L+M-24: Building a Dataset for Language+Molecules @ ACL 2024

Carl Edwards[1], **Qingyun Wang**[1], **Lawrence Zhao**[2] and **Heng Ji**[1]
[1]University of Illinois Urbana-Champaign [2]Yale University
{cne2, qingyun4, hengji}@illinois.edu, larry.zhao@yale.edu

Abstract

Language-molecule models have emerged as an exciting direction for molecular discovery and understanding. However, training these models is challenging due to the scarcity of molecule-language pair datasets. At this point, datasets have been released which are 1) small and scraped from existing databases, 2) large but noisy and constructed by performing entity linking on the scientific literature, and 3) built by converting property prediction datasets to natural language using templates. In this document, we detail the *L+M-24* dataset, which has been created for the Language + Molecules Workshop shared task at ACL 2024. In particular, *L+M-24* is designed to focus on three key benefits of natural language in molecule design: compositionality, functionality, and abstraction.[1]

1 Introduction

The world faces an enormous number of problems in the coming decades on scales of complexity never-before-seen, in areas such as climate change, healthcare, and pandemics. To address these issues, we need to discover inventive scientific solutions which are scalable, flexible, and inexpensive. Broadly speaking, many of these problems will require molecular solutions from the chemistry domain, such as developing new drugs (e.g. kinase inhibitors (Ferguson and Gray, 2018)), materials (e.g. organic photovoltaics (Kippelen and Brédas, 2009)), and chemical processes (Zhong et al., 2023). These solutions exist in extremely large search spaces, which makes AI tools a necessity.

Language-molecule models have emerged as an exciting direction for molecular discovery and understanding (Edwards et al., 2021; Zeng et al., 2022; Edwards et al., 2022; Su et al., 2022; Liu et al., 2022; Xu et al., 2023; Christofidellis et al., 2023; Liu et al., 2023b; Luo et al., 2023; Zhao et al., 2023c; Seidl et al., 2023). However, training these models is challenging due to the scarcity of molecule-language pair datasets. At this point, datasets have been released which are 1) small and scraped from existing databases (Edwards et al., 2021; Zeng et al., 2023; Liu et al., 2023a,c; Pei et al., 2023), 2) large but noisy and constructed by performing entity linking on the scientific literature (Zeng et al., 2022; Su et al., 2022), and 3) template-based built on prediction datasets (Zhao et al., 2023a; Fang et al., 2023). Approaches utilizing pseudo-data have also been attempted (Chen et al., 2023a). These approaches have helped remedy the problem of data scarcity in this domain; however, these approaches frequently ignore key benefits of natural language: 1) compositionality, 2) abstraction, and 3) functionality (Zhang et al., 2023). To this end, for the Language + Molecules Workshop at ACL 2024, we release *L+M-24*, which we construct to test these three goals, particularly compositionality, using recently released data sources (Zhao et al., 2023b; Kosonocky et al., 2023; Wishart et al., 2023). *L+M-24* is divided into four categories with important applications in the small-molecule domain: 1) Biomedical, 2) Light and Electricity, 3) Human Interaction and Organoleptics, and 4) Agriculture and Industry. Improving understanding of these applications can have important implications in problems such as drug discovery, climate issues, more efficient and green industrial processes, and improved food production.

2 Task Formulation

The dataset is primarily intended for language↔molecule translation, which consists of two tasks: generating 1) a caption given a molecule and 2) a molecule given a description.

[1]The dataset, finetuned baseline, and evaluation code are released publicly at github.com/language-plus-molecules/LPM-24-Dataset through HuggingFace.

Proceedings of the 1st Workshop on Language + Molecules (L+M 2024), pages 1–10

2.1 Designing for Compositionality, Abstraction, and Function

Overall, we focused on four primary categories of importance: 1) Biomedical, 2) Light and Electricity, 3) Human Interaction and Organoleptics, and 4) Agriculture and Industry. These categories and three properties from each are displayed in Table 1. The biomedical category is focused on drug properties, functions, and interaction with proteins. Light and electricity is focused on the ability for a molecule to produce or absorb light or electricity. Human interaction and organoleptics focuses on the effect and experience molecules cause in humans. Agriculture and industry focuses on molecules used in industrial processes and food production.

Based on our data sources (below), the properties we have selected already encode a large degree of functionality, enhanced by our manual curation. Further, since these properties are generally short phrases indicating functionality, they are also abstract and apply to many molecules (e.g., "insecticide"). For compositionality, we explicitly select certain pairs of properties which we hold out of the dataset. For example, a molecule may share two properties which are desirable together (e.g., low toxicity and fungicidal). *L+M-24* will help to evaluate whether model's can generalize to unseen compositions of properties.

Biomedical	**Human Interaction**
anti neoplastic	pungent
glaucoma treatment	bitter
capillarigenic	nephrotoxic
Light and Electricity	**Agriculture and Industry**
photoelectric conversion	herbicide
photopolymerization	emulsifier
dielectric	carcinogen

Table 1: Example properties in the dataset. Antineoplastic drugs are used to treat cancer. Glaucoma is a group of eye diseases. Capillarigenic means producing or causing capillaries. Pungent means having a strong taste or smell. Nephrotoxic is toxicity in the kidneys. Photoelectric conversion is the conversion of light into electricity. Photopolymerization is the process through which monomers are linked together through a photochemical reaction. A dielectric is a poor conductor of electricity but can be polarized. A herbidicde is toxic to plants. An emulsifier stabilizes an emulsion. A carcinogen is an agent capable of causing cancer. The full property list and number of occurrences is available in the online data repository.

3 Data Sources

We constructed our dataset using three different databases. We will first describe the process we used to extract information from each, followed by our overall strategy for adding hierarchy into the dataset. We want to deeply thank the authors of these resources for making them publicly available for the community.

3.1 PubChem

We used properties extracted from PubChem (Kim et al., 2016, 2019) as described in (Zhao et al., 2023c). Properties from this approach include odor, taste, and decomposition. We note these properties consist of molecule-specific descriptions, which the other data sources do not provide.

3.2 Chemical Function (CheF)

Here, we used functional properties extracted from patent literature by Kosonocky et al. (2023). This allowed us to capture molecules from the patent literature in addition to the scientific literature. Here, we started with CheF prefinal_v3[2]. We created a set of properties from both CheF's property summarizations and from the ChatGPT summarization source. For the summarization source, we also applied the WordNet lemmatizer (Bird et al., 2009) for deduplication. After obtaining a list of properties, we removed properties pertaining to less than 100 molecules. We then kept properties falling into the categories of "X-icide", "anti-X", "X treatment", "X modulators", "X inhibitors", "X agonists", "X antagonists", "light", and "electricity." We manually removed uninformative labels which were too broad or didn't describe enough function. Further, we manually corrected errors in label naming and duplication.

3.3 ChemFOnt: the chemical functional ontology resource

In addition to CheF, we also take advantage of another new chemical function data resource: ChemFOnt (Wishart et al., 2023). From this datasource, we collect three categories: health effect relations, organoleptic effect relations, and role relations.

4 Dataset Details

To convert these properties to natural language, we follow a template-based procedure using GPT-4 (OpenAI, 2023) generated compositional templates.

[2]obtained via personal communication.

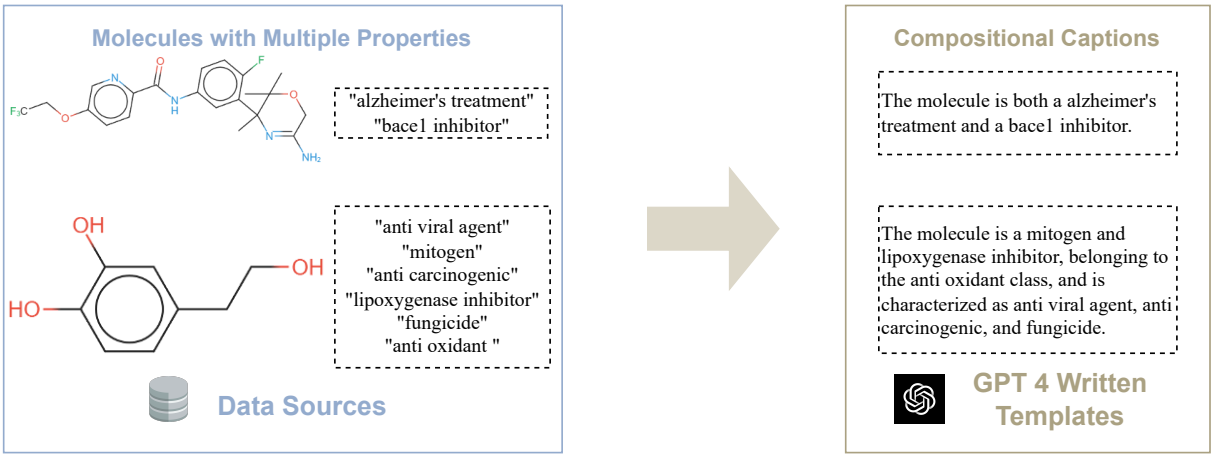

Figure 1: Example descriptions created for molecules from the training set.

4.1 Template Generation

We utilize GPT-4 (OpenAI, 2023) to generate specific templates for each combinations of molecular properties. Specifically, we manually write six templates: "*The molecule is a <0>.*"; "*It belongs to the <1> class of molecules.*"; "*It has an effect on <2>.*"; "*It impacts <3>.*"; "*The molecule is <4>.*"; and "*The molecule has a <5>.*". Subsequently, we use GPT-4 to generate a unique sentence template for each possible combination by rephrasing up to six combinations of the six initial templates as a single sentence. Ultimately, this process results in the generation of 917 distinct templates. The templates were manually checked and corrected to have a matching standard. The prompts and in-context examples for GPT-4 are given in the Appendix.

4.2 Converting Templates to Descriptions

For all properties in *L+M-24*, we first assigned them to possible templates based on their category or by individual consideration. Certain properties (e.g., polymerization, decomposition) were expressed in sentence format, so we did not use templates. Given a molecule with n properties, we first looked for a template that had the correct slots (e.g., <0>, <2>, and <2>) for its properties. When we found possible templates, we picked one at random and used it to generate a sentence for the molecule's properties. If there were no matching templates, we split the properties into two separate equal-sized groups and tried with each group. We return the concatenation of the two sentence templates as the molecule description. Note this process can repeat multiple times.

We note that we are also releasing a version of the dataset with 5 captions for each molecule. In this case, we split group sizes at random. Further, we split sentences apart 50% of the time (even when there were matching templates) to increase caption diversity.

4.3 Splitting

Duplicate molecules are merged using RDKit (Landrum, 2021) and molecules which cannot be processed are removed. We split the data by first examining property combinations. 20% of combinations are witheld into the evaluation set. From molecules in the remaining 80%, we keep 80% for training and put 20% in evaluation. The evaluation set is split into two tasks: molecule captioning and molecule generation. For each task, only one modality will be released prior to the shared task results.

The training set consists of 160,492 molecule-description pairs. For the evaluation set, both molecule generation and captioning contain 21,839 pairs. Further, special splits are released for the training set which allow for validation using the training data. They are constructed using the same procedure as the official evaluation dataset.

5 Evaluation Metrics

Overall, we adopt the evaluation metrics proposed by Edwards et al. (2022). However, we include invalid molecules in the calculations of FTS metrics (setting the score to zero for invalid molecules). We also add a uniqueness metric to the generated molecules for held-out combinations of properties (Polykovskiy et al., 2020). Further, we also look at property-specific precision, recall, and F-1 scores. These scores are calculated by matching tokenized names in the generated captions. These scores are

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							11.30
MolT5-Small	70.9	51.2	74.5	55.8	54.4	70.1	10.79
MolT5-Base	73.8	53.5	75.0	55.9	53.9	71.8	8.53
MolT5-Large	76.9	55.6	77.7	58.0	55.7	74.3	10.06
Meditron-7B	79.2	57.6	79.7	60.2	57.5	75.7	11.91

Table 2: Molecule captioning results on the validation split of *L+M-24*. Rouge scores are F1 values.

	Overall			Biomedical			Light+Electro			Human Interaction			Agr.+Industry			Held-out Combos		
Model	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
MolT5-Small	84.83	8.24	7.88	85.13	23.23	23.33	62.42	4.85	3.27	96.77	0.57	0.56	95.00	4.32	4.36	0.00	0.00	0.00
MolT5-Base	64.11	9.94	9.46	79.58	23.89	24.02	16.08	5.82	3.36	63.94	5.01	5.18	96.85	5.05	5.27	0.00	0.00	0.00
MolT5-Large	59.57	12.49	11.71	70.27	26.99	26.87	16.96	10.90	7.39	62.77	5.99	6.27	88.29	6.06	6.31	0.00	0.00	0.00
Meditron-7B	33.60	16.33	16.81	57.19	33.96	35.27	26.51	16.48	17.49	29.54	7.52	7.07	21.18	7.35	7.40	12.35	0.29	0.56

Table 3: Property-specific molecule captioning results on the validation split of *L+M-24*.

Model	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
	X-icides			**Toxins**			**Light**			**Electricity**			**X-inhibitors**			**anti-X**		
MolT5-Small	100.00	0.00	0.00	100.00	0.00	0.00	24.85	9.69	6.54	100.00	0.00	0.00	3.42	0.43	0.09	1.96	0.00	0.00
MolT5-Base	100.00	0.00	0.00	67.45	8.51	8.84	28.00	11.51	6.52	4.17	0.12	0.20	2.20	0.58	0.11	9.70	0.23	0.15
MolT5-Large	100.00	0.00	0.00	69.42	10.29	10.85	15.77	12.28	8.16	18.14	9.52	6.62	8.90	2.28	1.13	4.32	1.16	0.61
Meditron-7B	100.00	0.00	0.00	48.79	11.75	11.05	29.10	20.64	20.64	23.93	12.33	14.34	35.69	19.91	22.65	14.79	9.34	8.98
	X-modulator			**X-agonist**			**X-antagonist**			**X-treatment**			**X-disease**			**X cancer**		
MolT5-Small	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00	55.49	1.99	1.70	87.44	50.08	49.94	71.86	21.03	24.27
MolT5-Base	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00	58.90	2.25	1.80	94.61	55.16	59.18	45.06	25.49	24.54
MolT5-Large	21.30	0.58	0.88	5.91	1.96	1.23	14.30	0.58	0.42	14.27	2.67	2.22	97.18	81.07	81.86	65.76	52.06	51.56
Meditron-7B	42.43	21.24	24.98	39.19	23.23	26.35	34.22	18.98	21.15	28.75	11.35	15.13	97.34	81.11	82.02	79.80	68.65	72.62

Table 4: Selected subproperty group-specific molecule captioning results on the validation split of *L+M-24*.

Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth	100.0	100.0	0.00	100.0	100.0	100.0	0.00	11.26	100.0
MolT5-Small	56.56	0.00	56.34	64.22	58.10	37.44	NaN	0.49	80.52
MolT5-Base	68.38	0.00	44.79	76.03	65.23	47.46	NaN	7.06	100.0
MolT5-Large	56.42	0.00	55.40	75.70	65.01	39.51	17.52	7.69	99.44
Meditron-7B	69.40	0.01	46.49	77.16	69.34	50.07	2.46	7.80	99.63

Table 5: Molecule generation results on the validation split of *L+M-24*. The FCD and Text2mol metrics are computed using only syntactically valid molecules. We found FCD suffers from numerical instability for the small and base models.

Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Uniqueness↑	Validity↑
Ground Truth	100.0	100.0	0.00	100.0	100.0	100.0	0.0	23.05	100.0	100.0
MolT5-Small	22.80	0.00	54.14	8.99	5.19	3.48	NaN	5.79	10.14	39.79
MolT5-Base	29.51	0.00	48.91	38.78	19.73	14.21	NaN	21.60	5.13	100.0
MolT5-Large	24.37	0.00	63.44	41.56	24.23	15.71	NaN	23.77	12.72	97.82
Meditron-7B	28.04	0.00	53.44	40.90	27.42	16.82	3.91	22.46	74.81	98.58

Table 6: Molecule generation results on the subset of held-out combinations from the validation split of *L+M-24* (2107 data points).

further aggregated across specific properties (e.g., inhibitors, X-icides, etc.) and the four broad categories. Aggregations are performed by averaging scores (i.e., macro-F1). We further compute these scores specifically for held-out combinations of properties.

6 Benchmarks

MolT5 models (Edwards et al., 2022) were finetuned for 20 epochs on the "split_train" data split and evaluated on the "split_valid", both of which are available online. Huggingface's transformers (Wolf et al., 2019) was used for finetuning with a learning rate of 2e-5 and weight decay of 0.01. A batch size of 128 was used for small and base models, and a batch size of 48 for large models. Further, Meditron-7B (Chen et al., 2023b) was finetuned for 5 epochs with a context length of 930, 2e-6 learning rate, and batch size of 8/16 (molecule/caption generation). Models are released online. Results for captioning are reported in Tables 2, 3 and 4. Tables 5, and 6 shows results for molecule generation.

Overall, the dataset proves to be fairly challenging for these naively finetuned models. On captioning, Meditron-7B achieves a maximum overall F-1 score of 16.81 for property identification (Table 3). However, overall it has a much higher precision than recall, indicating the model only labels

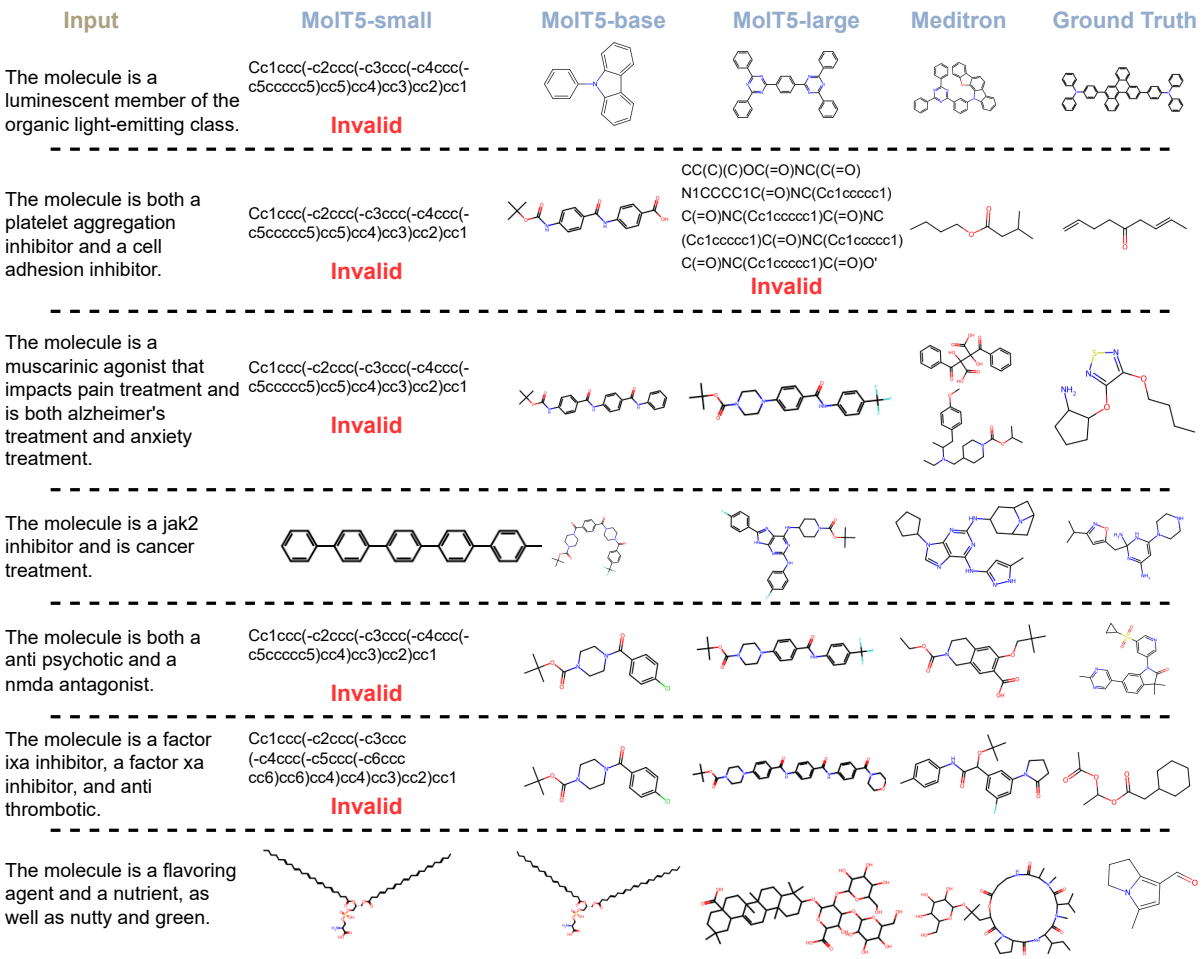

Figure 2: Examples of molecules generated by different models for never-before-seen property combinations.

a molecule with a certain property when having higher confidence. Certain classes of molecules, such as X-icides, are never identified (Table 4). Other classes, such as toxins or electricity, show emergent behavior as model size scales. Interestingly, the models appear to be fairly capable at linking molecules to certain diseases or cancers. We find that, likely due to poor performance on individual properties, only the largest model succeeds on predicting held-out combos, and with poor results. Additionally, we find that the Text2Mol metric, as trained on ChEBI-20, shows poor domain transfer to *L+M-24*.

The models are able to capture a number of useful properties, such as electroluminescence, diabetes treatment, non-alcoholic fatty liver disease, and emulsifiers. In some cases, the model captures important characteristics about the molecule but uses differing language. This poses a challenge for our evaluation metrics. For example, a molecule identified in the ground truth as an anti tumor agent is identified as being a cancer treatment by the model. In particular, the models appear to struggle with rarer properties, which are common in our dataset formulation and in the chemical domain as a whole. They also struggle with identifying molecule-protein interactions (e.g., "monoamine reuptake inhibitor"), although Meditron shows a large performance jump.

For the molecule generation task, we also find the dataset to be challenging. We show results generated by different models on never-before-seen property combinations in Figures 2 and 3. We believe the difficulty is for two reasons. First, common property combinations may have structurally very different molecules which exhibit those properties, making evaluation difficult. Second, the model may not grasp rare properties well. Overall, this results in the naively finetuned models producing similar outputs to many different prompts. Further, as expected, performance falls on unseen property combinations and larger models prove

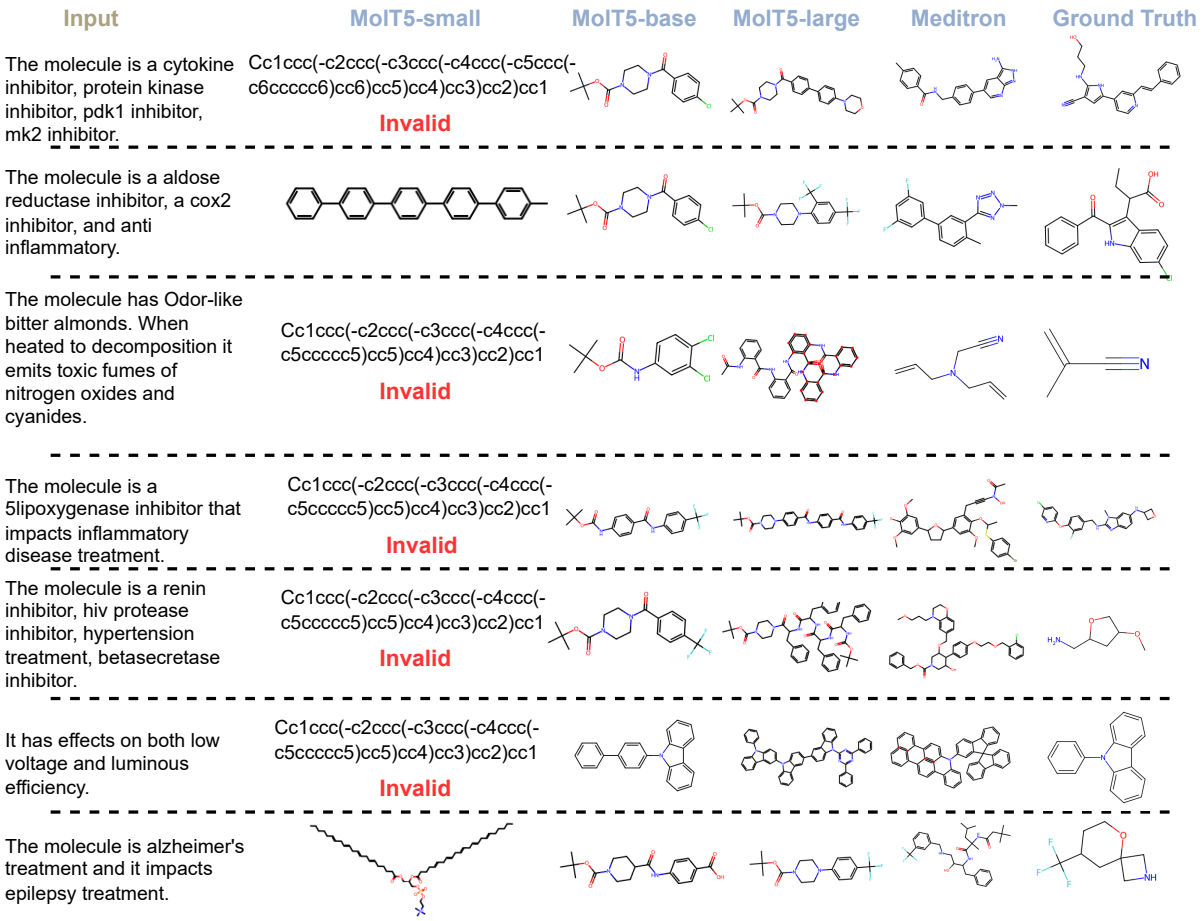

Figure 3: Examples of molecules generated by different models for never-before-seen property combinations.

more effective (Table 6).

7 Future Directions

Overall, this dataset proves to be quite challenging. We find that some specific properties in particular are challenging for the model. This may be because the model understands these properties, but is unwilling to use them in its descriptions due to the training procedure. This limitation may be addressed with more sophisticated decoding algorithms or by better finetuning methods. Future work will also likely benefit from incorporating other modalities, such as proteins, to provide better understanding to the model for some property types. Notably, certain properties display what may be emergent behavior; scaling training data or model size may yield non-linear improvements.

In this dataset, we focus on composition, abstraction, and function. Future work may also wish to integrate other recent trends: instruction-following and dialogue (Fang et al., 2023; Cao et al., 2023; Zeng et al., 2023; Zhao et al., 2024; Zhang et al., 2024; Yu et al., 2024), tool use (Boiko et al., 2023; Bran et al., 2023), additional molecule representations (e.g., 3D (Tang et al., 2023)), additional modalities (Xu et al., 2023), or molecule editing (Su et al., 2022). Further, we note the need for improved evaluation metrics, especially in the case of molecule generation for function where there may be many possible outputs. Specific methods for improving compositionality may be another fruitful avenue for research (Yellinek et al., 2023). It may also be interesting to use molecule-language instruction-following models within larger search frameworks, such as ChemReasoner (Sprueill et al., 2023, 2024).

8 Conclusion

In this manuscript, we describe the process for creating the *L+M-24* dataset. *L+M-24* is designed to focus on three key benefits of natural language in molecule design: compositionality, functionality,

and abstraction. It is the featured shared task at the First Language + Molecules Workshop at ACL 2024.

Acknowledgements

We would like to thank the authors and creators of our data sources for freely allowing the use of their work. We would like to acknowledge NVIDIA Corporation's contributions to this work through a grant of NVIDIA A100 Tensor Core GPUs. This work is supported by the Molecule Maker Lab Institute: an AI research institute program supported by NSF under award No. 2019897, by DOE Center for Advanced Bioenergy and Bioproducts Innovation U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DESC0018420, and by AI Algriculture: the Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021- 32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture. 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 National Science Foundation, the U.S. Department of Energy, and the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *ArXiv preprint*, abs/2304.05332.

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*.

Yuhan Chen, Nuwa Xi, Yanrui Du, Haochun Wang, Chen Jianyu, Sendong Zhao, and Bing Qin. 2023a. From artificially real to real: Leveraging pseudo data from large language models for low-resource molecule discovery. *arXiv preprint arXiv:2309.05203*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. *arXiv preprint arXiv:2301.12586*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Fleur M Ferguson and Nathanael S Gray. 2018. Kinase inhibitors: the road ahead. *Nature reviews Drug discovery*, 17(5):353–377.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, et al. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109.

Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, et al. 2016. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

Bernard Kippelen and Jean-Luc Brédas. 2009. Organic photovoltaics. *Energy & Environmental Science*, 2(3):251–261.

Clayton W Kosonocky, Claus O Wilke, Edward M Marcotte, and Andrew D Ellington. 2023. Mining patents with large language models demonstrates congruence of functional labels and chemical structures. *arXiv preprint arXiv:2309.08765*.

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

Pengfei Liu, Yiming Ren, and Zhixiang Ren. 2023a. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *arXiv preprint arXiv:2308.06911*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2022. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789*.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023b. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023c. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798*.

Yizhen Luo, Kai Yang, Massimo Hong, Xingyi Liu, and Zaiqing Nie. 2023. Molfm: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*.

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.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. 2023. Enhancing activity prediction models in drug discovery with the ability to understand human language. *ArXiv preprint*, abs/2303.03363.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*.

Henry W Sprueill, Carl Edwards, Mariefel V Olarte, Udishnu Sanyal, Heng Ji, and Sutanay Choudhury. 2023. Monte carlo thought search: Large language model querying for complex scientific reasoning in catalyst design. *arXiv preprint arXiv:2310.14420*.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *ArXiv preprint*, abs/2209.05481.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark B Gerstein. 2023. Mollm: A unified language model to integrate biomedical text with 2d and 3d molecular representations. *bioRxiv*, pages 2023–11.

David S Wishart, Sagan Girod, Harrison Peters, Eponine Oler, Juan Jovel, Zachary Budinski, Ralph Milford, Vicki W Lui, Zinat Sayeeda, Robert Mah, et al. 2023. Chemfont: the chemical functional ontology resource. *Nucleic Acids Research*, 51(D1):D1220–D1229.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hanwen Xu, Addie Woicik, Hoifung Poon, Russ B Altman, and Sheng Wang. 2023. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications*, 14(1):738.

Nir Yellinek, Leonid Karlinsky, and Raja Giryes. 2023. 3vl: using trees to teach vision & language models compositional concepts. *arXiv preprint arXiv:2312.17345*.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

Zheni Zeng, Bangchen Yin, Shipeng Wang, Jiarui Liu, Cheng Yang, Haishen Yao, Xingzhi Sun, Maosong Sun, Guotong Xie, and Zhiyuan Liu. 2023. Interactive molecular discovery with natural language. *arXiv preprint arXiv:2306.11976*.

Zijie Zeng, Xinyu Li, Dragan Gasevic, and Guanliang Chen. 2022. Do deep neural nets display human-like attention in short answer scoring? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 191–205, Seattle, United States. Association for Computational Linguistics.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. 2024. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. 2023. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023a. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *bioRxiv*, pages 2023–05.

Lawrence Zhao, Carl Edwards, and Heng Ji. 2023b. What a scientific language model knows and doesn't know about chemistry. In *NeurIPS 2023 AI for Science Workshop*.

Wenyu Zhao, Dong Zhou, Buqing Cao, Kai Zhang, and Jinjun Chen. 2023c. Adversarial modality alignment network for cross-modal molecule retrieval. *IEEE Transactions on Artificial Intelligence*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. 2024. Chemdfm: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818*.

Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xianrui Zhong, Xuan Liu, Hongxiang Li, Jinfeng Xiao, Minhao Jiang, Vivian Hu, Xuan Wang, Heng Ji, Martin Burke, Huimin Zhao, and Jiawei Han. 2023. Reaction miner: An integrated system for chemical reaction extraction from textual data. In *Proc. The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP2023) Demo Track*.

A Prompts and examples for GPT4

- Prompts: You are an expert in the chemical domain whose task is to create templates to describe the properties of molecules. You will be challenged with a list of different cases. Each case willl have a list of **templates**, and a **question**. Each template will describe certain properties. Your goal is to generate a new template in a sentence based on all the previous templates.

- Case 1: Templates: - The molecule is a <0>. - It belongs to the <1> class of molecules. Answer: The molecule, characterized as a <0>, falls under the <1> category of chemical compounds.

- Case 2: Templates: - It has an effect on <2>. - It impacts <3>. Answer: It impacts <2> and has an effect on <3>.

- Case 3: Templates: - The molecule is <4>. - The molecule has a <5>. Answer: The molecule is <4>. and has a <5>.

- Case 4: Templates: - The molecule is a <0_1>. - The molecule is a <0_2>. Answer: The molecule is a <0_1> and exhibits <0_2> properties.

- Case 5: Templates: - It belongs to the <1_1> class of molecules. - It belongs to the <1_2> class of molecules. Answer: The molecule is in the <1_1> class of compounds, characterizing it as a member of the <1_2> family.

B Additional Dataset Statistics

Here, we give a brief description of properties in the dataset. Table 7 shows the number of property-molecule pairs for different property classes. Figure 4 breaks the dataset down into different property classes. More details can be found in the dataset repository.

Group	Property-Molecule Pair Count
Total	1512865
Biomedical	776712
anti-X	24884
Modulators	2787
Inhibitors	23257
Agonists	1161
Antagonists	3172
Treatments	53070
Disease	316380
Cancer	41456
Inducers	31
Preventive	0
Blocker	47
Drug	260
X-genic	172
X-tropic	17
X-lytic	84
Relaxant	40
Binder	4
Stimulant	60
Depressant	52
health_effect_relations	309532
Light and Electricity	14077
Light	11069
Electricity	3008
Human Interaction	27457
Toxins	1070
organoleptic_effect_relations	20501
Agric. and Industry	694619
X-icides	809
role_relation	693648

Table 7: Number of property-molecule pairs for different property groups.

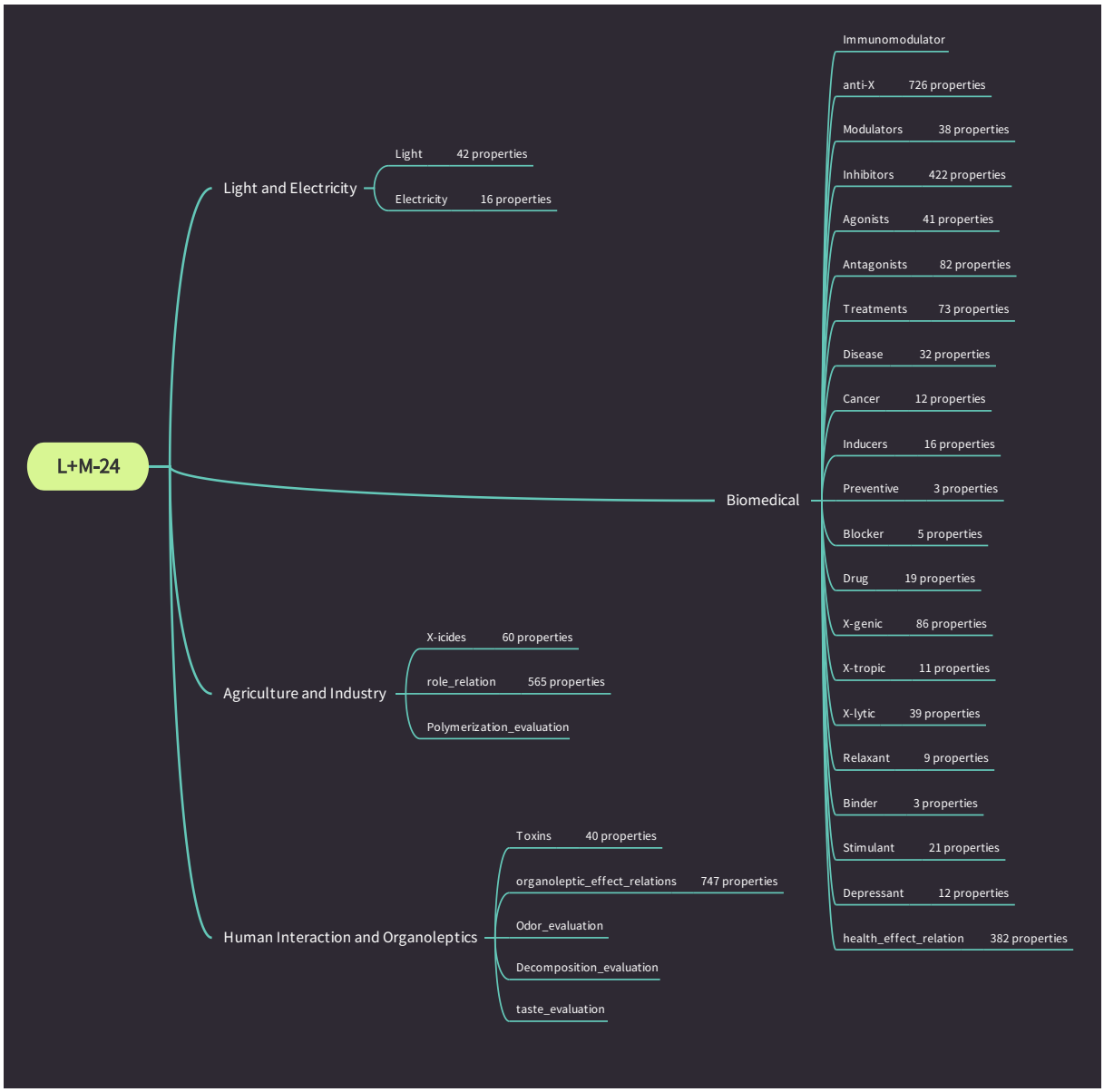

Figure 4: Breakdown of different property classes in *L+M-24*.