Enhanced BioT5+ for Molecule-Text Translation: A Three-Stage Approach with Data Distillation, Diverse Training, and Voting Ensemble

Qizhi Pei[1], Lijun Wu[2]*, Kaiyuan Gao[3], Jinhua Zhu[4], Rui Yan[1,5]*
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Microsoft Research [3]Huazhong University of Science and Technology
[4]University of Science and Technology of China
[5]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education
{qizhipei,ruiyan}@ruc.edu.cn apeterswu@gmail.com
im_kai@hust.edu.cn teslazhu@mail.ustc.edu.cn

Abstract

This paper presents our enhanced BioT5+ method for the Language + Molecules shared task at the ACL 2024 Workshop. The task involves "translating" between molecules and natural language, including molecule captioning and text-based molecule generation using the *L+M-24* dataset. Our method consists of three stages. In the first stage, we distill data from various models. In the second stage, combined with *extra* version of the provided dataset, we train diverse models for subsequent voting ensemble. We also adopt Transductive Ensemble Learning (TEL) to enhance these base models. Lastly, all models are integrated using a voting ensemble method. Experimental results demonstrate that BioT5+ achieves superior performance on *L+M-24* dataset. On the final leaderboard[1], our method (team name: **qizhipei**) ranks **first** in the text-based molecule generation task and **second** in the molecule captioning task, highlighting its efficacy and robustness in translating between molecules and natural language. The pre-trained BioT5+ models are available at https://github.com/QizhiPei/BioT5.

1 Introduction

With the development of Large Language Models (LLMs) (Touvron et al., 2023a,b; OpenAI, 2023; Taori et al., 2023; Chowdhery et al., 2023), the integration of molecules with natural language has garnered increasing attention in recent research efforts (Edwards et al., 2021, 2022; Zeng et al., 2022; Luo et al., 2023; Tang et al., 2023; Liu et al., 2023b; Zhao et al., 2023; Liu et al., 2023a,d,c; Pei et al., 2023, 2024a). Notably, two critical generative tasks have emerged: molecule captioning (*i.e., mol2text*) and text-based molecule generation (*i.e., text2mol*) (Edwards et al., 2022). These tasks are pivotal for biologists and chemists, as they facilitate the interpretation and creation of molecular structures through natural language descriptions.

Table 1: Statistics of *L+M-24* dataset. We use $\mathcal{B}$ to represent molecule-text paired datasets and $\mathcal{D}$ to represent datasets only containing molecules or text.

Split	Symbol	*mol2text*	*text2mol*
#Training	$\mathcal{B}$	126,864	126,864
#Training-*extra*	$\mathcal{B}_+$	533,953	533,953
#Validation	$\mathcal{B}_{valid}$	33,696	33,696
#Test	D_{test}	21,942	21,805

To leverage the advantages of natural language for molecular design and understanding (Zhang et al., 2024; Liao et al., 2024; Pei et al., 2024b; AI4Science and Quantum, 2023), Language + Molecules Workshop at ACL 2024 has been organized. A shared molecule-text translation task and the corresponding paired dataset are presented to accelerate research in this field.

1.1 Dataset Description

In the provided *L+M-24* dataset, each sample is a molecule-text pair, with the molecule represented by SMILES (Weininger, 1988; Weininger et al., 1989) and the text generated from collected molecular properties based on templates written by GPT-4 (OpenAI, 2023). An *extra* version of *L+M-24* is also available, with each molecule having five additional captions. We use the training split of this version (*i.e.*, training-*extra*) and remove dupli-

* Corresponding authors: Lijun Wu (apeterswu@gmail.com) and Rui Yan (ruiyan@ruc.edu.cn)

[1]https://language-plus-molecules.github.io/#leaderboard

cates. The fundamental statistics of the *L+M-24* are shown in Table 1, with more details about its construction described in Edwards et al. (2024).

1.2 Task Description

Mol2text The goal of the *mol2text* task is to generate a caption for a given molecule. Participants are required to submit generated captions for the test split of *mol2text*. Evaluation metrics include widely used text generation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), in addition to the Text2Mol metric (Edwards et al., 2021, 2022). These metrics assess the similarity between the generated molecular captions and the ground truth. Classification metrics including Precision, Recall, and F-1 value are also used to evaluate property-specific *mol2text* results.

Text2mol The goal of the *text2mol* task is to generate a molecule that fits a given description. Participants are required to submit the generated molecule SMILES for the test split of *text2mol*. Evaluation metrics include BLEU (Papineni et al., 2002), exact match percentages, Levenshtein distance, fingerprints (FTS) similarity score (MACCS (Durant et al., 2002), RDK (Landrum et al., 2023), Morgan (Rogers and Hahn, 2010)), FCD score (Preuer et al., 2018), Text2Mol (Edwards et al., 2021, 2022) score, and validity. These metrics evaluate the similarity between the generated molecule and the ground truth.

1.3 Overview of our Method

Our proposed method, enhanced version of BioT5+, is designed to tackle the *mol2text* and *text2mol* tasks using a comprehensive three-stage approach. The first stage involves data distillation, where we generate synthetic datasets from trained models to enrich the training data. In the second stage, we perform diverse training by fine-tuning various models on different combinations of distilled and extra datasets. We also employ Transductive Ensemble Learning (TEL) to further enhance these models by leveraging unlabeled data. In the final stage, we integrate these models using a voting ensemble method, which selects the best predictions based on perplexity scores. This multi-faceted strategy ensures that our models are robust, diverse, and capable of achieving superior performance across both tasks.

2 Methodology

In this section, we give a detailed introduction to our three-stage methodology.

Notations. We use SELFIES (Krenn et al., 2020) as the sequence representation of the molecule. Compared to SMILES, SELFIES is a more robust molecular representation, which is beneficial for molecule generation tasks such as *text2mol*, as it ensures the generation of 100% valid molecules. The SMILES in the *L+M-24* dataset are converted to corresponding SELFIES using *selfies* toolkit[2]. Let M and T denote molecular SELFIES and text descriptions, respectively, and $\mathcal{M}$ and $\mathcal{T}$ denote the corresponding collection of all sequences. Let $\mathcal{B} = \{(m_i, t_i)\}_{i=1}^{|\mathcal{B}|}$ represent the molecule-text pairs from the training split of *L+M-24*, and $\mathcal{B}_+ = \{(m_i, t_i)\}_{i=1}^{|\mathcal{B}_+|}$ represent the molecule-text pairs from training-*extra* split, where $m_i \in \mathcal{M}$, $t_i \in \mathcal{T}$, and $|\mathcal{B}|$ and $|\mathcal{B}_+|$ represent the size of $\mathcal{B}$ and $\mathcal{B}_+$, respectively. Let $\mathcal{D}^m = \{m_j\}_{j=1}^{|\mathcal{D}^m|}$ denote the collection of molecules from the PubChem (Kim et al., 2019) database, where $m_j \in \mathcal{M}$ and $|\mathcal{D}^m| = 800$K represents the number of sampled molecules. The text in $\mathcal{T}$ follows a specific format, so we directly use $\mathcal{D}^t = \{t_j\}_{j=1}^{|\mathcal{B}_+|}$, where the text $t_j \in \mathcal{B}_+$.

Our goal is to develop a *mol2text* translation model $f : \mathcal{M} \mapsto \mathcal{T}$, which generates a caption from T for a given molecule from M, and a reverse *text2mol* translation model $g : \mathcal{T} \mapsto \mathcal{M}$. In this paper, all models follow the T5 (Raffel et al., 2020)-large architecture. Our method consists of the following three stages:

Stage-1: Data Distillation. First, we train a *mol2text* translation model f_0 and a *text2mol* translation model g_0 on $\mathcal{B}$. Then, we use f_0 and g_0 to build the synthetic dataset:

$$\begin{aligned}\overline{\mathcal{B}}_{self} = &\{(m, f_0(m)) \mid m \in \mathcal{D}_m\} \\ &\cup \{\, (g_0(t),\ t\ ) \mid t \in \mathcal{D}_t\ \}\,.\end{aligned}$$

To further improve the diversity of the distilled data, we also use the officially provided Meditron-7B (Chen et al., 2023) *mol2text* model[3] f_{med} to build another synthetic dataset:

$$\overline{\mathcal{B}}_{med} = \{(m, f_{med}(m)) \mid m \in \mathcal{D}_m\}\,.$$

[2] https://github.com/aspuru-guzik-group/selfies

[3] https://huggingface.co/language-plus-molecules/Meditron7b-smiles2caption-LPM24

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**
BioT5+	**79.8**	**57.9**	**81.2**	**61.7**	**58.4**	**77.7**	11.36

Table 2: Results for *mol2text* task on the validation set of *L+M-24*.

	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	29.83	3.48	3.12	15.13	4.18	4.29	12.42	4.85	3.27	46.77	0.57	0.56	45.00	4.32	4.36	0.00	0.00	0.00
MolT5-Base	35.36	5.18	4.69	14.58	4.84	4.97	16.08	5.82	3.36	**63.94**	5.01	5.18	**46.85**	5.05	5.27	0.00	0.00	0.00
MolT5-Large	33.32	7.72	6.95	15.27	7.94	7.82	16.96	10.90	7.39	62.77	5.99	6.27	38.29	6.06	6.31	0.00	0.00	0.00
Meditron-7B	25.27	11.56	16.8	23.86	14.91	35.00	26.51	16.48	17.49	29.54	7.52	7.07	21.18	7.35	7.40	12.35	0.29	0.56
BioT5+	**35.50**	**20.69**	**20.93**	**56.91**	**38.22**	**39.27**	**36.20**	**27.43**	**28.22**	29.46	**9.09**	**8.42**	19.41	**8.03**	**7.82**	**17.61**	**0.73**	**1.40**

Table 3: Results for property-specific *mol2text* task on the validation set 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	0.00	0.00	0.00	0.00	0.00	0.00	24.85	9.69	6.54	0.00	0.00	0.00	3.42	0.43	0.09	1.96	0.00	0.00
MolT5-Base	0.00	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	0.00	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	0.00	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
BioT5+	0.00	0.00	0.00	47.93	**13.42**	**12.55**	**38.29**	**30.32**	**30.68**	**34.12**	**24.53**	**25.76**	**48.00**	**31.05**	**32.58**	**33.96**	**13.04**	**15.34**
	X-modulator			**X-agonist**			**X-antagonist**			**X-treatment**			**X-disease**			**X cancer**		
MolT5-Small	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	55.49	1.99	1.70	87.44	50.08	49.94	71.86	21.03	24.27
MolT5-Base	0.00	0.00	0.00	0.00	0.00	0.00	0.00	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
BioT5+	**55.32**	**42.83**	**44.76**	**53.02**	**36.96**	**37.09**	**50.06**	**32.79**	**34.18**	**46.83**	**18.99**	**24.47**	96.61	**81.75**	**82.05**	77.77	**73.48**	**75.25**

Table 4: Results for selected subproperty group-specific *mol2text* task on the validation set of *L+M-24*.

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MolT5-Small	66.82	48.29	72.80	54.44	53.33	68.14
MolT5-Base	69.83	50.56	73.34	54.55	52.86	69.86
MolT5-Large	73.63	53.20	75.79	56.47	54.42	72.16
Meditron-7B	75.16	54.72	77.97	58.75	56.33	73.69
BioT5+	**75.58**	**54.77**	**79.41**	**59.89**	**57.46**	**75.43**

Table 5: Results for *mol2text* task on the test set of *L+M-24*.

Table 6: Model combinations. $\mathcal{B}_+ \rightarrow \mathcal{B}$ means the model is first trained on $\mathcal{B}_+$ followed by $\mathcal{B}$.

Model	Dataset	Initialization
f_0, g_0	$\mathcal{B}$	BioT5+
f_1, g_1	$\mathcal{B} \cup \mathcal{B}_+$	BioT5+
f_2, g_2	$\mathcal{B} \cup \overline{\mathcal{B}}_{self}$	BioT5+
f_3, g_3	$\mathcal{B} \cup \overline{\mathcal{B}}_{med}$	BioT5+
f_4, g_4	$\mathcal{B}_+ \rightarrow \mathcal{B}$	BioT5+
f_1^*, g_1^*	$\overline{\mathcal{B}}^*$	f_1, g_1
f_2^*, g_2^*	$\overline{\mathcal{B}}^*$	f_2, g_2
f_3^*, g_3^*	$\overline{\mathcal{B}}^*$	f_3, g_3
f_4^*, g_4^*	$\overline{\mathcal{B}}^*$	f_4, g_4

In summary, despite $\mathcal{B}$, we have three additional synthetic datasets: $\mathcal{B}_+$, $\overline{\mathcal{B}}_{self}$, and $\overline{\mathcal{B}}_{med}$.

Stage-2: Diverse Training. Based on the datasets mentioned above, we train various types of *mol2text* and *text2mol* models on different combinations of these datasets, as shown in Table 6. We first train $\{f_i\}_{i=1}^4$ based on the distilled datasets in Stage-1. Then we adopt the Transductive Ensemble Learning (TEL) method to get $\{f_i^*\}_{i=1}^4$, which involves predicting labels for unlabeled data and subsequently fine-tuning models on these predictions to enhance performance (Wang et al., 2020). Taking the *mol2text* models as an example (the *text2mol* models follow a similar process), for each f_i in $\{f_i\}_{i=1}^4$, we select τ top-performing checkpoints $\{f_{ij}\}_{j=1}^\tau$ based on their validation BLEU scores from its training trajectory. We use $\{f_{ij}\}_{j=1}^\tau$ to caption the molecules from $\mathcal{B}_{valid}$ and $\mathcal{D}_{test}$, resulting in two synthetic datasets:

$$\overline{\mathcal{B}}_{valid,i} = \{(m, f_{ij}(m)) \mid m \in \mathcal{B}_{valid}, 1 \leq j \leq \tau\},$$
$$\overline{\mathcal{B}}_{test,i} = \{(m, f_{ij}(m)) \mid m \in \mathcal{D}_{test}, 1 \leq j \leq \tau\}.$$

Then we fine-tune model f_i^* on $\overline{\mathcal{B}}^* = \cup_{i=1}^4 \{\overline{\mathcal{B}}_{valid,i} \cup \overline{\mathcal{B}}_{test,i}\}$, where f_i^* is initialized from f_i. f_i^* generally performs better

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
BioT5+*	**73.97**	**0.01**	**40.87**	77.69	70.51	**51.58**	3.22	**13.83**	**100.0**
BioT5+	73.10	**0.01**	41.47	**78.06**	**70.93**	51.49	3.29	13.73	**100.0**

Table 7: Results for *text2mol* task on the validation set of *L+M-24*. * denotes model from TEL in Stage-2.

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
BioT5+*	**33.35**	**0.10**	**43.65**	41.52	28.05	17.53	**3.52**	22.91	51.05	**100.0**
BioT5+	31.89	**0.10**	46.14	**42.57**	**29.50**	**18.01**	3.88	**23.77**	48.22	**100.0**

Table 8: Results for *text2mol* task on the subset of held-out combinations from the validation set of *L+M-24*. * denotes model from TEL in Stage-2.

Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Validity↑
MolT5-Small	55.44	0.0	57.21	63.06	56.83	36.69	nan	81.03
MolT5-Base	67.04	0.0	45.71	74.61	63.7	46.29	nan	99.89
MolT5-Large	55.31	0.0	56.47	74.14	63.4	38.54	17.63	99.12
Meditron	68.84	**0.01**	46.47	75.59	67.66	48.72	**2.44**	99.54
BioT5+	**73.17**	**0.01**	**41.05**	**76.05**	**68.70**	**50.05**	3.13	**100.0**

Table 9: Results for *text2mol* task on the test set of *L+M-24*.

than f_i as f_i^* due to its ability to leverage the collective knowledge and complementary strengths of ensemble learning, leading to improved generalization and robustness. The comparison between f_i and f_i^* is shown in Table 10. In total, as shown in Table 6, we obtain eight types of models $\{f_i\}_{i=1}^4$ and $\{f_i^*\}_{i=1}^4$ in Stage-2.

Stage-3: Voting Ensemble. In the final stage, we combine the strengths of the models trained in Stage-2 through a voting ensemble approach. This method leverages multiple models to improve the reliability and accuracy of the predictions. We illustrate this process using the *mol2text* test dataset as an example, but the same methodology applies to *text2mol* and validation datasets.

Let $\mathcal{F} = \{\hat{f}_j\}_{j=1}^{|\mathcal{F}|}$, where $\hat{f}_j$ is derived from the Stage-2 models in Table 6. Each $\hat{f}_j$ generates captions for the molecules in $\mathcal{D}_{test}$, resulting in a corresponding set of datasets:

$$\mathcal{S} = \left\{ \overline{\mathcal{D}}_{test,j} = \left\{ (m, \hat{f}_j(m)) \mid m \in \mathcal{D}_{test} \right\} \mid \hat{f}_j \in \mathcal{F} \right\}.$$

For each dataset in $\mathcal{S}$, we compute the perplexity (PPL) score of each caption using all models in $\mathcal{F}$. The perplexity of a model $\hat{f}_j$ on a molecule-text pair (m, t) is defined as:

$$\text{PPL}_{\hat{f}_j}(m, t) = \exp\left(-\frac{1}{N} \sum_{i=1}^{N} \log P_{\hat{f}_j}(t_i \mid t_{<i}, m) \right),$$

where N is the length of the caption t, and $P_{\hat{f}j}(t_i \mid t_{<i}, m)$ is the probability of the i-th token in the caption given the preceding tokens and the molecule m. Next, we average the PPL scores across all models in $\mathcal{F}$ for each molecule-text pair in each dataset in $\mathcal{S}$. The average perplexity for a given molecule-text pair (m, t) in the dataset $\overline{\mathcal{D}}_{test,j}$ is calculated as:

$$\overline{\text{PPL}}(m, t) = \frac{1}{|\mathcal{F}|} \sum_{k=1}^{|\mathcal{F}|} \text{PPL}_{\hat{f}_k}(m, t).$$

Finally, for each molecule m in the test dataset $\mathcal{D}_{test}$, we select the caption $\hat{t}(m)$ with the lowest average PPL from each dataset in $\mathcal{S}$ as the final prediction: $\hat{t}(m) = \arg\min_{(m,t)\in\mathcal{S}} \overline{\text{PPL}}(m, t)$. This selection process ensures that we leverage the most reliable caption according to the ensemble's evaluation. By using this voting ensemble approach, we improve the robustness and accuracy of the predictions, leveraging the strengths of multiple models trained in Stage-2.

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
BioT5+†	79.53	57.44	80.69	61.10	57.98	76.86
BioT5+*	79.63	57.70	80.95	61.18	58.11	77.06
BioT5+	79.79	57.87	81.23	61.70	58.38	77.71

Table 10: Ablation results for *mol2text* task on the validation set of *L+M-24*. † denotes the model before TEL in Stage-2. * denotes the model after TEL in Stage-2.

Model Configuration. Following Pei et al. (2024a), we pre-train a large version of BioT5+ with 789M parameters, which is an enhanced version of BioT5 (Pei et al., 2023) with improved molecular understanding capabilities. As in Table 6, model $\{f_i\}_{i=0}^4$ are fine-tuned from this pre-trained BioT5+ model, and model $\{f_i^*\}_{i=1}^4$ are fine-tuned from $\{f_i\}_{i=1}^4$. We employ a greedy decoding strategy for all results, which selects the token with the highest probability at each time step without incorporating randomness or exploring multiple hypotheses.

3 Experiments

In this section, we present our main results for the *mol2text* and *text2mol* tasks. Baseline results on the validation set are derived from Edwards et al. (2024), and test set results are sourced from the official leaderboard. An ablation study is also conducted in Section 4 to demonstrate the efficacy of our methodology.

Mol2text. Results on the validation set are shown in Table 2, and results for the test set are shown in Table 5. Our method achieves the best performances on all metrics except for Text2Mol (Edwards et al., 2021, 2022) metric, with a BLEU-2 of 79.80 on the validation set and 75.58 on the test set. For the Text2Mol score on the validation set, both Meditron (Chen et al., 2023) and our method exceed the ground truth score (11.30), with our method slightly underperforming Meditron. The property-specific and selected subproperty group-specific results on the validation set are presented in Table 3 and Table 4, where our method also outperforms the baselines in nearly all metrics. These results show that the generated captions of our method are highly accurate.

Text2mol. Unlike *mol2text*, our voting ensemble in Stage 3 for the *text2mol* task does not improve all metrics simultaneously. Therefore, we also report the BioT5+* results which is the model from TEL in Stage-2. Results on the validation and test sets are presented in Table 7 and 9. Results on the subset of held-out combinations from the validation set are shown in Table 8. Our method achieves superior performances in most metrics, demonstrating its efficacy and generalization ability.

4 Ablation Study

To validate the effectiveness of our TEL training and voting ensemble, we conduct an ablation study for the *mol2text* task on the validation set of *L+M-24*. The results, shown in Table 10, indicate that the model after TEL (BioT5+*) yields better results than model before TEL (BioT5+†). The BioT5+ model, derived from voting ensemble in Stage-3, achieves the best results overall.

5 Conclusion

In this paper, we introduce our enhanced BioT5+ model for the shared task of the Language + Molecules Workshop at ACL 2024. We adopt a three-stage approach: data distillation, diverse training, and voting ensemble. Our method effectively leverages diverse datasets and advanced ensemble techniques to enhance model performance in both molecule captioning and text-based molecule generation tasks. Experimental results show that our approach achieves superior performance across various evaluation metrics, highlighting the potential of our enhanced BioT5+ model for integrating molecules and text.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC Grant No. 62122089), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China. Qizhi Pei is supported by the Outstanding Innovative Talents Cultivation Funded Programs 2023 of Renmin University of China.

References

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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.

Carl Edwards, Tuan Manh 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 375–413. Association for Computational Linguistics.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+m-24: Building a dataset for language+ molecules@ acl 2024. *arXiv preprint arXiv:2403.00791*.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 595–607. Association for Computational Linguistics.

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.

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 et al. 2023. Rdkit: Open-source cheminformatics. GitHub release.

Chang Liao, Yemin Yu, Yu Mei, and Ying Wei. 2024. From words to molecules: A survey of large language models in chemistry. *arXiv preprint arXiv:2402.01439*.

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

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

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023b. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023c. Molxpt: Wrapping molecules with text for generative pre-training. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023d. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15623–15638. Association for Computational Linguistics.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023. Molfm: A multimodal molecular foundation model. *CoRR*, abs/2307.09484.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024a. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *arXiv preprint arXiv:2402.17810*.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, Yue Wang, Zun Wang, Tao Qin, and Rui Yan. 2024b. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, Singapore. Association for Computational Linguistics.

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. *J. Chem. Inf. Model.*, 58(9):1736–1741.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2020. Transductive ensemble learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6291–6298.

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.

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):862.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. 2024. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. 2023. GIMLET: A unified graph-text model for instruction-based molecule zero-shot learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.