Next Token Prediction is All You Need

Yunjae Won
Graduate School of AI, KAIST
Seoul, South Korea
yunjae.won@kaist.ac.kr

Jinho Park
Graduate School of AI, KAIST
Seoul, South Korea
binlepain178@kaist.ac.kr

ABSTRACT

This progress report concerns several tasks regarding metabolic reaction prediction, namely, reaction direction prediction, missing metabolite prediction, and destination prediction. While, at a glance, each task seems to require separate solutions, we notice that these tasks all require the same fundamental knowledge, understanding the underlying metabolic reaction equation (e.g., $A + B \rightarrow C$). We thus unify the three separate tasks into a single next-token-prediction task, and approach the problem using a single unified transformer decoder-only architecture. This enables parameter sharing between tasks, and also allows data augmentation between tasks. Experimental results demonstrate the effectiveness of our approach.

1 INTRODUCTION

The three tasks is defined as the following:

- **Reaction Direction Prediction**: Given a set of source and destination metabolites, determine its reaction direction, TRUE ($\rightarrow$) or FALSE ($\leftarrow$). The metabolites are described by an unique id with a range of [0,3472]. Task performance is evaluated by accuracy.
- **Missing Metabolite Prediction**: Given a set of source and destination metabolites, determine the missing source metabolite, listing top 10 candidates in the order of estimated likelihood. The metabolites are described by an unique id range of [0,3470]. Evaluation is done using the Hits@10 and MRR@10 (Mean Reciprocal Rank) metric.
- **Destination Prediction**: Given a set of source, determine the set of destination metabolite. The metabolites are described by an unique id range of [0,3472]. Evaluation is done using the average precision, recall, and F1-Score metrics.

All three tasks involve 3,007 samples of training data. We tackle the three tasks in a single-shot manner, formulating all three tasks as a unified next-token-prediction task. The contributions of this project are the following:

- Our proposed *NTP* is enables parameter sharing between tasks, reducing the net computational requirement compared to a naive approach of constructing a task-wise deep neural network architecture.
- It also allows data augmentation between tasks, encouraging cross-task knowledge transfer.

2 PROPOSED METHOD

The main motivation of our method is that all three tasks require understanding the underlying metabolic reaction equation, namely the form of $A + B \rightarrow C$. Inspired by recent deep-learning based approaches for modeling metabolic reactions [2], we use a single decoder-only transformer architecture [6] to address all three tasks.

Given a set of metabolite ids $\mathcal{M} = \{0, ..., 3472\}$, we are concerned with the source metabolites $\mathcal{S} \subset \mathcal{M}$ and destination metabolites $\mathcal{D} \subset \mathcal{M}$. We define the sequence of source metabolites $S = [s_0, s_1, ...]$ where $s_i \in \mathcal{S}$ is ordered by its value in ascending order. Likewise, we define the sequence of destination metabolites $D = [d_0, d_1, ...]$ where $d_i \in \mathcal{D}$. We also define a set of special tokens as the following:

- $< TASK\_ID_0 >$: A task descriptor ID indicating the first task, the direction prediction task.
- $< TASK\_ID_1 >$: A task descriptor ID indicating the second task, the missing metabolite prediction task.
- $< TASK\_ID_2 >$: A task descriptor ID indicating the last task, the destination prediction task.
- $< DELIMITER >$: A delimiter that indicates the end of sequence of source metabolites, and also indicates the start of sequence of destination metabolites.
- $< EOS >$: An end-of-sequence token that indicates the end of sequence of destination metabolites.
- <→>: A token indicating that the reaction direction is TRUE ($\rightarrow$).
- <←>: A token indicating that the reaction direction is FALSE ($\leftarrow$).

We train our transformer architecture θ using standard cross-entropy loss [1], where we define the target token to be predicted as the following:

- **Reaction Direction Prediction**: Given a sequence of $[< TASK\_ID_0 >, s_0, s_1, ..., < DELIMITER >, d_0, d_1, ..., < EOS >]$, predict the next target token, which is one of <→> or <←>.
- **Missing Metabolite Prediction**: Given a sequence of $[< TASK\_ID_1 >, s_0, s_1, ..., < DELIMITER >, d_0, d_1, ..., < EOS >]$, predict the next target token, which is an element of $\mathcal{S}$.
- **Destination Prediction**: Given a sequence of $[< TASK\_ID_2 >, s_0, s_1, ..., < DELIMITER >]$, predict the next sequence of tokens $[d_0, d_1, ..., < EOS >]$.

We also employ a data augmentation strategy. From the **Reaction Direction Prediction** dataset, we generate three new datasets:

- **(1)** We switch the source metabolites and the destination metabolites, and flip the reaction direction accordingly.
- **(2)** For each sequence with the TRUE reaction direction, we randomly sample one source metabolite and augment the **Missing Metabolite Prediction** dataset.
- **(3)** For each sequence with the TRUE reaction direction, we utilize the destination metabolites to augment the **Destination Prediction** dataset.

From the **Missing Metabolite Prediction** dataset, we generate three new datasets:

Table 1: Training results for Reaction Direction Prediction task. Our NTP method significantly outperforms the baseline methods. The last row reports the performance of a single model trained on all three tasks at once with additional data augmentation techniques.

Method	Accuracy [%]
Random	50
Counting	67.29
NTP (*Ours, w.o. Data Augmentation*)	81.55
NTP (*Ours, with Data Augmentation*)	**83.88**

Table 2: Training results for Missing Metabolite Prediction task. Our NTP method significantly outperforms the baseline methods. The last row reports the performance of a single model trained on all three tasks at once with additional data augmentation techniques.

Method	Hits@10 [%]	MRR@10
Random Guessing	0.28	0.0007
Random Guessing based on Interactions	11.29	0.04
NTP (*Ours, w.o. Data Augmentation*)	**47.6**	**0.41**
NTP (*Ours, with Data Augmentation*)	47.29	0.40

- **(1)** For each source metabolite, we generate new data samples by considering the selected source metabolite as the missing node.
- **(2)** From the training dataset, since we can combine the missing node and the source metabolites and reconstruct the original source metabolites, we utilize this to augment the **Reaction Direction Prediction** dataset samples with the TRUE reaction directions.
- **(3)** Likewise, we utilize the reconstructed source metabolites to augment the **Destination Prediction** dataset.

From the **Destination Prediction** dataset, we generate two new datasets:

- **(1)** Since we are provided with the source and destination metabolites from the training dataset, we convert it to the **Reaction Direction Prediction** dataset by labeling each samples as the TRUE reaction direction.
- **(2)** Likewise, we sample one source metabolite and consider it as the missing node to augment the **Missing Metabolite Prediction** dataset.

By doing so, we effectively generate 21,139 new samples from the original training dataset with 9,021 samples.

3 EXPERIMENTS

We implemented our method using Pytorch [4], using a gpt-2 [5] architecture with the following configuration:

- `embd_pdrop = 0, attn_pdrop = 0, resid_pdrop = 0`
- `n_positions = 256`
- `n_embd = 768, n_layer = 12, n_head = 12`
- `activation_function = gelu_new`

Table 3: Training results for Destination Prediction task. Our NTP method significantly outperforms the baseline methods. The last row reports the performance of a single model trained on all three tasks at once with additional data augmentation techniques.

Method	Avg. Precision [%]	Avg. Recall [%]	Avg. F1-Score [%]
Random Guessing	0.068	0.07	0.07
Random Guessing based on Interactions	4.69	4.30	4.16
NTP (*Ours, w.o. Data Augmentation*)	**35.1**	35	34.64
NTP (*Ours, with Data Augmentation*)	**35.1**	**35.32**	**34.94**

We train our model for 10 epochs using the AdamW optimizer [3], with maximum learning rate 2e-5 linearly warmed up for the first 10% steps, and decayed under a cosine scheduler.

We compare our method with the baselines described in the assignment documentation, with an additional three different transformer models each separately trained on the training dataset of the three tasks, without any data augmentation techniques. This allows us to directly measure the effect of parameter sharing between tasks, allowing cross-task knowledge transfer.

Table 1 shows the accuracy of each methods on the reaction prediction prediction task. Table 2 shows the Hits10 and MRR10 of each methods on the missing metabolite prediction task. Finally, Table 3 describes the average precision, recall, and F1-Score fo each methods on the destination prediction task. Our method significantly outperforms the baseline methods provided by the assignment documentation. We also find that our method performs comparably to the models separately trained on each tasks' dataset without any data augmentation techniques. While we could not observe any significant cross-task knowledge transfer effects, we were able to handle all three tasks using a single transformer architecture, which performed comparably to training three different models on each datasets.

4 CONCLUSIONS

The proposed method *NTP* enables approaching the three metabolic reaction prediction tasks using a single unified transformer architecture. It has the following advantages:

- It enables parameter sharing between tasks, reducing the computational burden.
- It enables cross-task knowledge transfer, allowing data-augmentation between different tasks.

Future research can investigate the effect of pre-training the embedding layer with graph neural network training techniques. We hope that our research paves way for investigating a foundational model specialized for bio-chemical reaction tasks.

REFERENCES

[1] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114, 1952.
[2] David Kreutter, Philippe Schwaller, and Jean-Louis Reymond. Predicting enzymatic reactions with a molecular transformer. *Chemical science*, 12(25):8648–8659, 2021.
[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library.

In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[5] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[6] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

A APPENDIX

A.1 Labor Division

The team performed the following tasks

- Exploratory Data Analysis [Park]
- Method Formulation [Park, Won]
- Experiments [Park, Won]
- Report Writing [Park, Won]