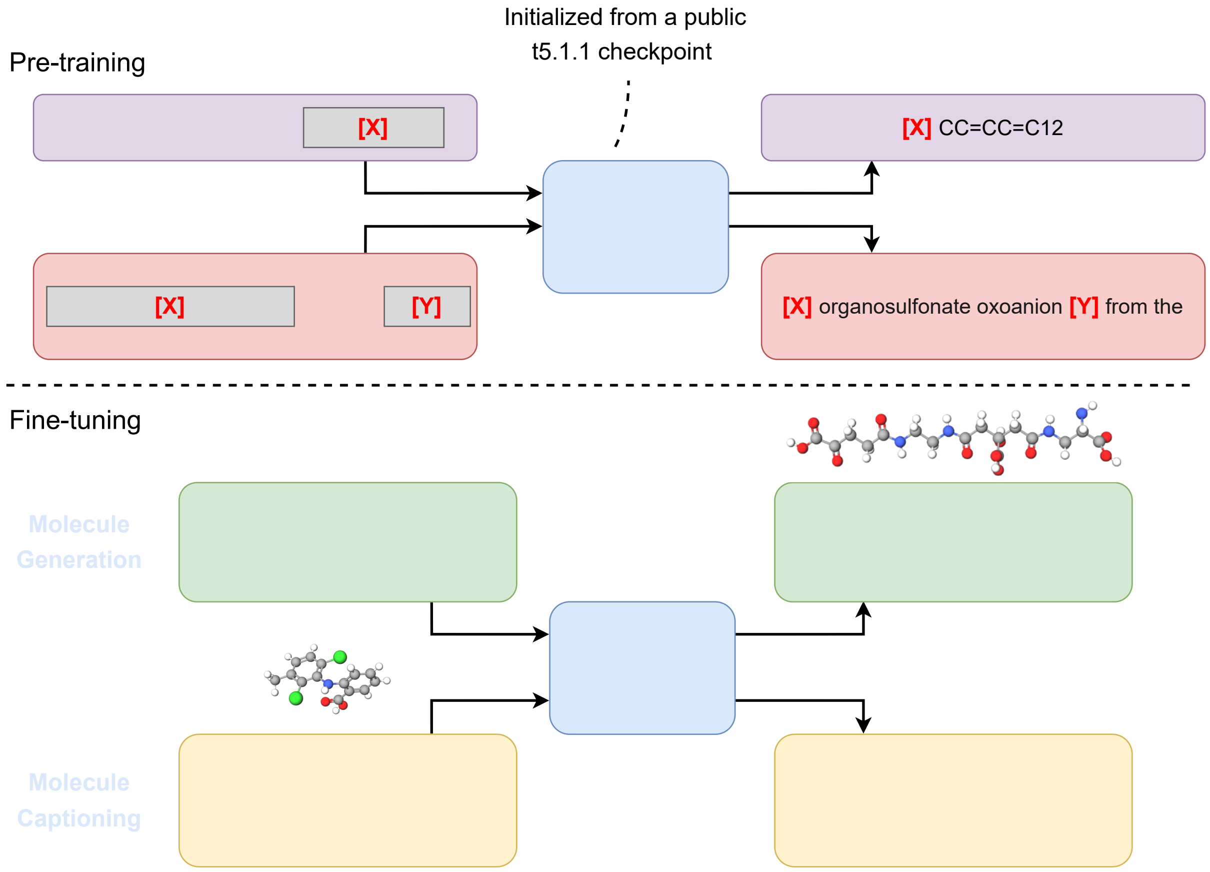

Pre-training
Initialized from a public t5.1.1 checkpoint
[X]
[X] CC=CC=C12
[X]
[Y]
[X] organosulfonate oxoanion [Y] from the
Fine-tuning
Molecule Generation
Molecule Captioning

Figure 3: A diagram of our framework. We first pre-train MolT5 on a large amount of data of both SMILES string and natural language using the "replace corrupted spans" objective (Raffel et al., 2020). After the pre-training stage, MolT5 can be easily fine-tuned for either the task of molecule captioning or generation (or both).

(Landrum, 2021) as in (Polykovskiy et al., 2020).

4 MolT5 – Multimodal Text-Molecule Representation Model

We can crawl a massive amount of text from the Internet. For example, Raffel et al. (2020) built a Common Crawl-based dataset that contains over 700 GB of reasonably clean and natural English text. On the other hand, over a billion molecules are also available from public databases such as ZINC-15 (Sterling and Irwin, 2015a). Inspired by the progress in large-scale pretraining (Ramesh et al., 2021), we propose a new self-supervised learning framework named **MolT5** (**Mol**ecular **T5**) to leverage the vast amount of unlabeled natural language text and molecule strings.

Figure 3 shows an overview of MolT5. We first initialize an encoder-decoder Transformer model (Vaswani et al., 2017) using one of the public checkpoints of T5.1.1[2], an improved version of T5 (Raffel et al., 2020). After that, we pretrain the model using the "replace corrupted spans" objective (Raffel et al., 2020). More specifically, during each pretraining step, we sample a minibatch comprising both natural language sequences and SMILES sequences. For each sequence, some words in the sequence are randomly chosen for corruption. Each consecutive span of corrupted tokens is replaced by a sentinel token (shown as [X] and [Y] in Figure 3). Then the task is to predict the dropped-out spans.[3]

Molecules (e.g. represented as SMILES strings) can be thought of as a language with a very unique grammar. Then, intuitively, our pretraining stage essentially trains a single language model on two monolingual corpora from two different languages, and there is no explicit alignment between the two corpora. This approach is similar to how some multilingual language models such as mBERT (Devlin et al., 2019) and mBART (Liu et al., 2020) were pretrained. As models such as mBERT demonstrate ex-

[2] `https://tinyurl.com/t511-ckpts`

[3] For more explanation of the pretraining task, we refer the readers to the original T5 paper (Raffel et al., 2020).

cellent cross-lingual capabilities (Pires et al., 2019), we also expect models pretrained using MolT5 to be useful for text-molecule translation tasks.

After the pretraining process, we can finetune the pretrained model for either molecule captioning or generation (depicted by the bottom half of Figure 3). In molecule generation, the input is a description, and the output is the SMILES representation of the target molecule. On the other hand, in molecule captioning, the input is the SMILES string of some molecule, and the output is a caption describing the input molecule.

5 Experiments and Results

5.1 Data

Pretraining Data As described in Section 4, the pretraining stage of MolT5 requires two monolingual corpora: one consisting of natural language text and the other consisting of molecule representations. We use the "Colossal Clean Crawled Corpus" (C4) (Raffel et al., 2020) as the pretraining dataset for the textual modality. For the molecular modality, we directly utilize the 100 million SMILES strings used in Chemformer (Irwin et al., 2021). As these strings were selected from the ZINC-15 dataset (Sterling and Irwin, 2015b), we refer to this pretraining dataset as ZINC from this point.

Finetuning and Evaluation Data We use ChEBI-20 (Edwards et al., 2021) as our gold standard dataset for finetuning and evaluation. It consists of 33,010 molecule-description pairs, which are separated into 80/10/10% train/validation/test splits. We use ChEBI-20 to finetune MolT5-based models and to train baseline models. Many captions in ChEBI-20 contain a name for the molecule at the start of the string (e.g., "Rostratin D is an organic disulfide isolated from ..."). To force the models to focus on the semantics of the description, we replace the molecule's name with "The molecule is [...]" (e.g., "The molecule is an organic disulfide isolated from ...").

5.2 Baselines

Any sequence-to-sequence model is applicable to our new tasks (i.e., molecule captioning and generation). We implement the following baselines:

1. **RNN-GRU** (Cho et al., 2014). We implement a 4-layer GRU recurrent neural network. The encoder is bidirectional.

2. **Transformer** (Vaswani et al., 2017). We train a vanilla Transformer model consisting of six encoder and decoder layers.

3. **T5** (Raffel et al., 2020). We experiment with three public T5.1.1 checkpoints[4]: small, base, and large. We finetune each checkpoint for molecule captioning or molecule generation using the t5x framework (Roberts et al., 2022).

We train the baseline models on ChEBI-20 using SMILES representations for the molecules. Molecule captioning and generation are trained with molecules as input/output and text as output/input. More information about the baselines and the hyperparameters is in the appendix.

5.3 Pretraining Process

We first initialize an encoder-decoder Transformer model using a public checkpoint of T5.1.1 (either *t5.1.1.small*, *t5.1.1.base*, or *t5.1.1.large*). We then pretrain the model on the combined dataset of C4 and ZINC (i.e., C4+ZINC) for 1 million steps. Each step uses a batch size of 256 evenly split between text and molecule sequences. After this, we finetune the pretrained model on ChEBI-20 for either molecule captioning or generation. The number of finetuning steps is 50,000.

5.4 Molecule Captioning

Table 1 shows the overall molecule captioning results. The pretrained models, either T5 or MolT5, are considerably better at generating realistic language to describe a molecule than the RNN and Transformer baselines. The RNN is more capable of extracting relevant properties from molecules than the Transformer, but it generally produces ungrammatical outputs. On the other hand, the Transformer produces grammatical outputs, but they tend to repeat the same properties, such as carcinogenic, regardless of whether they apply. For this reason, the Text2Mol scores are much lower for the Transformer model, since its outputs match the given molecule much less frequently. We speculate that the ChEBI-20 dataset is too small to effectively train a Transformer without large-scale pretraining. We find that our additional pretraining of MolT5 results in a reasonable increase over T5 in captioning performance on both the traditional NLG metrics and our Text2Mol metric for each model size. Finally, we refer the reader to Section H in

[4] https://tinyurl.com/t511-ckpts

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							0.609
RNN	0.251	0.176	0.450	0.278	0.394	0.363	0.426
Transformer	0.061	0.027	0.204	0.087	0.186	0.114	0.057
T5-Small	0.501	0.415	0.602	0.446	0.545	0.532	0.526
MolT5-Small	0.519	0.436	0.620	0.469	0.563	0.551	0.540
T5-Base	0.511	0.423	0.607	0.451	0.550	0.539	0.523
MolT5-Base	0.540	0.457	0.634	0.485	0.578	0.569	0.547
T5-Large	0.558	0.467	0.630	0.478	0.569	0.586	0.563
MolT5-Large	**0.594**	**0.508**	**0.654**	**0.510**	**0.594**	**0.614**	**0.582**

Table 1: Molecule captioning results on the test split of CheBI-20. Rouge scores are F1 values.

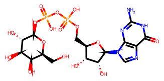
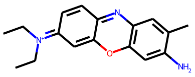
	Input	RNN	Transformer	T5	MolT5	Ground Truth
1		the molecule is a gdp - d - glucoside - a - [...]	the molecule is the stable isotope of helium with relative atomic mass 3. 016029. the least abundant (0. 000137 atom percent) isotope of naturally occurring helium.	The molecule is a GDP-D-glucose in which the anomeric centre of the pyranose fragment has alpha-configuration. It is a GDP-D-glucose and a ribonucleoside 5'-diphosphate-alpha-D-glucose. It is a conjugate acid of a GDP-alpha-D-glucose(2-).	The molecule is a GDP-L-galactose in which the anomeric oxygen is on the same side of the fucose ring as the methyl substituent. It has a role as a plant metabolite and a mouse metabolite. It is a conjugate acid of a GDP-beta-L-galactose(2-).	The molecule is a GDP-L-galactose having beta-configuration at the anomeric centre of the L-galactose fragment. It is a conjugate acid of a GDP-beta-L-galactose(2-).
2	^{39}Ar	the molecule is stable metallic [...]	the molecule is the stable isotope of thallium with relative atomic mass 202. 9723. the least abundant (29. 524 atom percent) isotope of naturally occurring thallium.	The molecule is the radioactive isotope of chromium with relative atomic mass 39.98286 and half-life of 138.376 days; the only naturally occurring isotope of chromium.	The molecule is the stable isotope of rubidium with relative atomic mass 44.955910, 100 atom percent natural abundance and nuclear spin 7/2.	The molecule is a trace radioisotope of argon with atomic mass of 38.964313 and a half-life of 269 years. It has a role as an isotopic tracer.
3		the molecule is a cationic fluorescent dye having 2, 3 - dimethyl - 1, 2, 3, 4, 6 - tetrahydro - 1h - 1, 2, 3, 4, 6 - tetrahydropyridin - 1 - yl] amino } amino group, respectively. it has a role as a fluorochrome.	the molecule is a deuterated compound that is is is is is an isotopologue of chloroform in which the four hydrogen atoms have been replaced by deuterium. it is a deuterated compound and an alpha, omega - dicarboxylic acid.	The molecule is a quaternary ammonium ion and a member of phenanthridines. It has a role as an intercalator and a fluorochrome.	The molecule is an organic cation that is phenoxazin-5-ium substituted by amino and methylamino groups at positions 3 and 7 respectively. The chloride salt is the histological dye 'azure C'.	The molecule is an organic cation that is phenoxazin-5-ium substituted by methyl, amino and diethylamino groups at positions 2, 3 and 7 respectively. The tetrachlorozincate salt salt is the histological dye 'brilliant cresyl blue'.

Figure 4: Example captions generated by different models.

the appendix for information about the statistical significance of our results.

Several examples of different models' outputs are shown in Figure 4 and Appendix Figure 9. In (1), MolT5's description matches best, identifying the molecule as a "GDP-L-galactose". MolT5 is usually able to recognize what general class of molecule it is looking at (e.g. cyclohexanone, maleate salt, etc.). In general, all models often look for the closest compound they know and base their caption on that. The argon atom, example (2) with SMILES '[39Ar]', is not present in the training dataset bonded to any other atoms (likely because it is an inert noble gas). All models recognize that (2) is a single atom, but they are unable to describe it. In (3), the models try to caption a histological dye. MolT5 captions the molecule as an azure histological dye, which is very close to the ground truth "brilliant cresyl blue", while T5 does not.

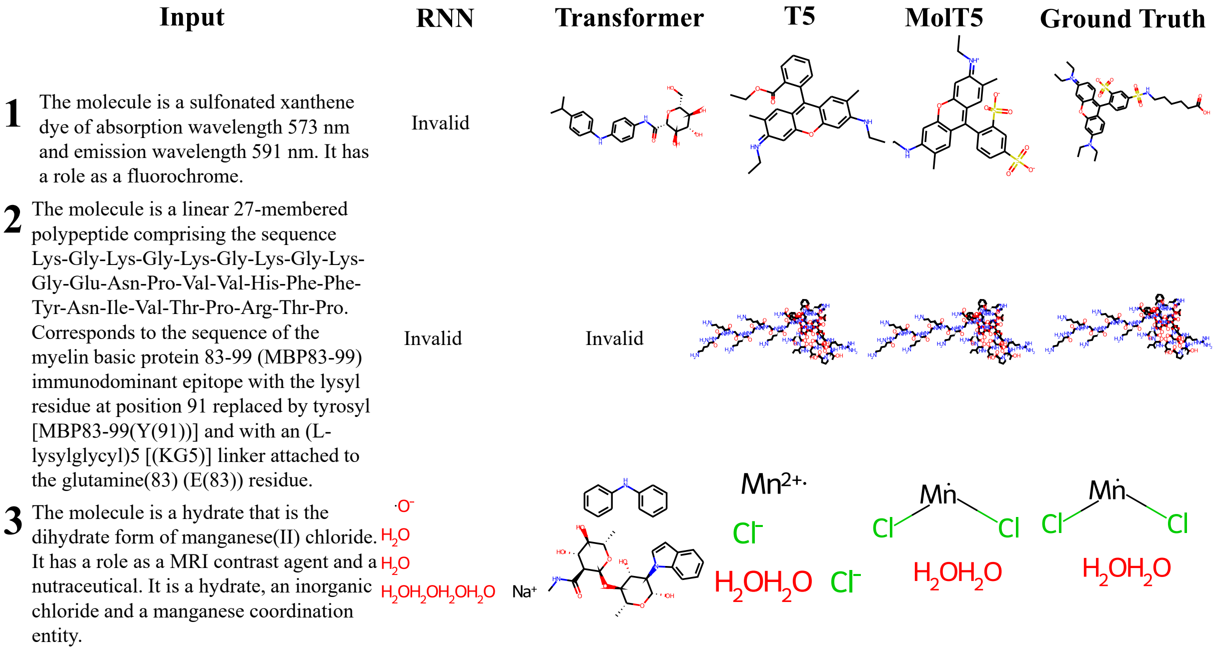
5.5 Text-Based de novo Molecule Generation

In the molecule generation task, the pretrained models also perform much better than the RNN and Transformer (Table 2). Although it is well known that scaling model size and pretraining data leads to significant performance increases (Kaplan et al., 2020), it was still surprising to see the results. For example, a default T5 model, which was only pretrained on text data, is capable of generating molecules which are much closer to the ground truth than the RNN and which are often valid. This trend also persists as language model size scales, since T5-large with 770M parameters outperforms the specifically pretrained MolT5-small with 60M parameters. Still, the pretraining in MolT5 slightly improves some molecule generation results, with especially large gains in validity. Finally, Section H in the appendix has information about the statistical significance of our results.

We show results for the models in Figure 5 and

Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth	1.000	1.000	0.0	1.000	1.000	1.000	0.0	0.609	1.0
RNN	0.652	0.005	38.09	0.591	0.400	0.362	4.55	0.409	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	11.32	0.277	**0.906**
T5-Small	0.741	0.064	27.703	0.704	0.578	0.525	2.89	0.479	0.608
MolT5-Small	0.755	0.079	25.988	0.703	0.568	0.517	2.49	0.482	0.721
T5-Base	0.762	0.069	24.950	0.731	0.605	0.545	2.48	0.499	0.660
MolT5-Base	0.769	0.081	24.458	0.721	0.588	0.529	2.18	0.496	0.772
T5-Large	0.854	0.279	16.721	0.823	0.731	0.670	1.22	0.552	0.902
MolT5-Large	**0.854**	**0.311**	**16.071**	**0.834**	**0.746**	**0.684**	**1.20**	**0.554**	0.905

Table 2: Molecule generation results on the test split of CheBI-20. Except for BLEU, Exact, Levenshtein, and Validity, other metrics are computed using only syntactically valid molecules, as in (Campos and Ji, 2021).



1 The molecule is a sulfonated xanthene dye of absorption wavelength 573 nm and emission wavelength 591 nm. It has a role as a fluorochrome.

2 The molecule is a linear 27-membered polypeptide comprising the sequence Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Glu-Asn-Pro-Val-Val-His-Phe-Phe-Tyr-Asn-Ile-Val-Thr-Pro-Arg-Thr-Pro. Corresponds to the sequence of the myelin basic protein 83-99 (MBP83-99) immunodominant epitope with the lysyl residue at position 91 replaced by tyrosyl [MBP83-99(Y(91))] and with an (L-lysylglycyl)5 [(KG5)] linker attached to the glutamine(83) (E(83)) residue.

3 The molecule is a hydrate that is the dihydrate form of manganese(II) chloride. It has a role as a MRI contrast agent and a nutraceutical. It is a hydrate, an inorganic chloride and a manganese coordination entity.

Figure 5: Examples of molecules generated by different models.

also in Figures 6, 7, and 8 in Appendix F, which we number by input description. Compared to T5, MolT5 is better able to understand instructions for manipulating molecules, as shown in examples (3, 4, 6, 7, 16, 18, 21). In many cases, MolT5 obtains exact matches with the ground truth (2, 3, 4, 6, 7, 8, 10, 12, 17, 20, 21). (3) is an interesting case, since it shows that MolT5 can understand crystalline solids like hydrates. (2) is another interesting example; it is the longest SMILES string, at 474 characters, which MolT5 is able to generate an exact match for. MolT5 understands peptides and can produce them from descriptions (2,15,17). It also shows this ability for saccharides (6, 21) and enzymes (8,20). MolT5 is able to understand rare atoms such as Ruthenium (5). However, in this case it still misses the atom's charge. Some example descriptions, such as (1), lack details so the molecules generated by MolT5 may be interesting to investigate.

5.6 Probing the Model

We conduct probing tests on the model for certain input properties, which are shown in Appendix J. Often, the model will generate molecules that it knows matches the input description from the finetuning data. It also creates solutions from these as well by adding various ions (e.g. ".[Na+]"). In some cases, it generates molecules not appearing in finetuning data (sometimes successfully sometimes not). For example, given the input "The molecule is a corticosteroid.", the first molecule generated is a well known corticosteroid called corticosterone. The fifth molecule generated is not present in the PubChem database. Based on a structure similarity search, it is most closely related to the androgenic steroid Fluoxymesterone and the corticosteroid Hydrocortisone.

6 Related Work

6.1 Multimedia Representation

Much recent work on multimedia representations falls into training large vision-language models (Su et al., 2020; Lu et al., 2019; Chen et al., 2020). CLIP (Radford et al., 2021) trains a zero-shot image classifier by using natural language labels which can be easily extended. A modification of CLIP's contrastive loss function, which follows (Sohn, 2016), is applied by Text2Mol (Edwards et al., 2021) for cross-modal retrieval between molecule and text pairs. Edwards et al. (2021) also released the ChEBI-20 dataset of molecule-description pairs, which is used for training and evaluation in this paper. Vall et al. (2021) leverage a contrastive loss between bioassay descriptions and molecules to predict activity between the two. Sun et al. (2021) uses cross-modal attention with molecule structures to improve chemical entity typing. Zeng et al. (2022) pretrain a language model to learn a joint representation between molecules and biomedical text via entity linking which they use for tasks such as relation extraction, molecule property prediction, and cross-modal retrieval like Text2Mol. Unlike our work, they do not explore generating text nor molecules. Vaucher et al. (2020) create a dataset of chemical equations and associated action sequences in natural language. Vaucher et al. (2021) then leverage this dataset to train a BART model which can plan chemical reaction steps. Their natural language generation is constrained to the specific reaction steps in their dataset– the main purpose of their model is to create the steps for a reaction rather than describing molecules.

6.2 Image Captioning and Text-Guided Image Generation

Image captioning has been studied extensively (Pan et al., 2004; Lu et al., 2018; Hossain et al., 2019; Stefanini et al., 2021). Many recent studies tend to pretrain Transformer-based models on massive text-image corpora (Li et al., 2020; Hu et al., 2022). Work has also been done in the biomedical domain (Pavlopoulos et al., 2019), a close cousin of the chemistry domain, where tasks tend to be focused on diagnosis of various image types such as x-rays (Demner-Fushman et al., 2016).

The reverse problem, text-guided image generation, has proven considerably more challenging (Khan et al., 2021). Several attempts have used GAN-based methods (Reed et al., 2016; Zhang et al., 2017; Xu et al., 2018). Recent work has shown remarkable results. DALL· E (Ramesh et al., 2021, 2022) can seamlessly fuse multiple concepts together to generate a realistic image.

6.3 Molecule Representation

Molecule representation has been a long-standing problem in the field of cheminformatics. Traditionally, fingerprinting methods have been a preferred technique to featurize molecule structural representations (Rogers and Hahn, 2010; Cereto-Massagué et al., 2015). These approaches do not allow representations to be learned from data. In recent years, advances in machine learning and NLP have been applied to this problem. A popular input for these algorithms has been SMILES strings (Weininger, 1988; Weininger et al., 1989), which are a computer-readable linearization of molecule graphs. Jaeger et al. (2018) use the Morgan fingerprinting algorithm to convert each molecule into a 'sentence' of its substructures, to which it applies the Word2vec algorithm (Mikolov et al., 2013a,b). Duvenaud et al. (2015) use neural methods to learn fingerprints. Other advances such as BERT (Devlin et al., 2019) have also been applied to the domain, such as MolBERT (Fabian et al., 2020) and ChemBERTa (Chithrananda et al., 2020), which use SMILES strings as inputs to pretrain a BERT-esque model. Work has been done to use the molecule graph structure and known reactions for learning representations (Wang et al., 2022). Schwaller et al. (2021b) trains a BERT model to learn representations of chemical reactions. Schwaller et al. (2021a) leverages unsupervised representation learning with Transformers to extract an organic chemistry grammar. Unlike existing work, MolT5's molecule representations allow for translation between molecules and natural language.

There has been particular interest in training generative models for de novo molecule discovery. Bagal et al. (2021) apply a GPT-style decoder for this task. Lu and Zhang (2022) apply a T5 model to SMILES strings for multitask reaction prediction problems. MegaMolBART[5] trains a BART model on 500M SMILES strings from the ZINC-15 dataset (Sterling and Irwin, 2015b)

[5] https://tinyurl.com/megamolbart

7 Conclusions and Future Work

In this work, we propose **MolT5**, a self-supervised learning framework for pretraining models on a vast amount of unlabeled text and molecule strings. Furthermore, we propose two new tasks: molecule captioning and text-guided molecule generation, for which we explore various evaluation methods. Together, these tasks allow for translation between natural language and molecules. Using **MolT5**, we are able to obtain high scores for both tasks.

8 Broader Impacts

Our proposed model and tasks will have the following broader impacts. 1) It will help to democratize molecular AI, allowing chemistry experts to take advantage of new AI technologies for discovering new life-changing drugs by interacting in the natural language, because it is most natural for humans to provide explanations and requirements in natural language. 2) Text-based molecule generation enables the ability to generate molecules with specific functions (such as taste) rather than properties, enabling the next generation of chemistry where custom molecules are used for each application. Specifically-designed molecular solutions have the potential to revolutionize fields such as medicine and material science. 3) Our models, whose weights we will release, will allow further research in the NLP community on the applications of multimodal text-molecule models.

8.1 Risks

MolT5, like other large language models, can potentially be abused. First, there may be biases learned by the model due to its large-scale training data. These biases may affect what type of molecules are generated when the model is prompted about certain diseases. Thus, any molecules discovered by usage of MoLT5 should strictly evaluated by standard clinical processes before being considered for medicinal use. Another risk is that the model may be used to discover potentially dangerous molecules instead of beneficial ones. It is difficult to predict what exact molecules may be discovered via usage of our work. However, while there is this unfortunate potential for misuse of the technology, knowledge of dangerous molecule's existence and structure is generally not harmful due to the requisite technical knowledge and laboratory resources required to synthesize them in any meaningful quantity. Overall, we believe these downsides are outweighed by the benefits to the research and pharmaceutical communities.

9 Limitations

Since this work focuses on a new application for large language models, many of the same limitations apply here. Namely, the model is trained on a large dataset collected from the Internet, so it may contain unintended biases. One limitation of our model is using SMILES strings – recent work (Krenn et al., 2020) proposes a string representation with validity guarantees. In practice, we found this to work poorly with pretrained T5 checkpoints (which were important from a computational perspective). We also note that some compounds in ChEBI-20 can cause validity problems in the default SELFIES implementation. We leave further investigation of this to future work. Finally, we stress that MolT5 was created for research purposes and generated molecules should not be used for medical purposes without careful evaluation by standard clinical testing first.

Acknowledgement

We would like to thank Martin Burke for his helpful discussion. This research is based upon work supported by the Molecule Maker Lab Institute: an AI research institute program supported by NSF under award No. 2019897 and No. 2034562. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. 2021. Molgpt: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Daniel Campos and Heng Ji. 2021. Img2smi: Translating molecular structure images to simplified molecular-input line-entry system. *arXiv preprint arXiv:2109.04202*.

Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.

Seyone Chithrananda, Gabe Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. 2002. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

Thomas Gaudelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy BR Hayter, Richard Vickers, Charles Roberts, Jian Tang, et al. 2021. Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics*, 22(6):bbab159.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17980–17989.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Bjerrum. 2021. Chemformer: A pre-trained transformer for computational chemistry. *ChemRxiv*.

Sabrina Jaeger, Simone Fulle, and Samo Turk. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Greg Landrum. 2021. Rdkit: Open-source cheminformatics software.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. Entity-aware image caption generation. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP2018)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of Chemical Information and Modeling*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Frederic P Miller, Agnes F Vandome, and John McBrewster. 2009. Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance.

Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. 2004. Automatic image captioning. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 3, pages 1987–1990. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A survey on biomedical image captioning. In *Proceedings of the second workshop on shortcomings in vision and language*, pages 26–36.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. 2020. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:1931.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2018. Fréchet chemnet distance: A metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58 9:1736–1741.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.

Ahmet Sureyya Rifaioglu, Heval Atas, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Doğan. 2018. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in Bioinformatics*, 20(5):1878–1912.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

Nadine Schneider, Roger A. Sayle, and Gregory A. Landrum. 2015. Get your atoms in order - an open-source implementation of a novel and robust molecular canonicalization algorithm. *Journal of chemical information and modeling*, 55 10:2111–20.

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. 2021a. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166.

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. 2021b. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912*.

T. Sterling and John J. Irwin. 2015a. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55:2324 – 2337.

Teague Sterling and John J Irwin. 2015b. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chenkai Sun, Weijiang Li, Jinfeng Xiao, Nikolaus Nova Parulian, ChengXiang Zhai, and Heng Ji. 2021. Fine-grained chemical entity typing with multimodal knowledge representation. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1984–1991. IEEE.

Taffee T Tanimoto. 1958. Elementary mathematical theory of classification and prediction.

Andreu Vall, Sepp Hochreiter, and Günter Klambauer. 2021. Bioassayclr: Prediction of biological activity for novel bioassays based on rich textual descriptions. *ELLIS Machine Learning for Molecule Discovery Workshop*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alain C Vaucher, Philippe Schwaller, Joppe Geluykens, Vishnu H Nair, Anna Iuliano, and Teodoro Laino. 2021. Inferring experimental procedures from text-based representations of chemical reactions. *Nature communications*, 12(1):1–11.

Alain C Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures. *Nature communications*, 11(1):1–11.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin Burke. 2022. Chemical-reaction-aware molecule representation learning. In *Proc. The International Conference on Learning Representations (ICLR2022)*.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

David Weininger, Arthur Weininger, and Joseph L Weininger. 1989. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):1–11.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

A Baselines and Hyperparameters

Any sequence-to-sequence model is applicable to our new tasks (i.e., molecule captioning and generation). We implement the following baselines:

1. **RNN-GRU** (Cho et al., 2014). We implement a 4-layer GRU recurrent neural network with a hidden size of 512. We use a learning rate of 1e-4 and a batch size of 128 for molecule generation. For caption generation, a batch size of 116 is used. The number of training epochs is 50. Additionally, the encoder is bidirectional. For training, teacher forcing is used 50% of the time, and gradient clipping to 50 is applied.

2. **Transformer** (Vaswani et al., 2017). We train a vanilla Transformer model consisting of six encoder and decoder layers. The number of training epochs is 40, the batch size is 16, and the learning rate is 1e-4. We use a linear decay with a warmup of 400 steps.

3. **T5** (Raffel et al., 2020). We experiment with three public T5.1.1 checkpoints[6]: small, base, and large. We finetune each checkpoint for molecule captioning or molecule generation using the open-sourced t5x framework (Roberts et al., 2022). The number of training steps is set to be 50,000. The dropout rate is set to be 0.0 for the small and base models, and it is set to be 0.1 for the large model. For other hyperparameters, we use the default values provided by the t5x framework.

We train the baseline models on the ChEBI-20 dataset using SMILES representations for the molecules. Molecule captioning and generation are trained with molecules as input/output and text as output/input. Sequences are limited to 512 tokens for input and output. During inference, a beam decoder with a beam size of 5 is used.

On the RNN and vanilla Transformer models, we use a character-split vocabulary for SMILES. For the text vocabulary, we use SciBERT's 31,090-token vocabulary (Beltagy et al., 2019).

B Reproducibility Checklist

The programs, trained models, and resources will be made publicly available. For training the RNN and Transformer baselines, we use NVIDIA Tesla V100 GPUs. For pretraining and finetuning T5-related models, we use TPUs.

When testing on a MacBook Pro that has no access to GPUs, the average inference time of our MolT5-Base molecule generation model is 2.24 seconds/query. The average inference time of our large MolT5-Base molecule captioning model is 9.86 seconds/query.

C Decoding with Huggingface Model

For ease of adoption, we converted our original models trained using the t5x framework (Roberts et al., 2022) to HuggingFace-based models (Wolf et al., 2019). We will release the converted models on HuggingFace (HF) Hub. Due to implementation differences, the HF-based models produce slightly different outputs from the original models. Therefore, we also report the numbers of the HF-based models in Table 3 and Table 4.

D High Validity Molecule Generation

To increase the validity score of the molecule generation models, we consider a high-validity decoding strategy. We use diverse beam search (Vijayakumar et al., 2016) with a beam width and beam group of 30 and a diversity penalty of 0.5. Then, we use RDKit (Landrum, 2021) to select the first valid beam. On rare occasions, the beam size exceeds memory limitations, so we iteratively reduce the beam size by 5 for that input and try again. In Table 4, MolT5-Small-HV, MolT5-Base-HV, and MolT5-Large-HV denote models that use this decoding process.

E Ablations

We perform ablations on MolT5-Small pretraining. For molecule captioning (Table 5), pretraining on both C4 and ZINC is clearly more beneficial than pretraining only on C4 or only on ZINC.

For molecule generation, at first glance, pretraining on C4+ZINC seems not to outperform pretraining only on C4 (Table 6). However, note that except for BLEU, Exact, Levenshtein, and Validity, other metrics in Table 6 are computed using only syntactically valid molecules. Table 7 shows the normalized molecule generation results. After normalization, we see that pretraining on C4+ZINC outperforms pretraining only on C4 or only on ZINC according to most metrics. Finally, pretraining only on ZINC increases the validity score substantially. However, this leads to decreased similarity of the generated molecules to the ground truths.

[6] https://tinyurl.com/t511-ckpts

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							0.609
RNN	0.251	0.176	0.450	0.278	0.394	0.363	0.426
Transformer	0.061	0.027	0.204	0.087	0.186	0.114	0.057
T5-Small	0.515	0.424	0.613	0.459	0.568	0.538	0.527
MolT5-Small	0.532	0.445	0.627	0.477	0.583	0.557	0.543
T5-Base	0.522	0.432	0.616	0.461	0.572	0.545	0.524
MolT5-Base	0.551	0.464	0.637	0.489	0.594	0.574	0.549
T5-Large	0.555	0.464	0.632	0.482	0.585	0.588	0.564
MolT5-Large	**0.588**	**0.502**	**0.650**	**0.507**	**0.604**	**0.614**	**0.582**

Table 3: HuggingFace model molecule captioning results for the different baseline models on the test split of CheBI-20. Rouge scores are F1 values.

Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth	1.000	1.000	0.0	1.000	1.000	1.000	0.0	0.609	1.0
RNN	0.652	0.005	38.09	0.591	0.400	0.362	4.55	0.409	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	11.32	0.277	0.906
T5-Small	0.740	0.061	30.05	0.798	0.681	0.623	1.77	0.541	0.597
MolT5-Small	0.749	0.082	28.816	0.780	0.654	0.601	1.35	0.535	0.725
MolT5-Small-HV	0.613	0.075	30.458	0.699	0.547	0.482	1.44	0.479	0.983
T5-Base	0.769	0.067	27.112	0.816	0.701	0.637	1.44	0.554	0.654
MolT5-Base	0.783	0.082	24.846	0.788	0.661	0.602	1.16	0.544	0.787
MolT5-Base-HV	0.661	0.073	28.276	0.721	0.579	0.509	1.38	0.501	0.979
T5-Large	0.856	0.285	16.845	0.877	0.794	0.732	0.40	0.587	0.959
MolT5-Large	**0.858**	**0.318**	**15.957**	**0.890**	**0.813**	**0.750**	**0.38**	**0.590**	0.958
MolT5-Large-HV	0.810	0.314	16.758	0.872	0.786	0.722	0.44	0.582	**0.996**

Table 4: HuggingFace model de novo molecule generation results for the different baseline models on the test split of CheBI-20. MolT5-Small-HV, MolT5-Base-HV, and MolT5-Large-HV are models that use a high-validity decoding process–see Appendix D.

Pretraining	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							0.609
C4-Only	0.523	0.433	0.616	0.463	0.571	0.545	0.530
ZINC-Only	0.519	0.434	0.619	0.466	0.573	0.548	0.538
C4+ZINC	0.532	0.445	0.627	0.477	0.583	0.557	0.543

Table 5: Pretraining ablation results of molecule captioning for MolT5-Small on the test split of CheBI-20. Rouge scores are F1 values.

Pretraining	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth							0.0	0.609	1.0
C4-Only	0.771	0.081	26.84	0.811	0.697	0.641	2.99	0.555	0.635
ZINC-Only	0.716	0.063	32.953	0.701	0.576	0.524	2.75	0.463	0.807
C4+ZINC	0.749	0.082	28.816	0.78	0.654	0.601	2.60	0.535	0.725

Table 6: Pretraining ablation results of molecule generation for MolT5-Small on the test split of CheBI-20.

Pretraining	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth							0.0	0.609	1.0
C4-Only	0.771	0.081	26.84	0.51499	0.44259	0.40704	4.71	0.35243	0.635
ZINC-Only	0.716	0.063	32.953	0.56571	0.46483	0.42287	3.41	0.37364	0.807
C4+ZINC	0.749	0.082	28.816	0.5655	0.47415	0.43572	3.59	0.38788	0.725

Table 7: Normalized pretraining ablation results of molecule generation for MolT5-Small on the test split of CheBI-20. Molecule-based results (FTS, FCD, Text2Mol) are normalized by multiplying by validity (for scores where higher is better) or dividing by validity (for scores where lower is better).

F More Examples

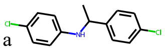
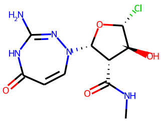
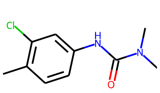
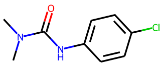
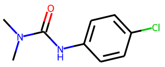
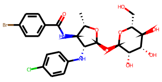
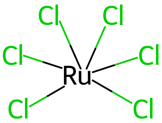
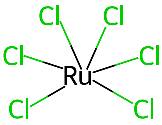
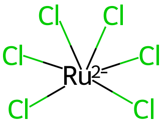
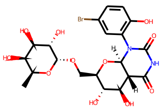
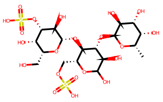
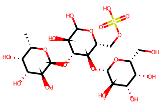
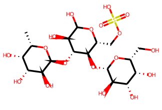
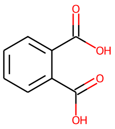
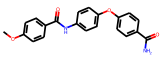
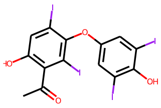
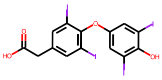
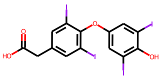
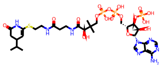
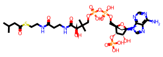
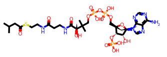
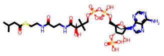
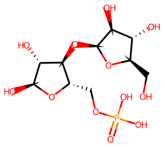
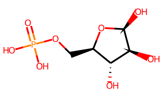
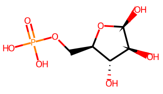
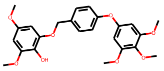
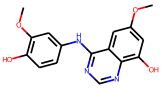
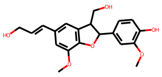
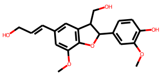
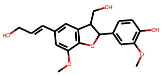
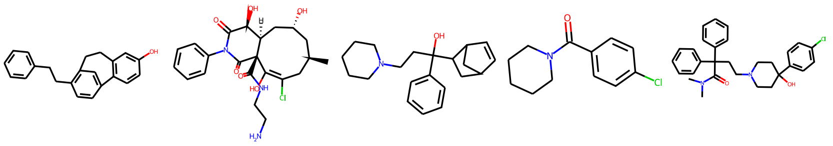
Input	RNN	Transformer	T5	MolT5	Ground Truth
4 The molecule is a member of the class of phhenylureas that is urea in which one of the nitrogens is substituted by a p-chlorophenyl group while the other is substituted by two methyl groups. It has a role as a herbicide, a xenobiotic and an environmental contaminant. It is a member of monochlorobenzenes and a member of phenylureas.					
5 The molecule is a perchlorometallate anion having six chlorines and ruthenium(IV) as the metal component. It is a perchlorometallate anion and a ruthenium coordination entity.	HCl HCl HCl				
6 The molecule is a trisaccharide derivative that consists of 6-sulfated D-glucose having an alpha-L-fucosyl residue attached at position 3 and a beta-D-galactosyl residue attached at position 4. It has a role as an epitope. It is a trisaccharide derivative and an oligosaccharide sulfate.	Invalid				
7 The molecule is a monocarboxylic acid that is thyroacetic acid carrying four iodo substituents at positions 3, 3', 5 and 5'. It has a role as a thyroid hormone, a human metabolite and an apoptosis inducer. It is an iodophenol, a 2-halophenol, a monocarboxylic acid and an aromatic ether.					
8 The molecule is a methylbutanoyl-CoA is the S-isovaleryl derivative of coenzyme A. It has a role as a mouse metabolite. It derives from an isovaleric acid and a butyryl-CoA. It is a conjugate acid of an isovaleryl-CoA(4-).	Invalid				
9 The molecule is an D-arabinose 5-phosphate that is beta-D-arabinofuranose attached to a phospahte group at position 5. It derives from a beta-D-arabinofuranose.	Invalid	Invalid			
10 The molecule is a guaiacyl lignin obtained by cyclodimerisation of coniferol. It has a role as a plant metabolite and an anti-inflammatory agent. It is a member of 1-benzofura primary alcohol, a guaiacyl lignin an member of guaiacols. It derives from a coniferol.					

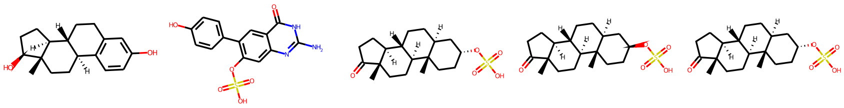
Figure 6: More examples of interesting molecules generated by different models.

Input **RNN** **Transformer** **T5** **MolT5** **Ground Truth**

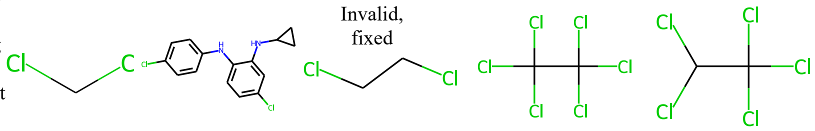
11 The molecule is a synthetic piperidine derivative, effective against diarrhoea resulting from gastroenteritis or inflammatory bowel disease. It has a ro as a mu-opioid receptor agonist, an antidiarrhoeal drug and an anticoronavi agent. It is a member of piperidines, a monocarboxylic acid amide, a member monochlorobenzenes and a tertiary alcohol. It is a conjugate base of a loperamide(1+).



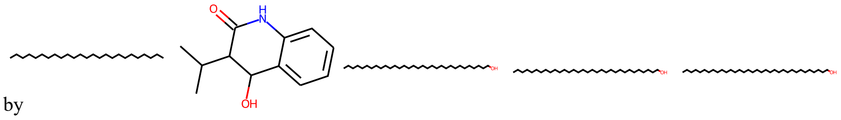
12 The molecule is a steroid sulfate that is the 3-sulfate of androsterone. It has a role as a human metabolite and a mouse metabolite. It is a 17-oxo steroid, a steroid sulfate and an androstanoid. It derives from an androsterone. It is a conjugate acid of an androsterone sulfate(1-). It derives from a hydride of a 5alpha-androstane.



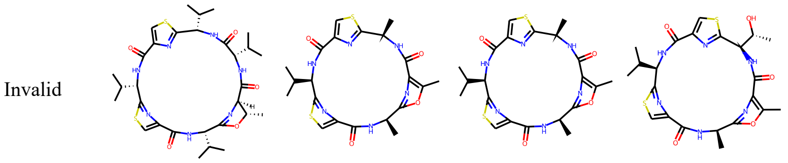
13 The molecule is a member of the class of chloroethanes that is ethane in which five of the six hydrogens are replaced by chlorines. A non-flammable, high-boiling liquid (b.p. 161-162°C) with relative density 1.67 and an odour resembling that of chloroform, it is used as a solvent for oil and grease, in metal cleaning, and in the separation of coal from impurities. It has a role as a non-polar solvent.



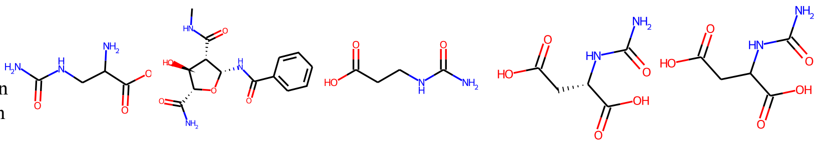
14 The molecule is an ultra-long-chain primary fatty alcohol that is tetratriacontane in which one of the terminal methyl hydrogens is replaced by a hydroxy group It has a role as a plant metabolite.



15 The molecule is an eighteen-membered homodetic cyclic peptide which is isolated from Oscillatoria sp. and exhibits antimalarial activity against the W2 chloroquine-resistant strain of the malarial parasite, Plasmodium falciparum. It has a role as a metabolite and an antimalarial. It is a homodetic cyclic peptide, a member of 1,3-oxazoles, a member of 1,3-thiazoles and a macrocycle.



16 The molecule is an N-carbamoylamino acid that is aspartic acid with one of its amino hydrogens replaced by a carbamoyl group. It has a role as a Saccharomyces cerevisiae metabolite, an Escherichia coli metabolite and a human metabolite. It is a N-carbamoyl-amino acid, an aspartic acid derivative and a C4-dicarboxylic acid. It is a conjugate acid of a N-carbamoylaspartate(2-).



17 The molecule is a tripeptide composed of glycine, glycine and L-alanine residues joined in sequence. It has a role as a metabolite.

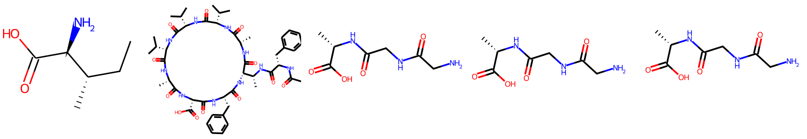
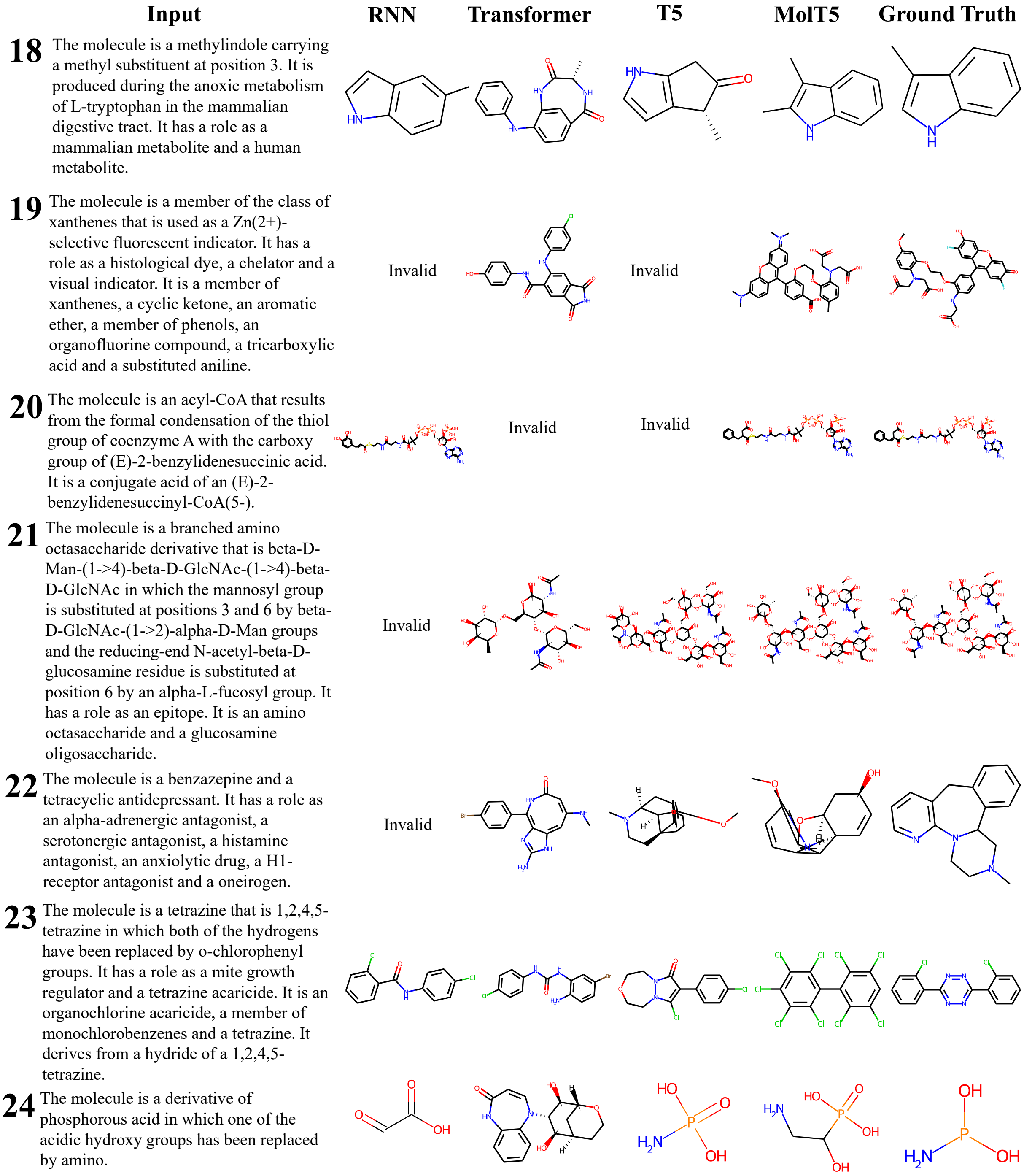


Figure 7: More examples of interesting molecules generated by different models.



	Input	RNN	Transformer	T5	MolT5	Ground Truth
18	The molecule is a methylindole carrying a methyl substituent at position 3. It is produced during the anoxic metabolism of L-tryptophan in the mammalian digestive tract. It has a role as a mammalian metabolite and a human metabolite.					
19	The molecule is a member of the class of xanthenes that is used as a Zn(2+)-selective fluorescent indicator. It has a role as a histological dye, a chelator and a visual indicator. It is a member of xanthenes, a cyclic ketone, an aromatic ether, a member of phenols, an organofluorine compound, a tricarboxylic acid and a substituted aniline.	Invalid		Invalid		
20	The molecule is an acyl-CoA that results from the formal condensation of the thiol group of coenzyme A with the carboxy group of (E)-2-benzylidenesuccinic acid. It is a conjugate acid of an (E)-2-benzylidenesuccinyl-CoA(5-).		Invalid	Invalid		
21	The molecule is a branched amino octasaccharide derivative that is beta-D-Man-(1->4)-beta-D-GlcNAc-(1->4)-beta-D-GlcNAc in which the mannosyl group is substituted at positions 3 and 6 by beta-D-GlcNAc-(1->2)-alpha-D-Man groups and the reducing-end N-acetyl-beta-D-glucosamine residue is substituted at position 6 by an alpha-L-fucosyl group. It has a role as an epitope. It is an amino octasaccharide and a glucosamine oligosaccharide.	Invalid				
22	The molecule is a benzazepine and a tetracyclic antidepressant. It has a role as an alpha-adrenergic antagonist, a serotonergic antagonist, a histamine antagonist, an anxiolytic drug, a H1-receptor antagonist and a oneirogen.	Invalid				
23	The molecule is a tetrazine that is 1,2,4,5-tetrazine in which both of the hydrogens have been replaced by o-chlorophenyl groups. It has a role as a mite growth regulator and a tetrazine acaricide. It is an organochlorine acaricide, a member of monochlorobenzenes and a tetrazine. It derives from a hydride of a 1,2,4,5-tetrazine.					
24	The molecule is a derivative of phosphorous acid in which one of the acidic hydroxy groups has been replaced by amino.					

Figure 8: More examples of interesting molecules generated by different models.

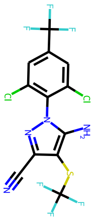
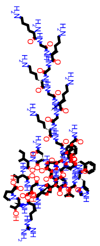
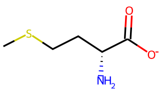
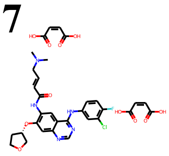
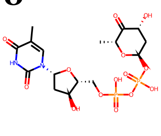
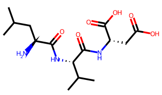
Input	RNN	Transformer	T5	MolT5	Ground Truth
4 	the molecule is an organofluorine compound that is 1, 2, 3, 4 - triazol - 1h - 1, 2, 4 - triazole which is substituted at positions 2, 3, and 5 by a 2, 3, 5 - triazol - 1 - yl group and at position 5 by a 2 - (trifluoromethyl) - 1, 3, 5 - triazol - 1 - yl group. it is an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound and a member of monochlorobenzenes.	the molecule is a deuterated compound that is is is is is is an isotopologue of chloroform in which the four hydrogen atoms have been replaced by deuterium. it is a deuterated compound, a gamma - lactam and an aliphatic sulfide.	The molecule is a member of the class of pyrazoles that is 1H-pyrazole that is substituted at positions 1, 3, 4, and 5 by 2,6-dichloro-4-(trifluoro methyl)phenyl, cyano, (trifluoromethyl)sulfinyl, and amino groups, respectively. It is a nitrile, a dichlorobenzene, a primary amino compound, a member of pyrazoles, a sulfoxide and a member of (trifluoromethyl) benzenes	The molecule is a member of the class of pyrazoles that is 1H-pyrazole that is substituted at positions 1, 3, 4, and 5 by 2,6-dichloro-4-(trifluoro methyl)phenyl, cyano, (trifluoromethyl)sulfinyl, and amino groups, respectively. It is a nitrile, a dichlorobenzene, a primary amino compound, a member of pyrazoles, a sulfoxide and a member of (trifluoromethyl) benzenes	The molecule is a member of the class of pyrazoles that is 1H-pyrazole that is substituted at positions 1, 3, 4, and 5 by 2,6-dichloro-4-(trifluoromethyl)phenyl, cyano, (trifluoromethyl) sulfanyl, and amino groups, respectively. It is a metabolite of the agrochemical fipronil. It has a role as a marine xenobiotic metabolite. It is a member of pyrazoles, a dichlorobenzene, a member of (trifluoromethyl)benzenes, an organic sulfide and a nitrile.
5 	the molecule is a fifteen - membered oligopeptide comprising glycyl, lysyl, lysyl, leucyl, lysyl, lysyl, leucyl, lysyl, leucyl, lysyl, leucyl, lysyl, leucyl, lysyl, lysyl, leucyl, lysyl, lysyl, leucyl, lysyl […] lysyl, glutaminyl, lysyl, prolyl, lysyl, lysyl, lysyl, lysyl, lysyl, lysyl, lysyl, leucyl, lysyl, lys	the molecule is a linear seventeen - membered polypeptide comprising the sequence glu - asn - pro - val - val - his - phe - phe - asn - ile - val - thr - pro. corresponds to the sequence of the myelin basic protein 83 - 99 (mbp83 - 99) immunodominant epitope with the valyl residue at position 91 replaced by tyrosyl [mbp83 - 99 (91)].	The molecule is a linear 27-membered polypeptide comprising the sequence Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Glu-Asn-Pro-Val-Val-His-Phe-Phe-Phe-Asn-Ile-Val-Thr-Pro-Arg-Thr-Pro. Corresponds to the sequence of the myelin basic protein 83-99 (MBP83-99) immunodominant epitope with the lysyl residue at position 91 replaced by phenylalanyl [MBP83-99(F(91))] and with an (L-lysylglycyl)5 [(KG5)] linker attached to the glutamine(83) (E(83)) residue.	The molecule is a linear 27-membered polypeptide comprising the sequence Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Glu-Asn-Pro-Val-Val-His-Phe-Phe-Phe-Asn-Ile-Val-Thr-Pro-Arg-Thr-Pro. Corresponds to the sequence of the myelin basic protein 83-99 (MBP83-99) immunodominant epitope with the lysyl residue at position 91 replaced by phenylalanyl [MBP83-99(F(91))] and with an (L-lysylglycyl)5 [(KG5)] linker attached to the glutamine(83) (E(83)) residue.	The molecule is a linear 27-membered polypeptide comprising the sequence Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Glu-Asn-Pro-Val-Val-His-Phe-Phe-Tyr-Asn-Ile-Val-Thr-Pro-Arg-Thr-Pro. Corresponds to the sequence of the myelin basic protein 83-99 (MBP83-99) immunodominant epitope with the lysyl residue at position 91 replaced by tyrosyl [MBP83-99(Y(91))] and with an (L-lysylglycyl)5 [(KG5)] linker attached to the glutamine(83) (E(83)) residue.
6 	the molecule is an l - alpha - amino acid anion resulting from the removal of a proton from the carboxylic acid group of (s) - 2 - hydroxy - l - cysteinyl - l - cysteine. it is a conjugate base of a (s) - 2 - hydroxy - l - methionine.	the molecule is the stable isotope of oxygen with relative atomic mass 15. 99. the most abundant (99. 76 atom percent) isotope of naturally occurring oxygen.	The molecule is the D-enantiomer of methioninate. It has a role as an Escherichia coli metabolite, a Saccharomyces cerevisiae metabolite and a bacterial metabolite. It is a conjugate base of a D-methionine. It is an enantiomer of a L-methioninate.	The molecule is the D-enantiomer of methioninate. It has a role as an Escherichia coli metabolite, a Saccharomyces cerevisiae metabolite and a plant metabolite. It is a conjugate base of a D-methionine. It is an enantiomer of a L-methioninate.	The molecule is the D-enantiomer of methioninate. It has a role as an Escherichia coli metabolite and a Saccharomyces cerevisiae metabolite. It is a conjugate base of a D-methionine. It is an enantiomer of a L-methioninate.
7 	the molecule is a sesquiterpene lactone. it has a role as an antineoplastic agent and a plant metabolite. it is a sesquiterpene lactone, an organic heterotricyclic compound and a secondary alcohol.	the molecule is the stable isotope of oxygen with relative atomic mass 15. 999131, 100 atom percent natural abundance and nuclear spin 3 / 2.	The molecule is a maleate salt obtained by combining acetophenazine with two molar equivalents of maleic acid. It has a role as a phenothiazine antipsychotic drug. It contains an acetophenazine.	The molecule is a maleate salt obtained by combining rosuvastatin with one molar equivalent of maleic acid. It has a role as an antineoplastic agent and a B-Raf inhibitor. It contains a rosuvastatin(1+).	The molecule is a maleate salt obtained by combining afatinib with two molar equivalents of maleic acid. Used for the first-line treatment of patients with metastatic non-small cell lung cancer. It has a role as a tyrosine kinase inhibitor and an antineoplastic agent. It contains an afatinib.
8 	the molecule is a dtdp - sugar having 4 - dehydro - 6, 6 - dideoxy - alpha - d - manno - oct - 2 - ulosonic acid. it has a role as an escherichia coli metabolite and a mouse metabolite. it is a conjugate acid of a dtdp - alpha - d - glucose (2 -).	the molecule is the stable isotope of helium with relative atomic mass 3. 016029. the least abundant (0. 000137 atom percent) isotope of naturally occurring helium.	The molecule is a dTDP-sugar having 4-dehydro-2,6-dideoxy-beta-L-glucose as the sugar component. It is a dTDP-sugar and a secondary alpha-hydroxy ketone. It derives from a dTDP-L-glucose.	The molecule is a dTDP-sugar having 4-dehydro-2,6-dideoxy-alpha-D-glucose as the sugar component. It is a dTDP-sugar and a secondary alpha-hydroxy ketone. It derives from a dTDP-D-glucose.	The molecule is a dTDP-sugar having 4-dehydro-2,6-dideoxy-alpha-D-glucose as the sugar component. It has a role as a bacterial metabolite. It is a dTDP-sugar and a secondary alpha-hydroxy ketone. It derives from a dTDP-D-glucose. It is a conjugate acid of a dTDP-4-dehydro-2,6-dideoxy-alpha-D-glucose(2-).
9 	the molecule is a tetrapeptide composed of l - asparagine, l - aspartyl, l - aspartic acid, and l - aspartic acid units joined in sequence by peptide linkages. it has a role as a metabolite. it derives from a l - glutamic acid.	the molecule is the stable isotope of oxygen with relative atomic mass 15. 99. the most abundant (99. 99 atom percent) isotope of naturally occurring oxygen.	The molecule is a tripeptide composed of two L-leucine units joined to L-aspartic acid by a peptide linkage. It has a role as a metabolite. It derives from a L-leucine and a L-aspartic acid.	The molecule is a tripeptide composed of L-leucine, L-valine and L-aspartic acid joined in sequence by peptide linkages. It has a role as a metabolite. It derives from a L-leucine, a L-valine and a L-aspartic acid.	The molecule is a tripeptide composed of L-leucine, L-valine and L-aspartic acid joined in sequence by peptide linkages. It has a role as a metabolite. It derives from a L-leucine, a L-valine and a L-aspartic acid.

Figure 9: More examples of interesting captions generated by different models.

G Testing Model Diversity with Retrieval

To test the diversity of generations, we apply a Text2Mol (Edwards et al., 2021) cross-modal retrieval model to the entire generated set of molecules or descriptions. In the case of molecules, we first take the molecules generated for our test set. We consider these molecules as our corpus and then use the descriptions (which were used to generate the molecules in the first place) as our queries. So, for each query we look at the rank of its generated molecule (the highest rank is 1). This process tests whether the Text2Mol retrieval model can differentiate between the generated (valid) molecules. Doing so means it can retrieve a specific molecule when given the description used to generate it. If the generative model did not sufficiently take the descriptions into consideration, then the retrieval model won't be able to distinguish between generated molecules and the scores will be very low (such as the transformer model, which frequently generates the same molecule/caption).

As an example, consider that we have 10 descriptions of molecules. For each description, we use a generative model to generate a molecule. Now, we treat these 10 generated molecules as our corpus. Using our retrieval model, we now consider each description as a query and try to retrieve the molecule that was generated from that description. If the retrieval model performs poorly, that means the molecules which were generated are difficult to distinguish from one another. By using this method with different generative models, we measure the relative diversity of generated molecules along with how well the generated molecules match the description.

Results are reported in Tables 8 and 9 for retrieving generated molecules from descriptions and for retrieving generated descriptions from molecules, respectively. We use the same Text2Mol model for retrieval here as in the Text2Mol metric. For description of metrics, see (Edwards et al., 2021). Results indicate that MolT5 model generations are sufficiently distinct to be retrievable. In contrast, the outputs of the captioning transformer are essentially indistinguishable for the retrieval model.

Model	Mean Rank	MRR	Hits@1	Hits@10	Hits@100	Validity
Ground Truth	4.9	0.735	60.4%	95.2%	99.5%	100%
RNN	106.7	0.192	10.45%	37.0%	74.4%	54.2%
Transformer	426.4	0.106	5.62%	19.8%	46.2%	90.6%
T5-Small	113.9	0.441	33.0%	64.1%	81.1%	60.7%
MolT5-Small	126.2	0.413	30.1%	62.2%	80.2%	72.1%
T5-Base	97.8	0.467	35.5%	67.1%	84.7%	66.0%
MolT5-Base	113.9	0.438	32.3%	65.2%	83.7%	77.2%
T5-Large	84.5	0.586	46.5%	81.2%	90.5%	90.2%
MolT5-Large	87.4	0.570	44.6%	80.1%	91.0%	90.5%

Table 8: Retrieval of generated molecules on the test split of CheBI-20.

Model	Mean Rank	MRR	Hits@1	Hits@10	Hits@100
Ground Truth	5.6	0.703	56.4%	94.3%	99.3%
RNN	137.0	0.160	7.45%	34.0%	73.6%
Transformer	1750	0.007	00.4%	01.2%	03.2%
T5-Small	60.3	0.414	28.4%	65.5%	88.1%
MolT5-Small	47.7	0.460	32.6%	70.7%	91.4%
T5-Base	69.2	0.414	28.9%	65.0%	87.0%
MolT5-Base	40.2	0.465	32.5%	72.5%	92.0%
T5-Large	29.4	0.499	35.5%	77.6%	94.2%
MolT5-Large	16.1	0.558	40.4%	84.2%	96.8%

Table 9: Retrieval of generated captions on the test split of CheBI-20.

H Statistical Significance

To strengthen the quantitative results, we conducted statistical tests between T5-Large and MolT5-Large. For molecule captioning, we carried out paired t-tests. The computed p-values and test statistics are:

- For ROUGE-1, the p-value is 1.53e-22. The test statistic is -9.841.
- For ROUGE-2, the p-value is 3.27e-26. The test statistic is -10.683.
- For ROUGE-L, the p-value is 3.58e-21. The test statistic is -9.509.
- For METEOR, the p-value is 2.02e-21. The test statistic is -9.57.
- For Text2Mol, the p-value is 1.053e-29. The test statistic is -11.431.

Note that for every metric above, the higher the score, the better the performance. Since all the test statistics are negative and the p-values are extremely small, MolT5-Large produces significant improvements over T5-Large on the task of molecule captioning.

For molecule generation, we conducted independent t-tests to compare between T5-Large and MolT5-Large:

- For MACCS FTS, the p-value is 0.008. The test statistic is -2.652.
- For RDK FTS, the p-value is 0.0092. The test statistic is -2.604.
- For Morgan FTS, the p-value is 0.0153. The test statistic is -2.426.
- For Levenshtein, the p-value is 0.064. The test statistic is 1.8544704091978725.
- For Text2Mol, the p-value is 0.168. The test statistic is -1.376724743237994.

Note that for Levenshtein, the lower the score, the better the performance. We see that the test statistics for all metrics except Levenshtein is negative. In addition, while the p-values now are typically larger than the ones computed for molecule captioning, the p-values for molecule generation are still reasonably small. Therefore, we can still conclude that MolT5-Large also produces significant improvements over T5-Large on the task of molecule generation.

I NLP Capabilities of MolT5

We finetune our MolT5-based models on some GLUE tasks and see similar results for MolT5 and T5. For example, our finetuned MolT5-base model achieved an accuracy score of 95.6% on SST-2. For comparison, T5-base achieved a score of 95.2%. Since our self-supervised learning framework uses a large amount of natural language text in addition to SMILES string, it is reasonable that our MolT5-based models still possess “typical” NLP capabilities.

J Model Probing Tests

To generate a variety of output molecules given a single input, we employ diverse beam search (Vijayakumar et al., 2016) with a beam width and beam group of 30 and a diversity penalty of 0.5. The goal of these tests (shown in the following figures) is to explore molecule outputs given very specific desired properties. Note that these brief input descriptions are out-of-distribution from the finetuning data. In the following figures, the top 10 valid molecules are shown for each prompt (order: left to right, top to bottom).

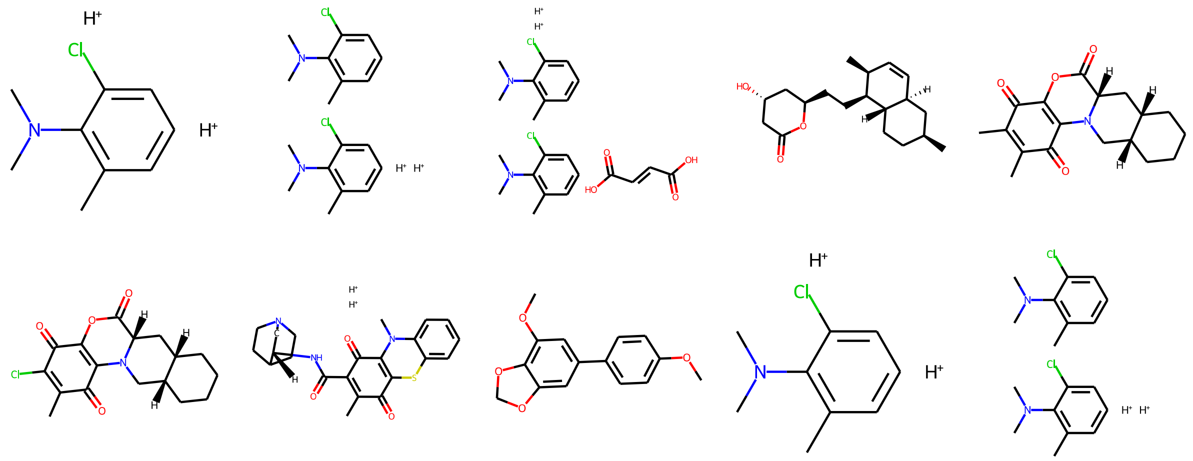


Figure 10: Input: The molecule displays antimalarial properties.

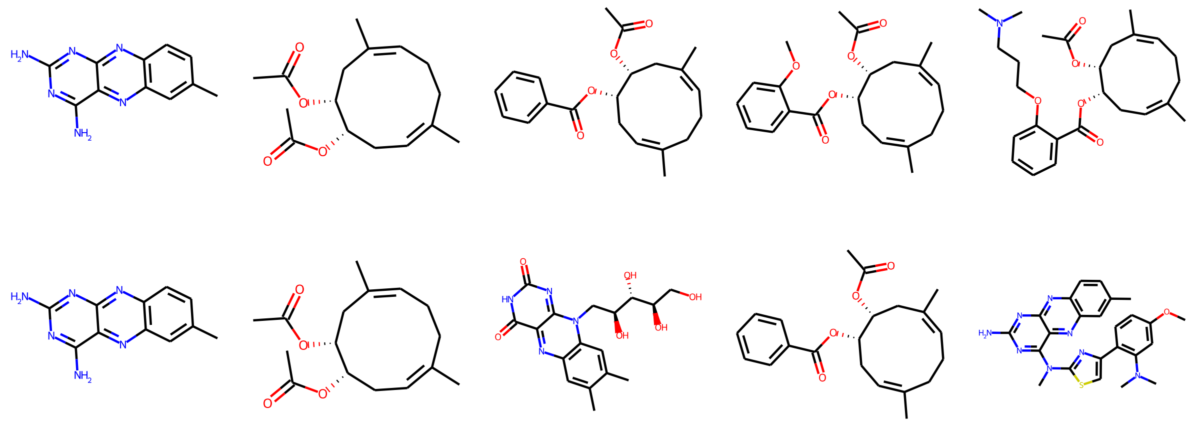


Figure 11: Input: The molecule is a apoptosis inducer.

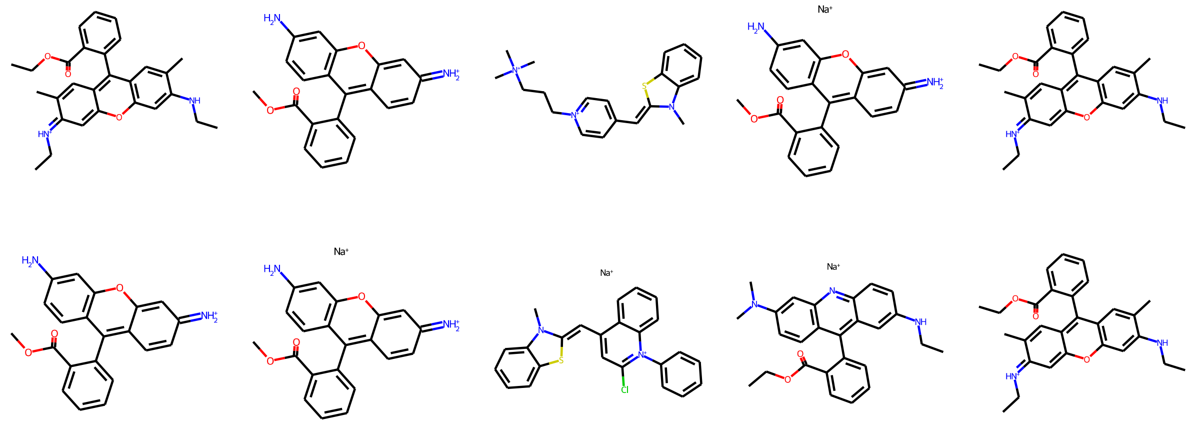


Figure 12: Input: The molecule is a blue dye.

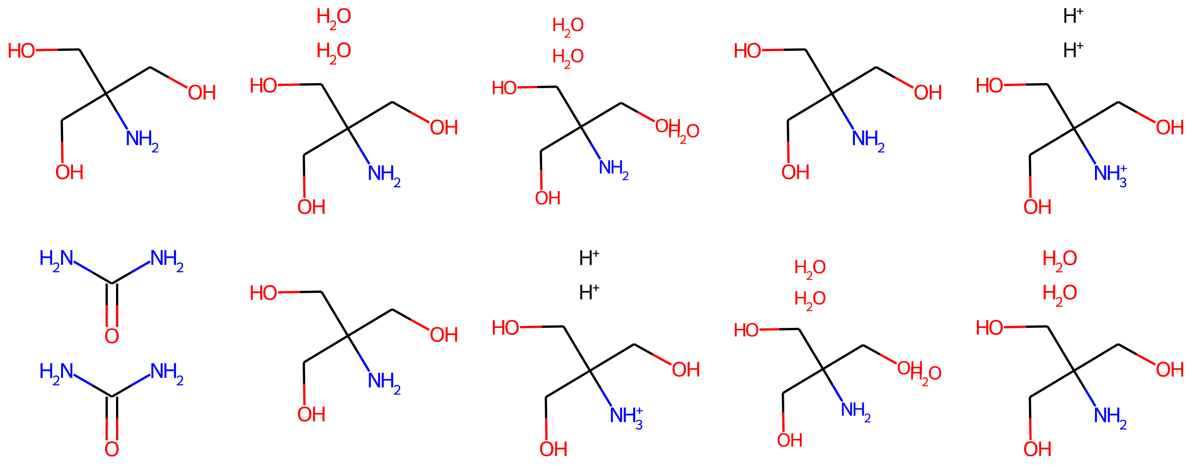


Figure 13: Input: The molecule is a coagulent.

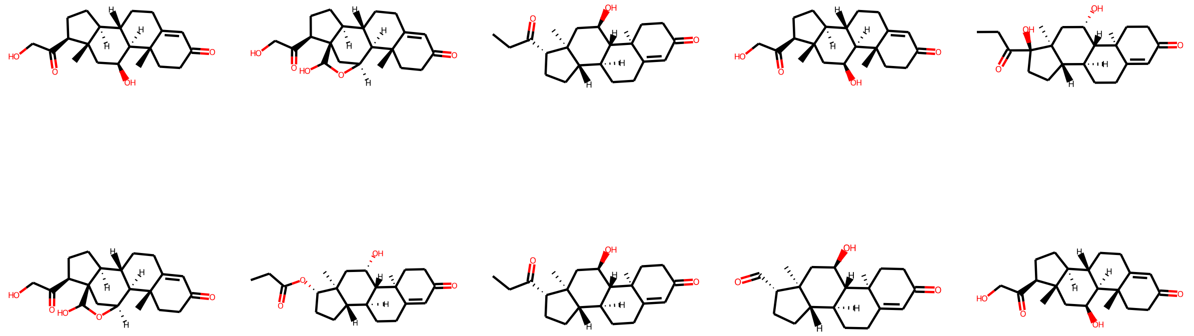


Figure 14: Input: The molecule is a corticosteroid.

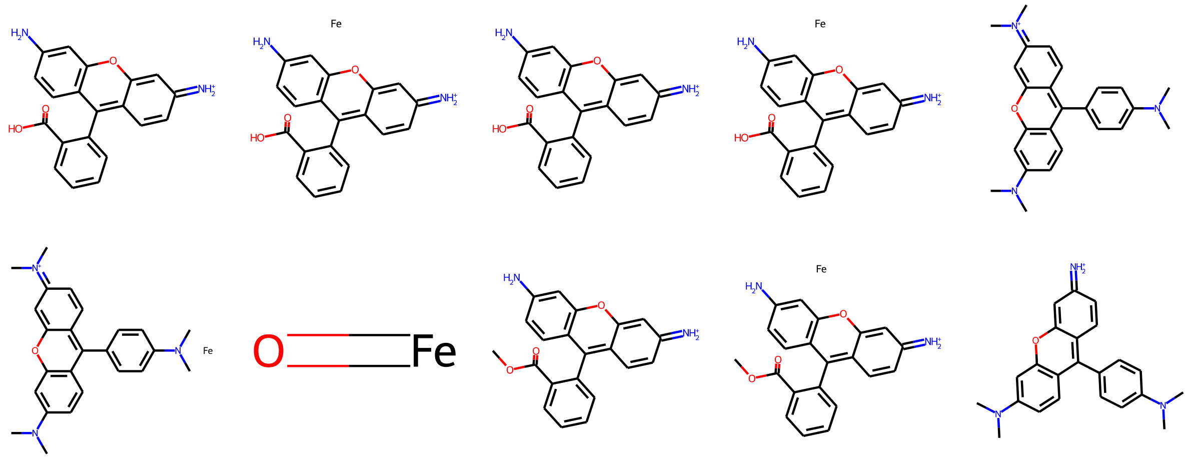


Figure 15: Input: The molecule is a fluorochrome.

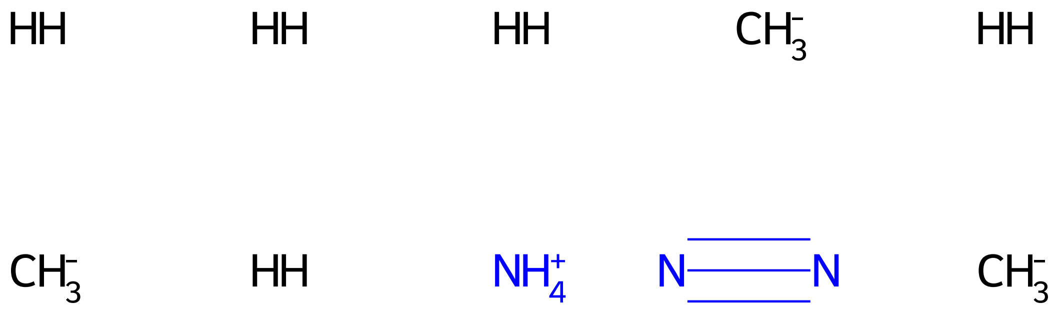


Figure 16: Input: The molecule is a gas at room temperature.

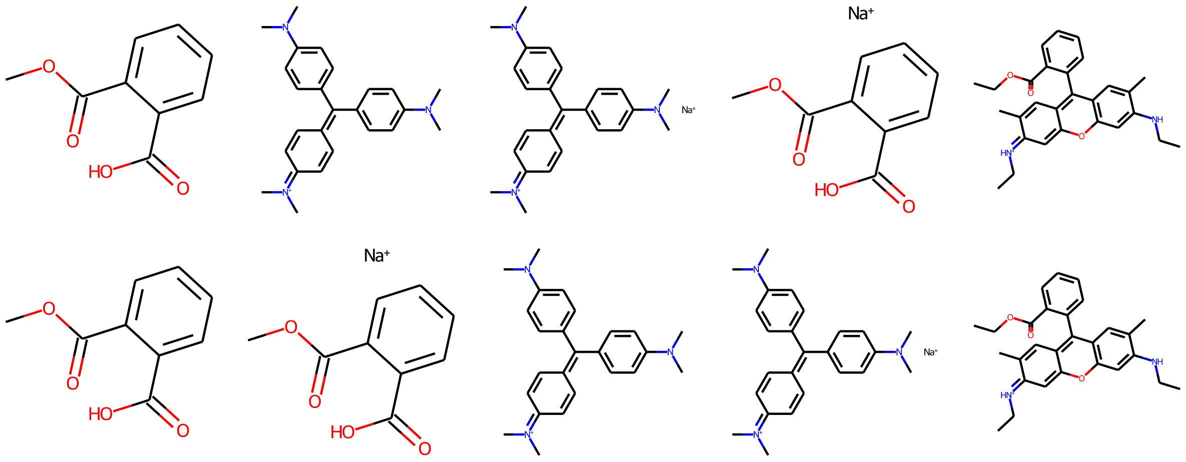


Figure 17: Input: The molecule is a green dye.

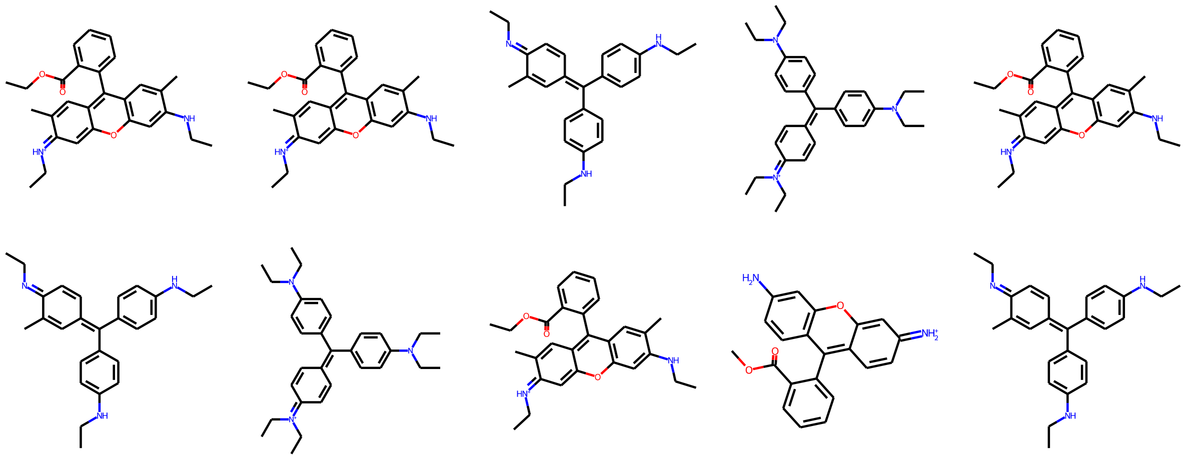


Figure 18: Input: The molecule is a histological dye.

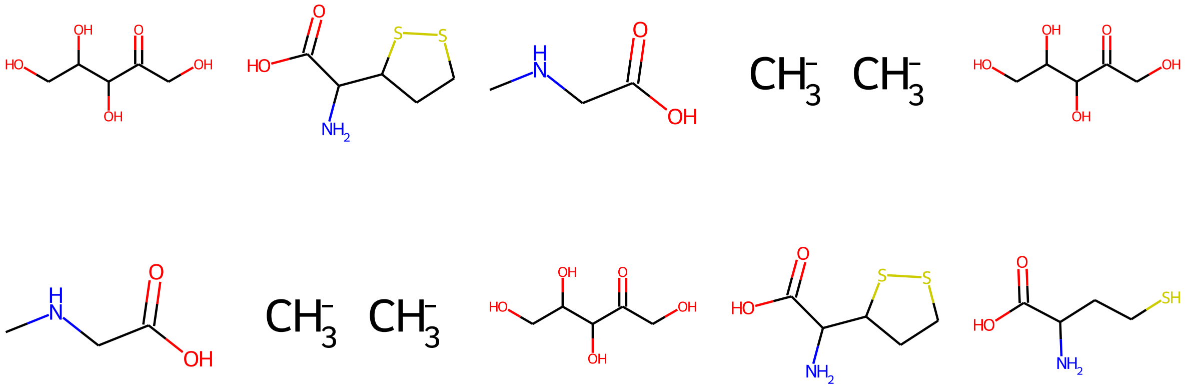


Figure 19: Input: The molecule is a human metabolite.

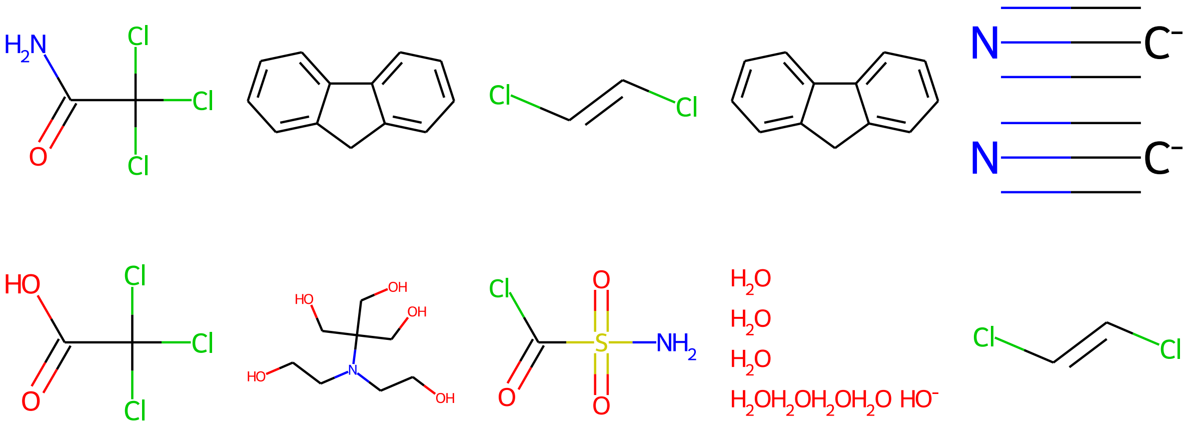


Figure 20: Input: The molecule is a hydrocarbon which tastes really cool.

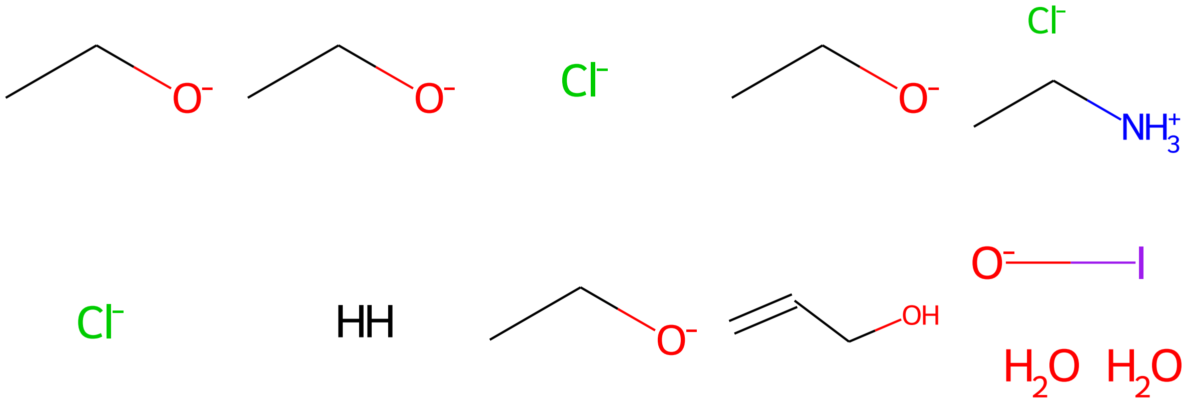


Figure 21: Input: The molecule is a liquid at room temperature.

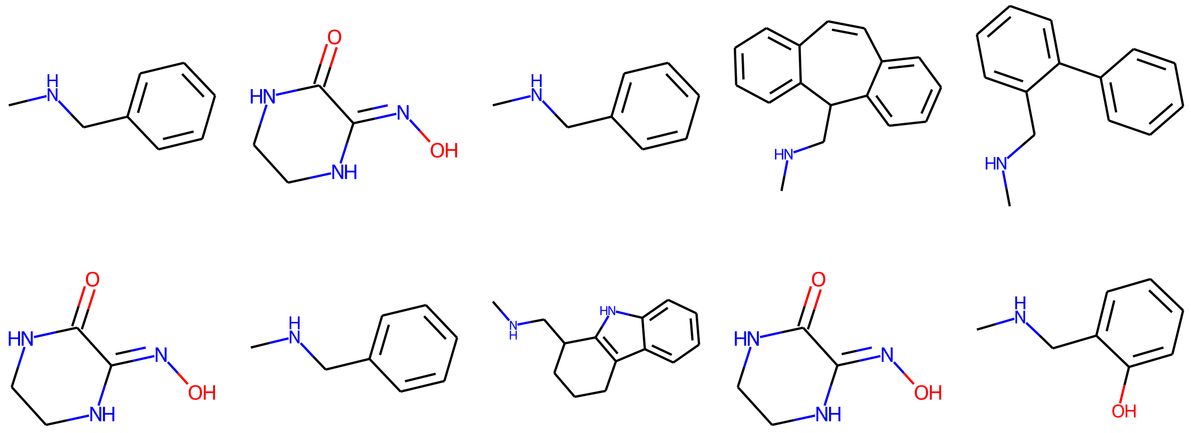


Figure 22: Input: The molecule is a macrocycle.

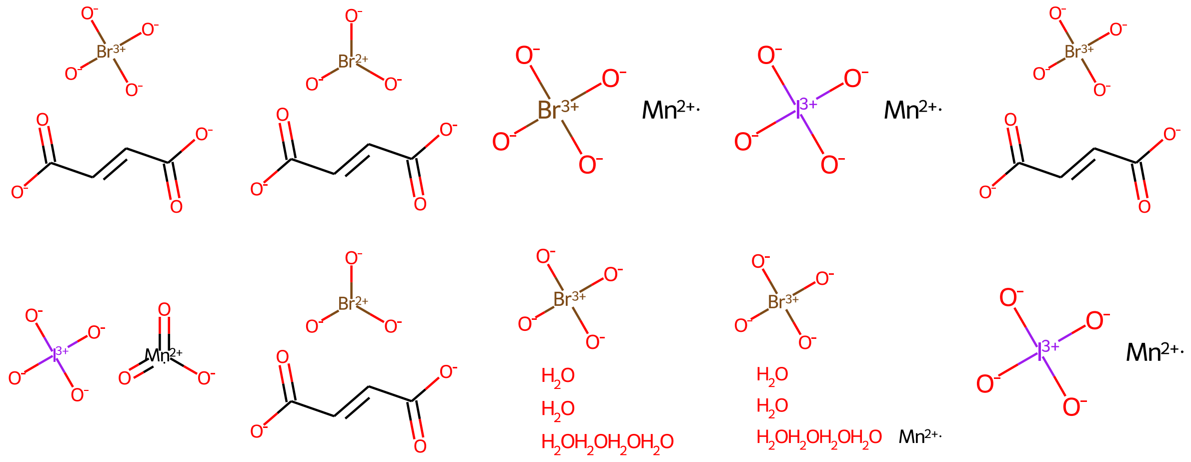


Figure 23: Input: The molecule is a maleate salt.

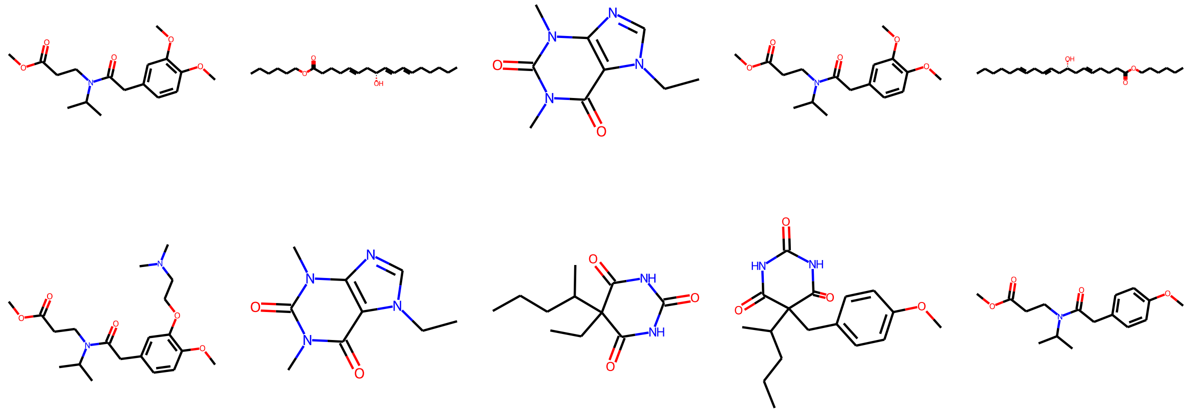


Figure 24: Input: The molecule is a neurotransmitter agent.

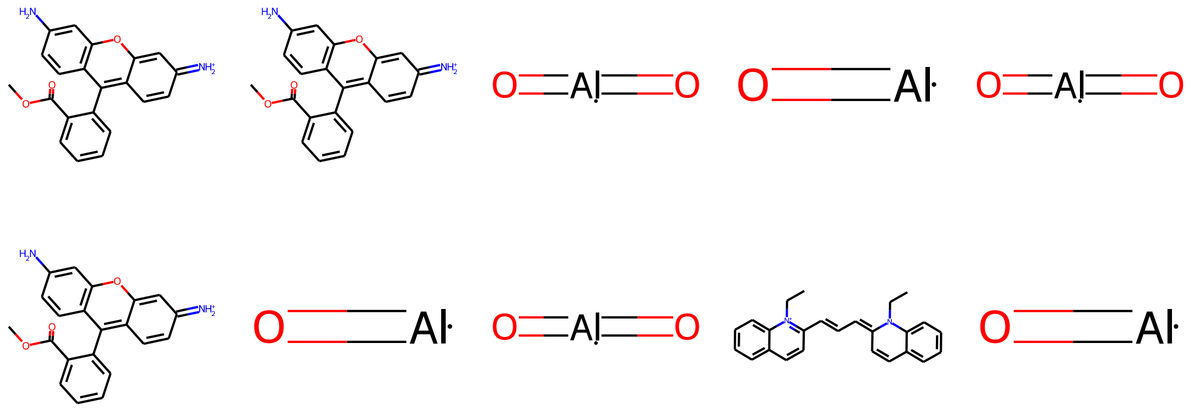


Figure 25: Input: The molecule is a orange dye.

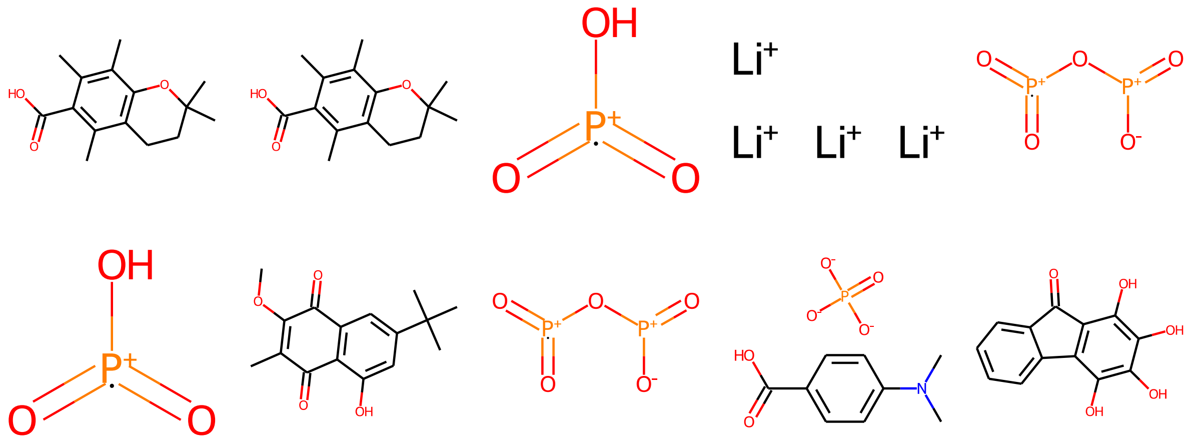


Figure 26: Input: The molecule is a photovoltaic.

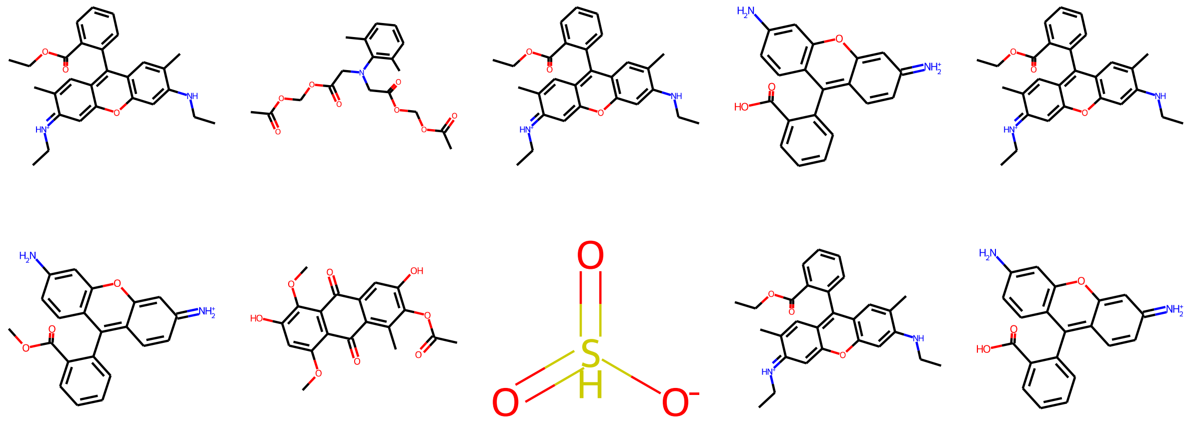


Figure 27: Input: The molecule is a pigment which converts sunlight into energy.

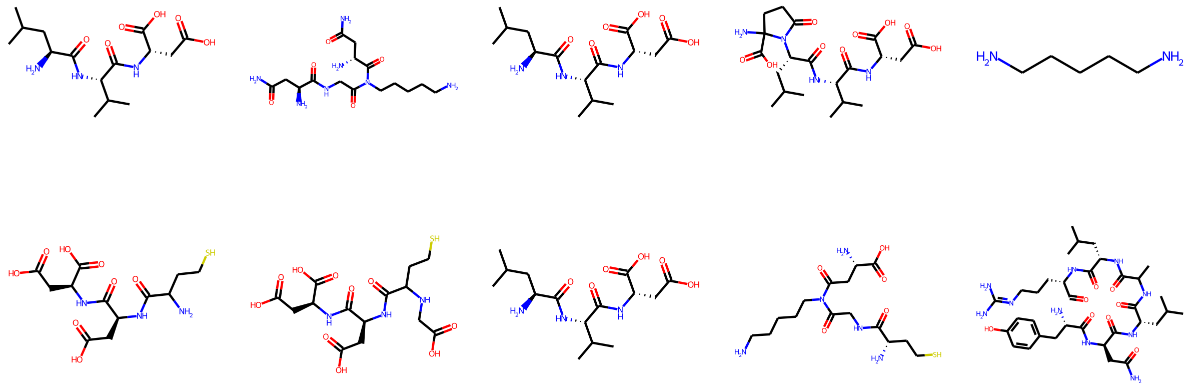


Figure 28: Input: The molecule is a polypeptide.

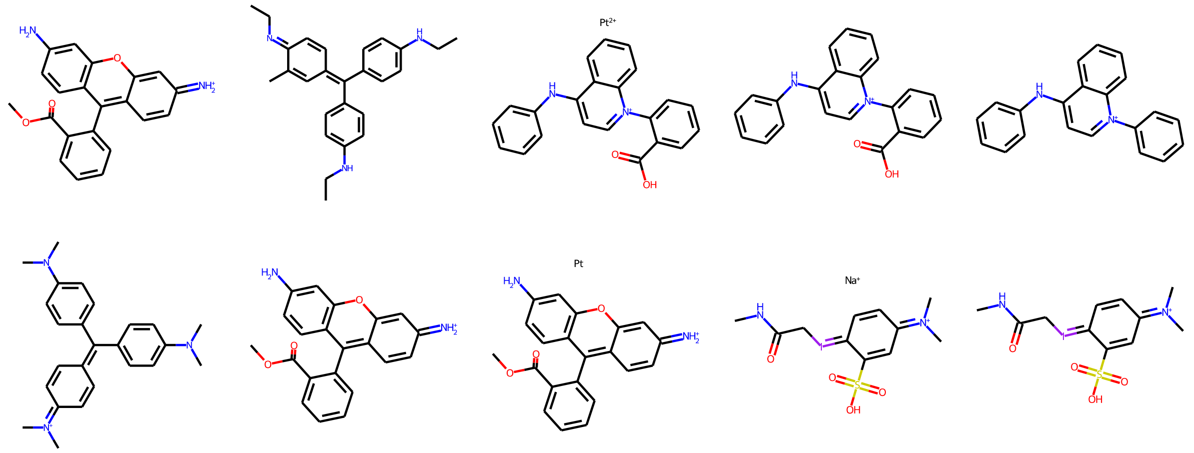


Figure 29: Input: The molecule is a purple dye.

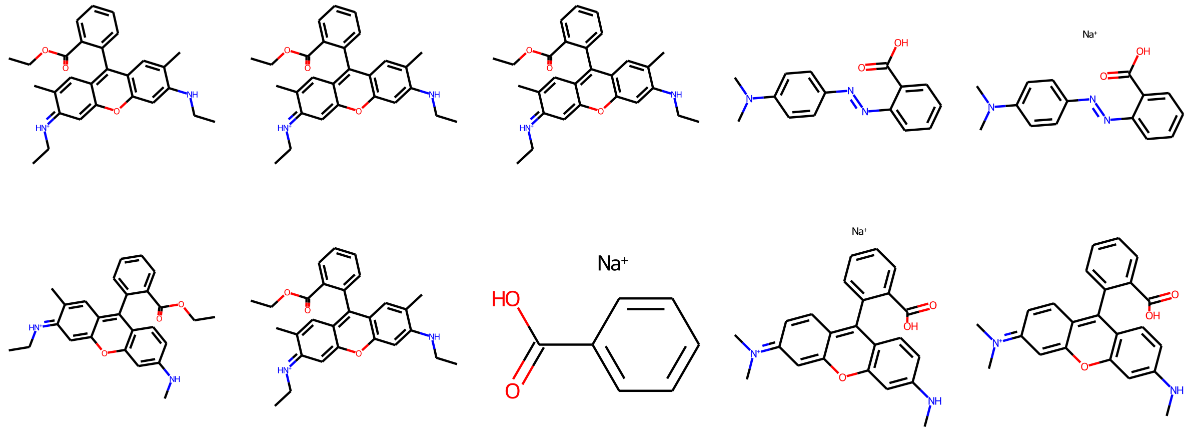


Figure 30: Input: The molecule is a red dye.

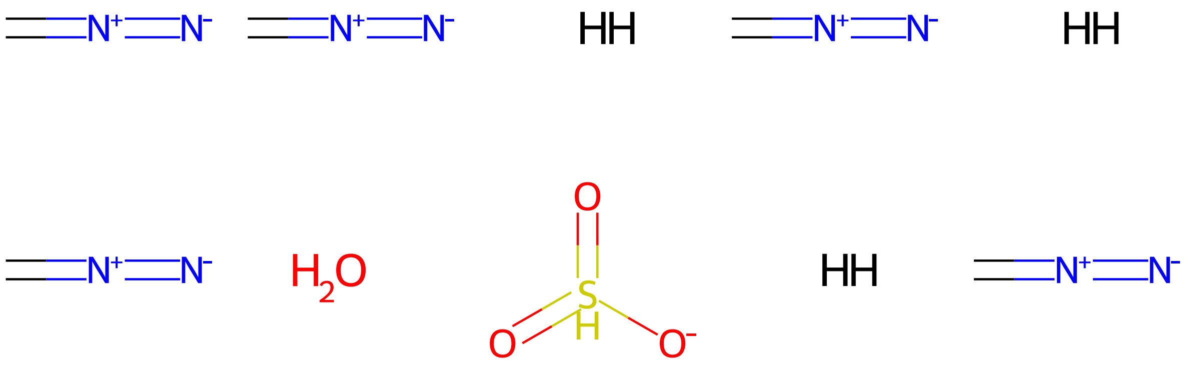


Figure 31: Input: The molecule is a solid at room temperature.

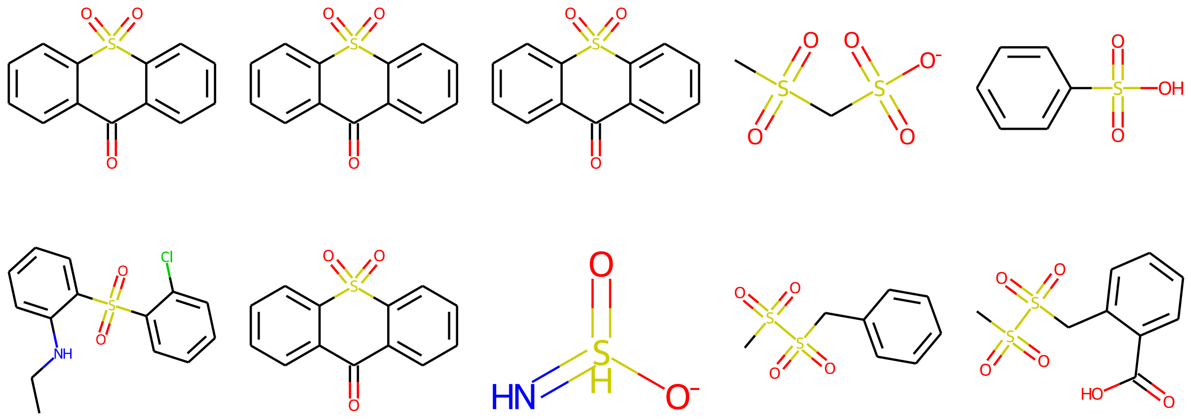


Figure 32: Input: The molecule is a sulfonated xanthene.

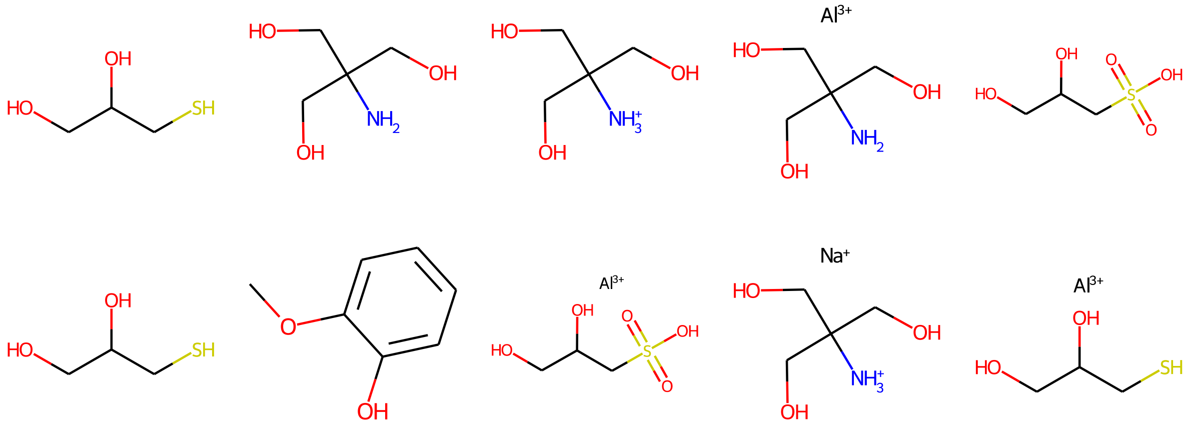


Figure 33: Input: The molecule is a sweet tasting sugar additive.

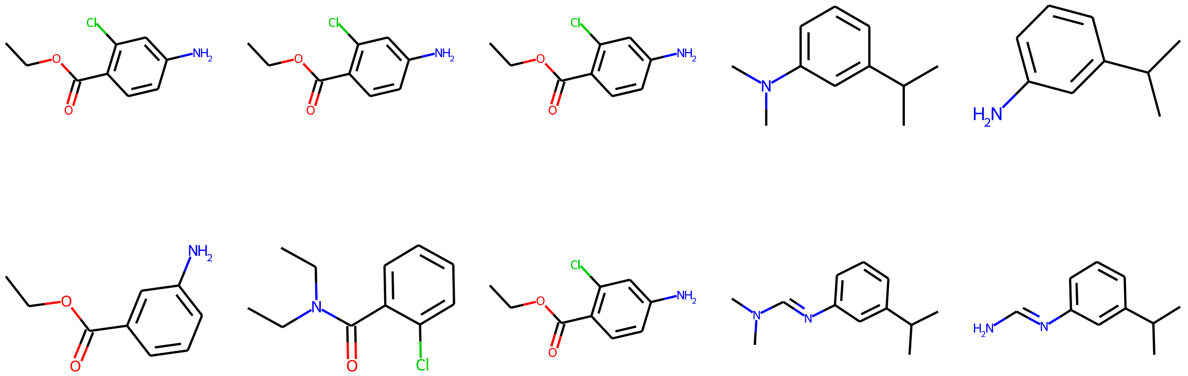


Figure 34: Input: The molecule is a topical anaesthetic.

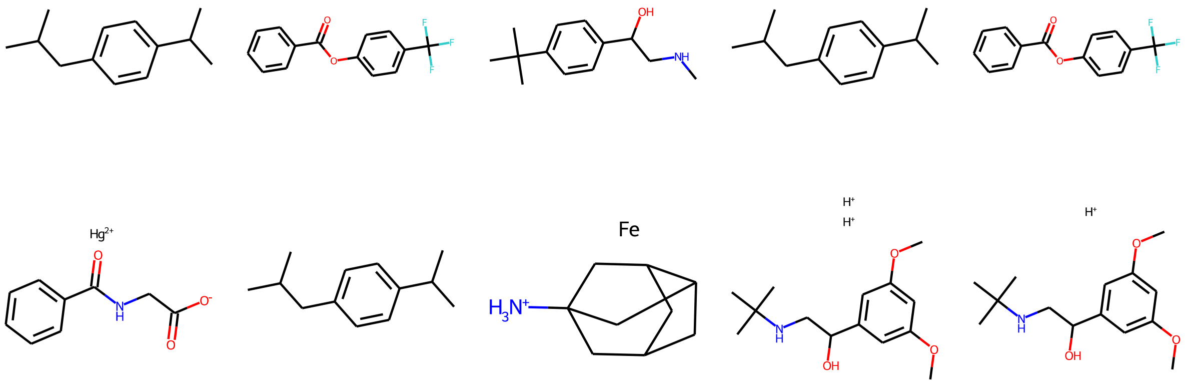


Figure 35: Input: The molecule is able to lower blood pressure.

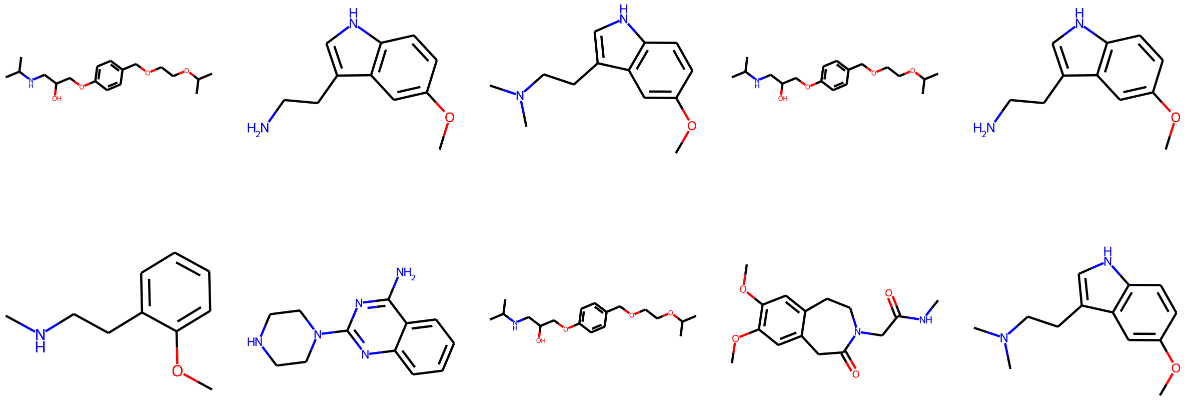


Figure 36: Input: The molecule is an adrenergic uptake inhibitor.

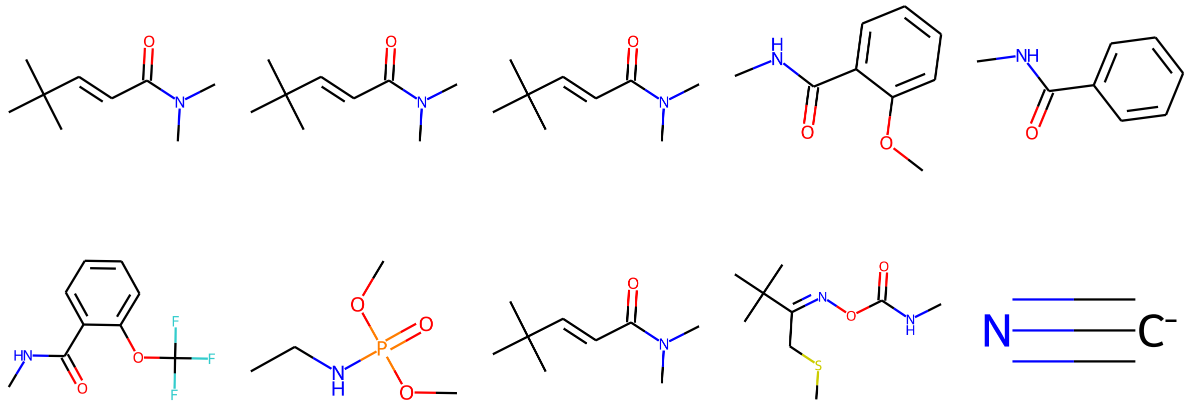


Figure 37: Input: The molecule is an agrochemical.

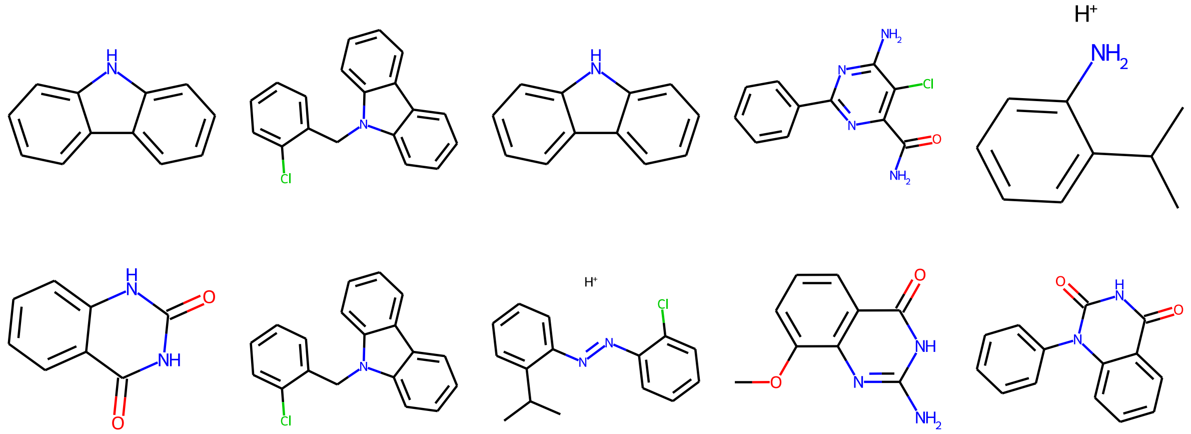


Figure 38: Input: The molecule is an anabolic agent.

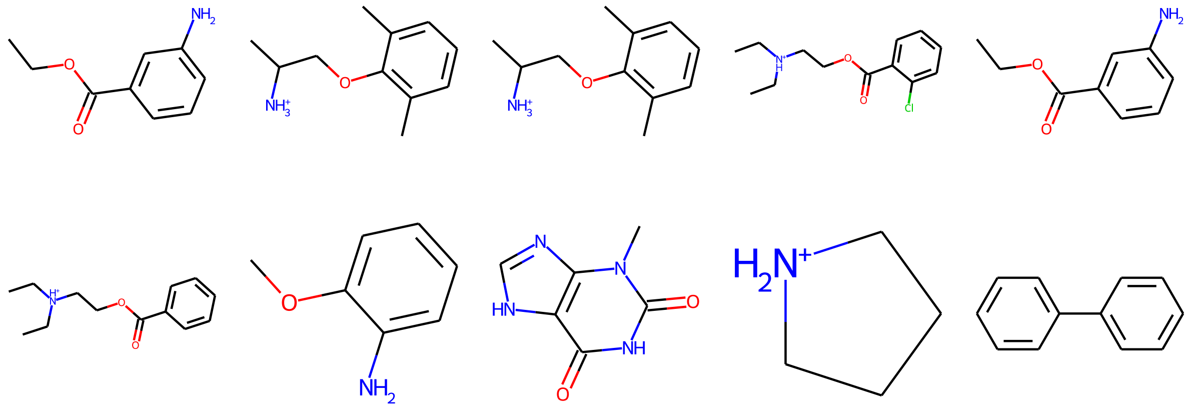


Figure 39: Input: The molecule is an analgesic.

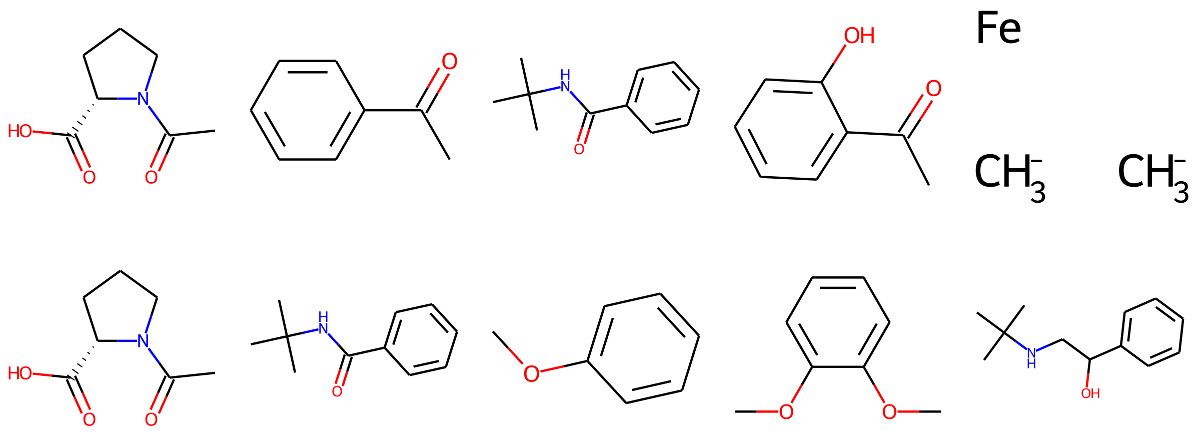


Figure 40: Input: The molecule is an angry man.

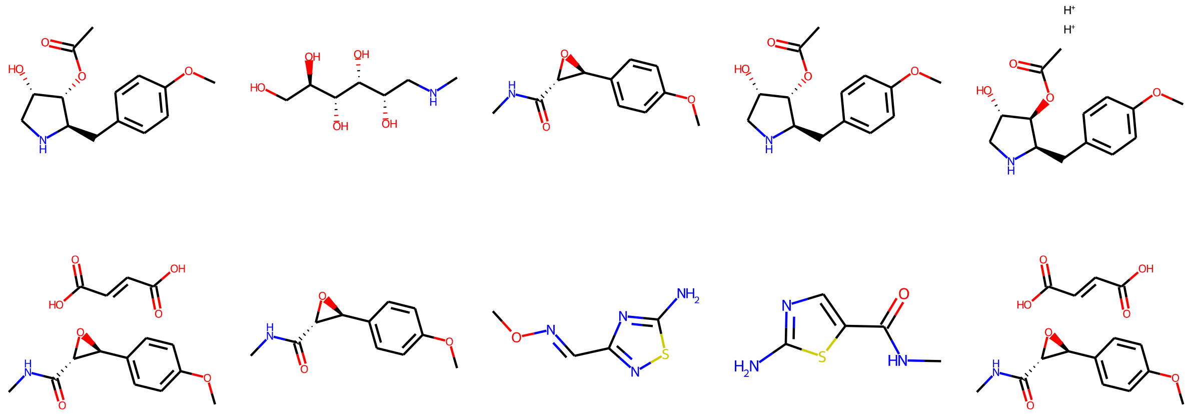


Figure 41: Input: The molecule is an antibiotic.

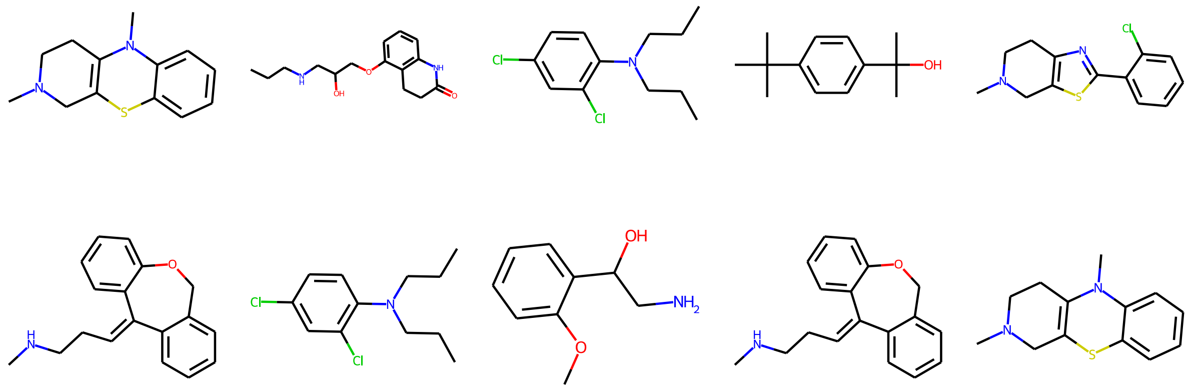


Figure 42: Input: The molecule is an antidepressant.

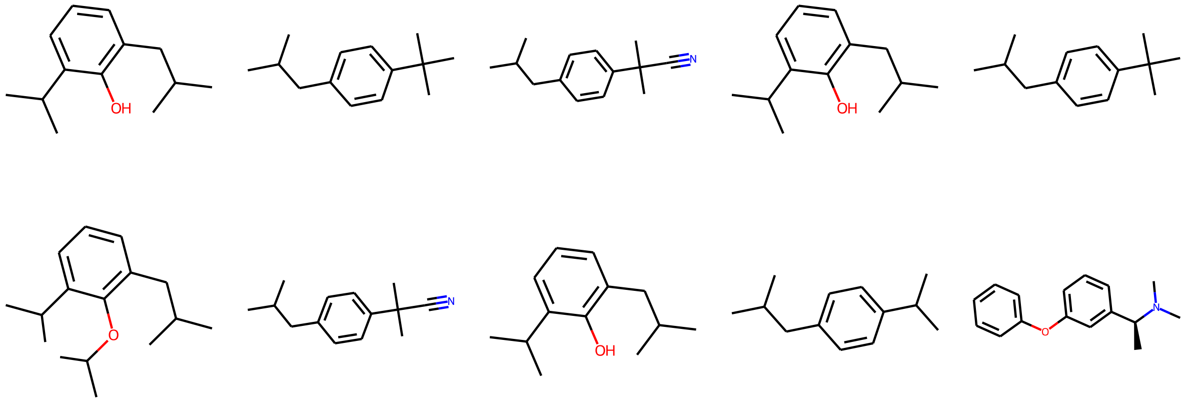


Figure 43: Input: The molecule is an anti-inflammatory agent.

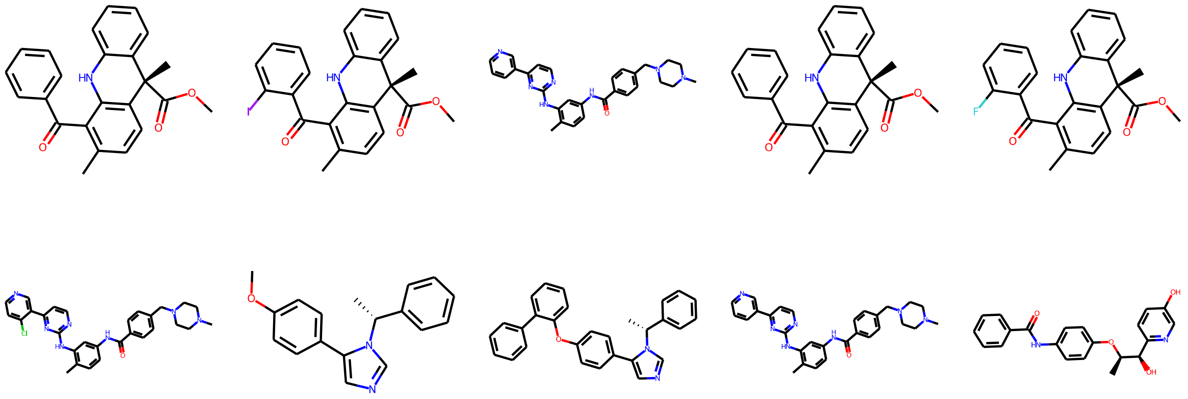


Figure 44: Input: The molecule is an antineoplastic agent.

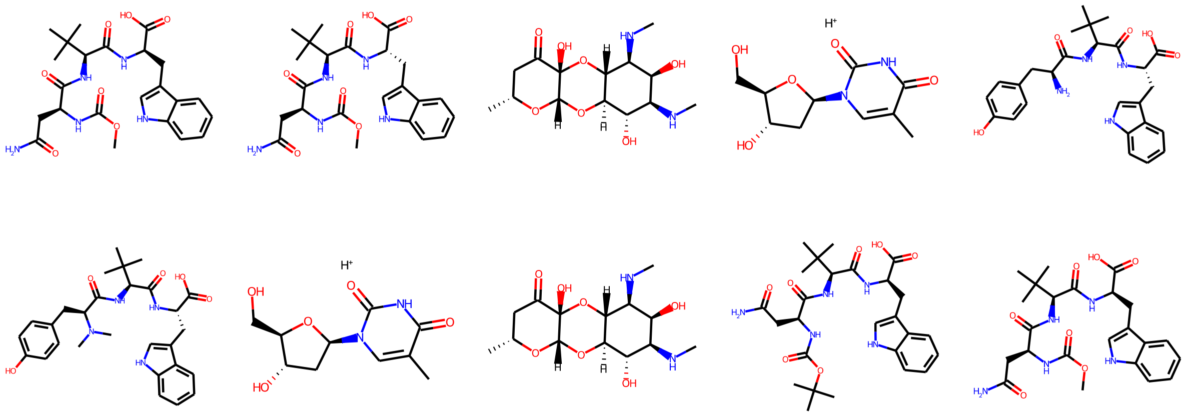


Figure 45: Input: The molecule is an antiplasmodial drug.

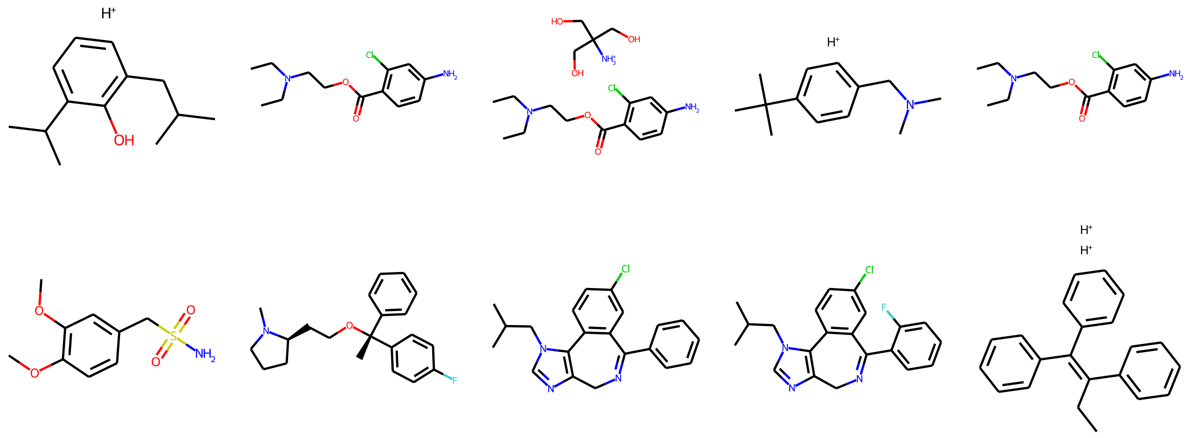


Figure 46: Input: The molecule is an antipruritic drug.

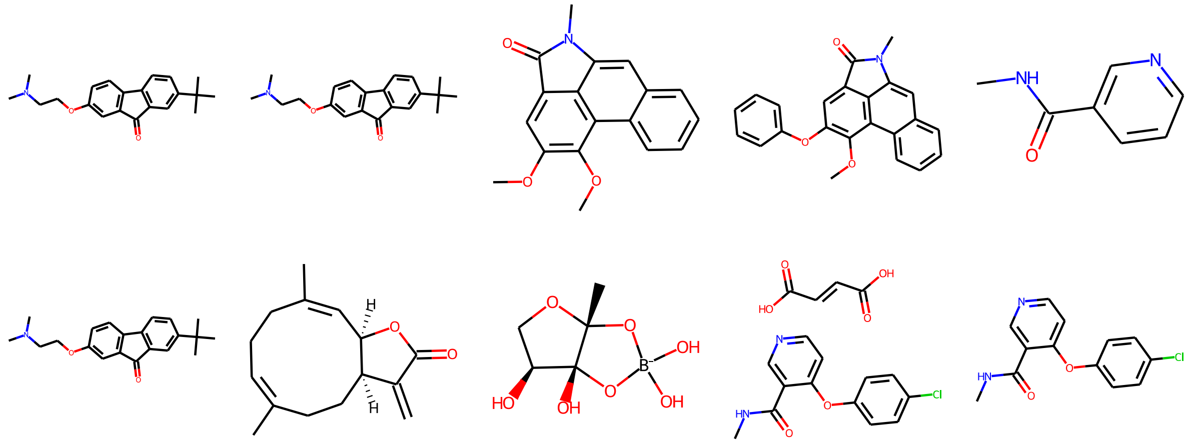


Figure 47: Input: The molecule is an antitubercular agent.

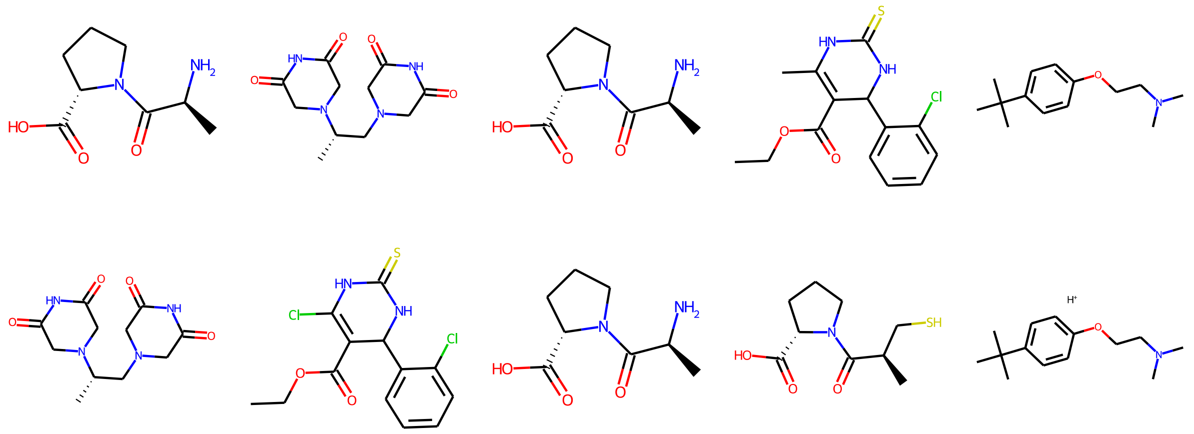


Figure 48: Input: The molecule is an anti-ulcer drug.

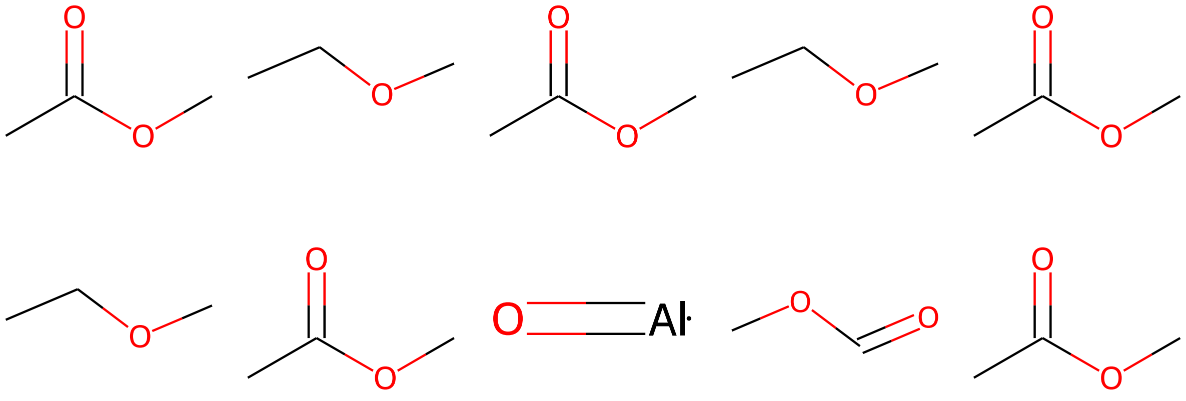


Figure 49: Input: The molecule is an aromatic ether.

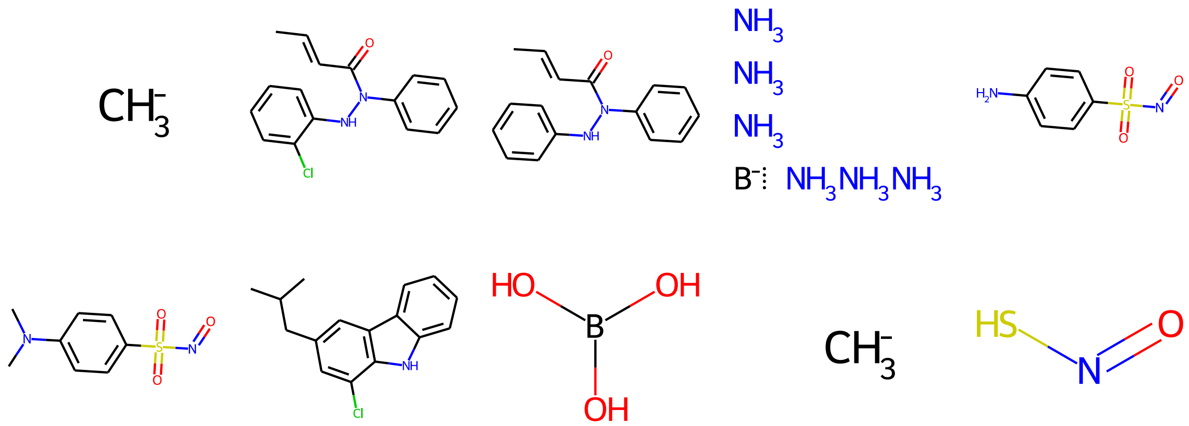


Figure 50: Input: The molecule is a catabolic agent.

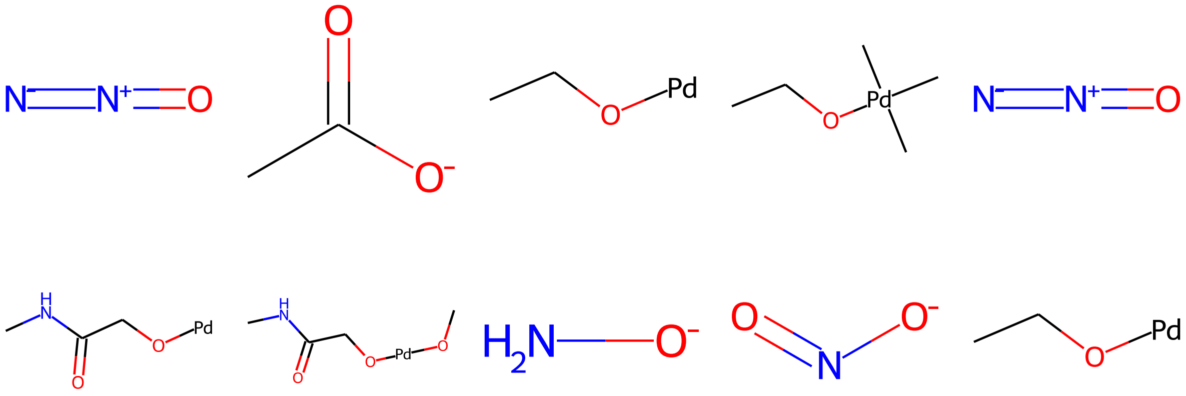


Figure 51: Input: The molecule is an explosive.

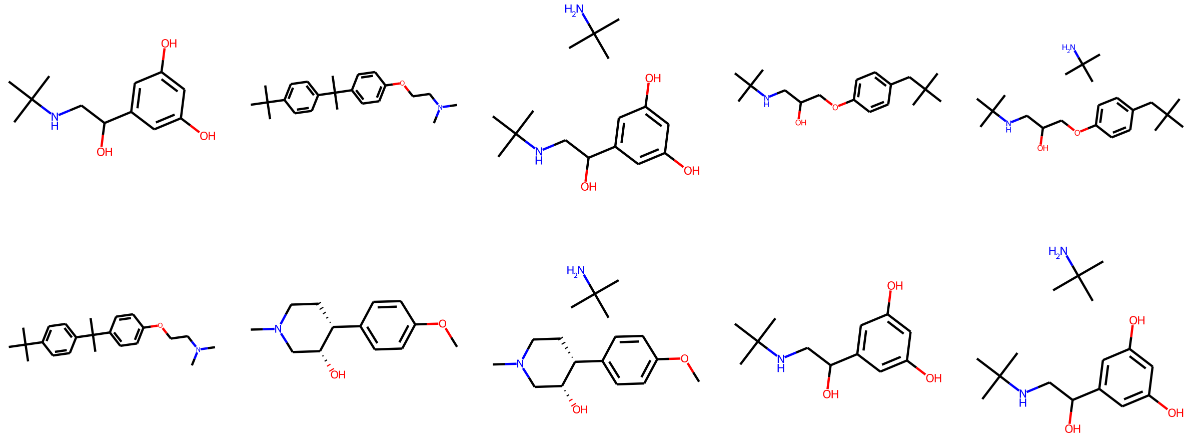


Figure 52: Input: The molecule is an inhibitor of the Parkinson’s disease.

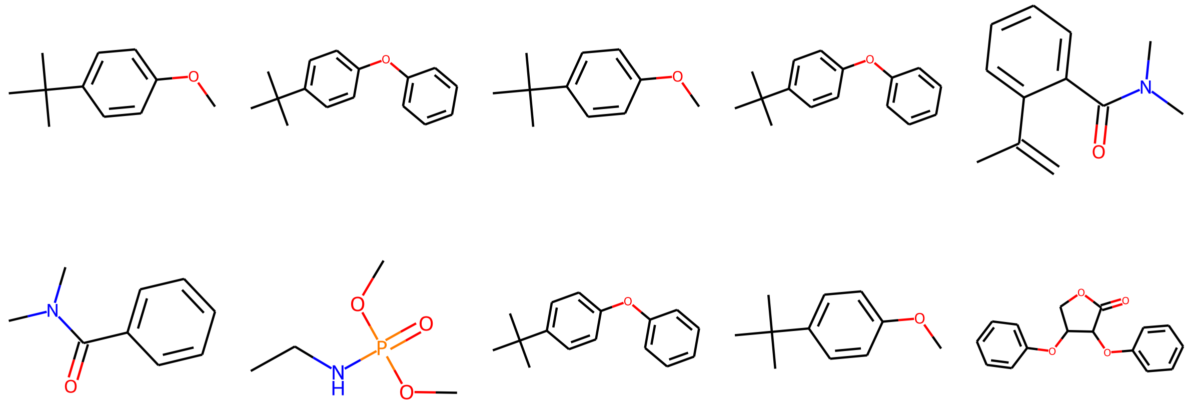


Figure 53: Input: The molecule is an insect attractant.

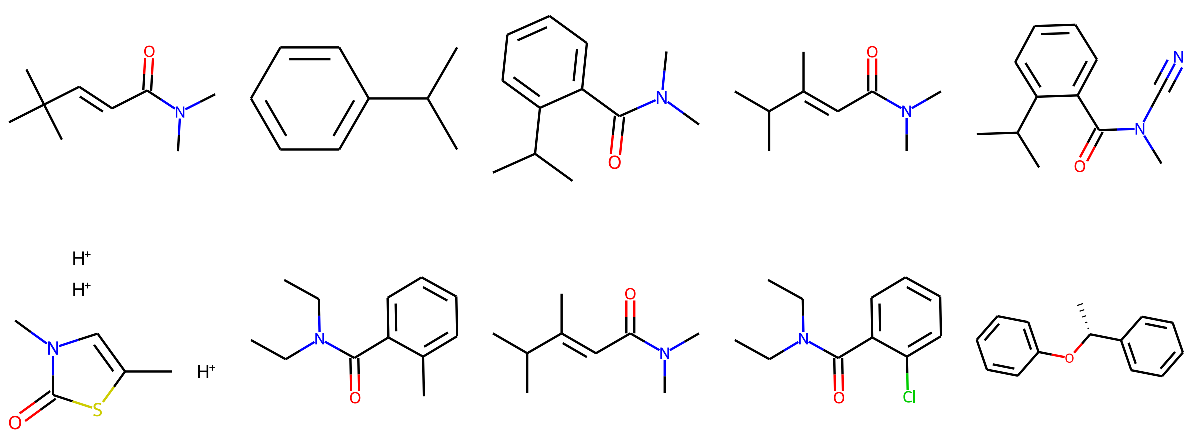


Figure 54: Input: The molecule is an insecticide.

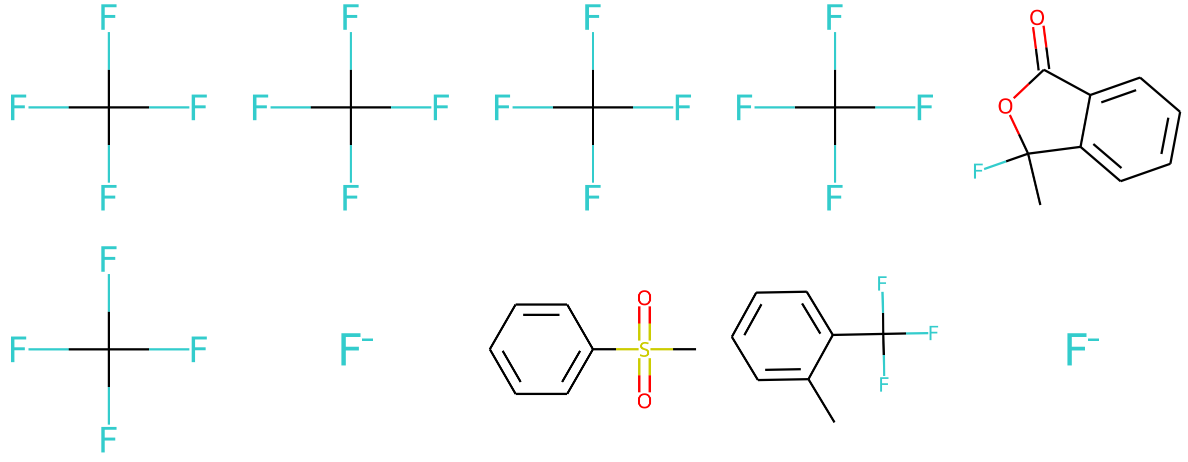


Figure 55: Input: The molecule is an organofluorine compound.

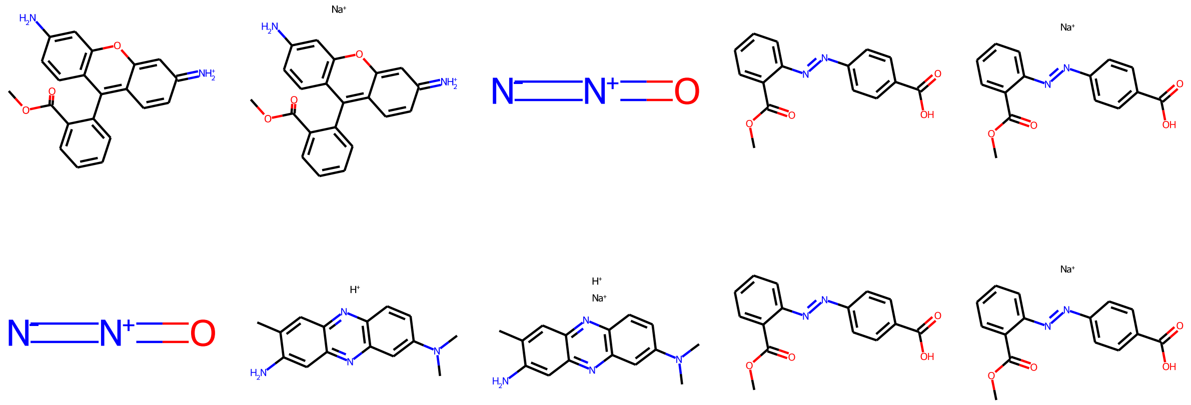


Figure 56: Input: The molecule is blue.

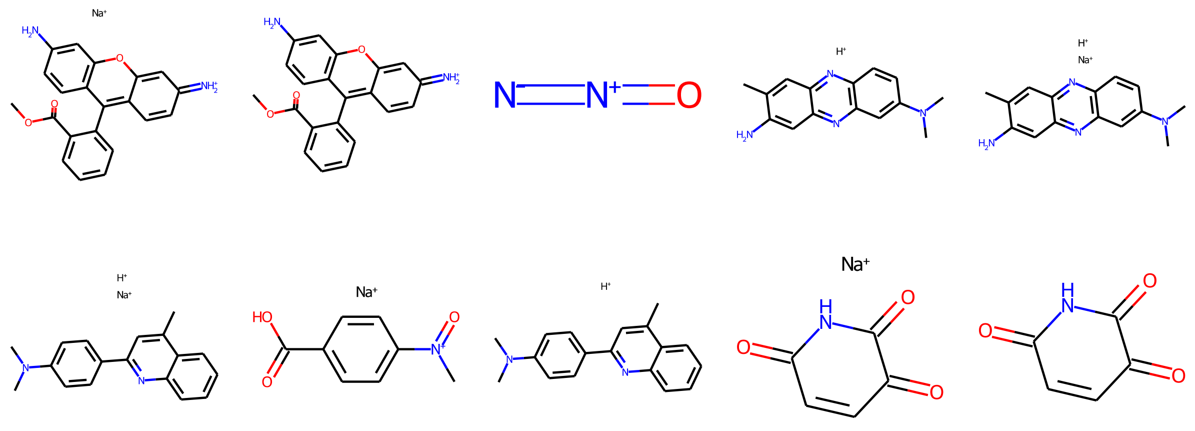


Figure 57: Input: The molecule is blue blue.

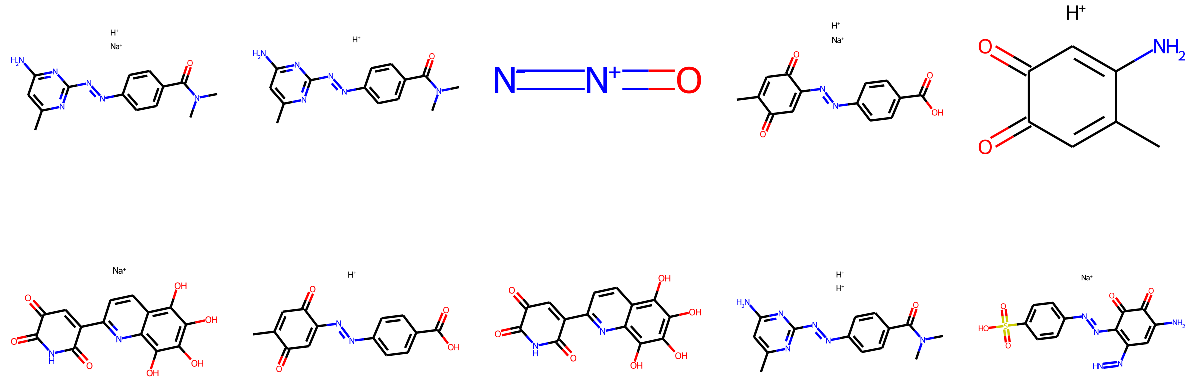


Figure 58: Input: The molecule is blue blue blue.

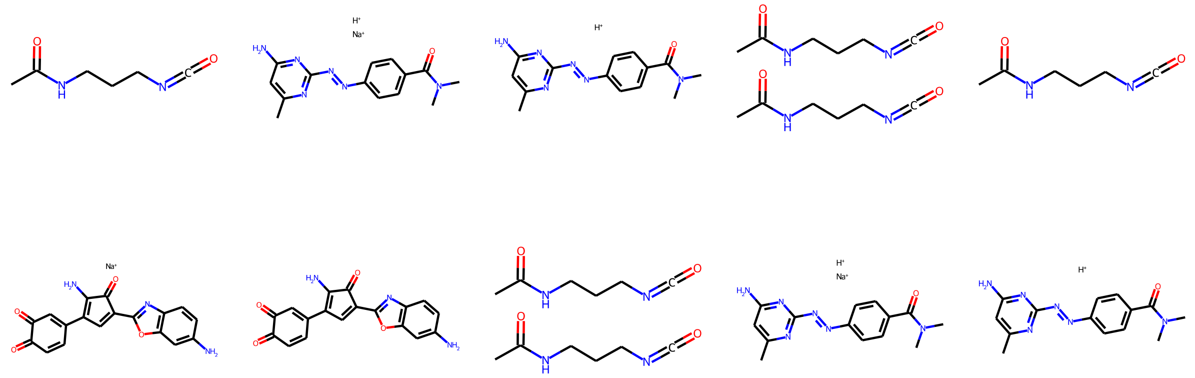


Figure 59: Input: The molecule is blue blue blue blue.

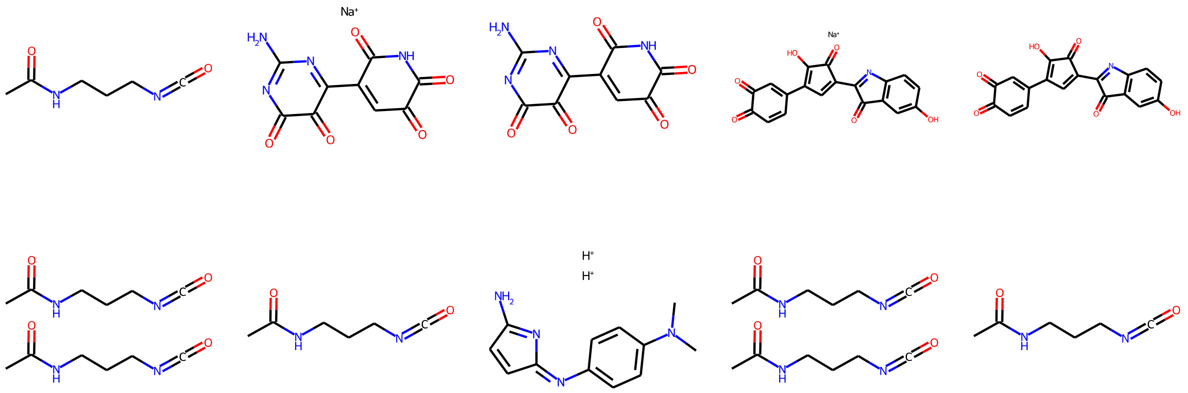


Figure 60: Input: The molecule is blue blue blue blue blue.

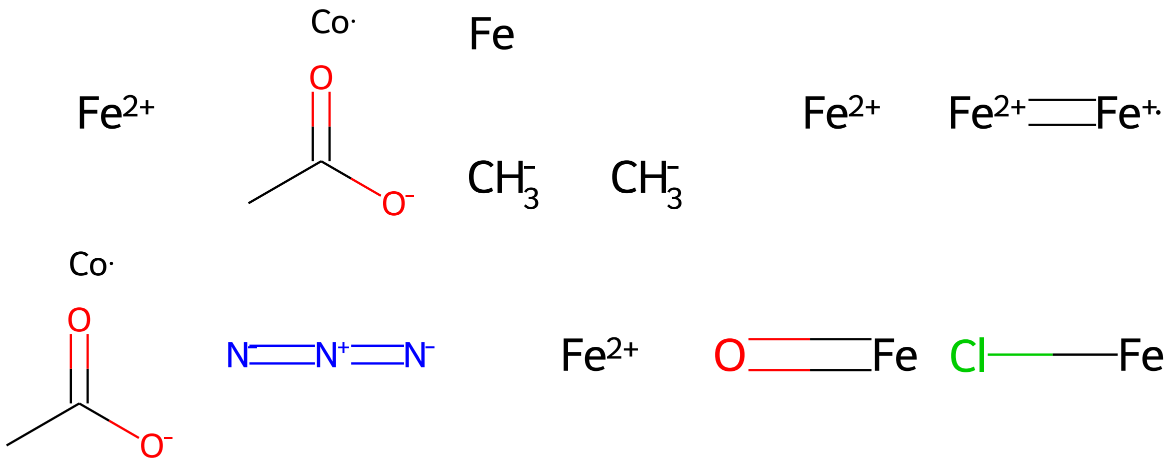


Figure 61: Input: The molecule is electrically conductive.