AI4Science Research Proposal Template

Star (Xinxin) Liu, Stellaire X. Robinhood
Department of Computer and Information Science, University of Pennsylvania, PA, USA
Department of Materials Science and Engineering, University of Pennsylvania, PA, USA

Abstract

The abstract paragraph should focus on the main points: what keywords, what research field, what problem, how do we solve, use what fashion techniques, and reach what score. We should **highlight** our novelty and strong improvement, as well as the view which potentially aspires future works.

1 Introduction

The summary of:

- related works, the conflicts or gap between previous works,
- why are they so important,
- and what we contribute to them.

1.1 The field/category of this work

1.2 Latest fashion in this field

1.3 Limitations of previous works in this field

LaTeX style

1.4 We solve what kind of gap/problems

1.5 And how to solve

The detail of methods and strategies can be find at Sections Section 4.1, SI Section B below.

2 Related Work

2.1 Encoder/Representation

2.2 General Model Architecture

2.3 Problem from where and How did they find it

2.4 Previous Tasks and Datasets

2.5 Data-preprocessing techniques

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

What we do after Encoding? E.g.: any embedding techniques

3 Background (Headings: first level)

Assuming the reviewer and editor are not specialist in our segment, we should give some explanation for the fundamentals. We cannot expect everybody understand every thing in our work, but at the very first we have to let them understand where our work is promising and fancy!!

4 Framework of Methodology

4.1 Find the science/engineering problems

And gives a general and original idea about how to solve it

4.2 Model Architecture Design

Representation Learning:
The key rule for designing a ReprL model is:

- Similar to traditional (CompSci) models
- Components are like, One-to-one correspondence: e.g. graphs' nodes &edges $\rightarrow$ molecules' atoms &bonds

Statistical Learning:
We should give **very rigorous** definitions for our datasets, sub-sets, sample, data-points, feature space, distributions, hierarchy, transformed space, assumptions, applicable conditions, …etc., to give a very clear assertion.

Algorithm:
Maybe we should introduce any recursive proof to clarify our algorithm/automata is context-free grammar (CFG) or optimal theoretically or something.

4.3 Training Strategy

Common strategies:

- Hyper-parameter selection:
  - Grid Search
  - Random Search
  - Bayesian Opt (e.g. most easy one, by optuna)
    - How do we choose the returned criteria like what losses how weighted
    - How do we choose the initial guess
    - How to penalize the worse-performed trial
  - Modified BayesOpt
  - Gaussian Process
  - Simulating Annealing
- Optimizer selection:
  - Adam
  - SGD
  - Any modified Adam/SGD

4.4 New Problems Solving

What we may meet when training or applying our new tools…

5 Experimental Evaluation

General describe why we use such evaluations in both scientific and algorithmic views…

5.1 Datasets and Tasks (Citations within the text)

In general, we should introduce **1-2** fashion tasks/benchmarks to prove our model/algorithm is *state-of-the-arts* on current Evaluation Framework.

Then, we had better design a **new and exclusive** benchmark to illustrate this model/algorithm has more general application scenarios then ever before.

5.1.1 Previous Benchmark 1:

For small molecule drug properties prediction benchmark, `ADMET TDC-2` may be found at

`https://tdcommons.ai/overview`

`https://github.com/mims-harvard/TDC`

Citation follows the format:

```
@article {Velez-Arce2024tdc,
  author = {Velez-Arce, Alejandro and Huang, Kexin and Li, Michelle and Lin, Xiang
and Gao, Wenhao and Fu, Tianfan and Kellis, Manolis and Pentelute, Bradley L. and
Zitnik, Marinka},
  title = {TDC-2: Multimodal Foundation for Therapeutic Science},
  elocation-id = {2024.06.12.598655},
  year = {2024},
  doi = {10.1101/2024.06.12.598655},
  publisher = {Cold Spring Harbor Laboratory},
  URL = {https://www.biorxiv.org/content/early/2024/06/21/2024.06.12.598655},
  journal = {bioRxiv}
}
```

5.1.2 Previous Benchmark 2:

For excample, SKEMPI v2.0:

> Justina Jankauskaitė, et al. (2019) investigated SKEMPI v2.0 …

5.1.3 Our Task and newly designed Benchmark

The main concern of xxx issue is xxx. And we find it is a general issue for these xxx models. So, we designed this xxx to show the potentials in xxx applications.

Table (and Figures) Styles

Table 1: Sample table title.

Sample Info		Size (μm)
Name	Description	
Dendrite	Input terminal	~ 100
Axon	Output terminal	~ 10
Soma	Cell body	up to 10^6

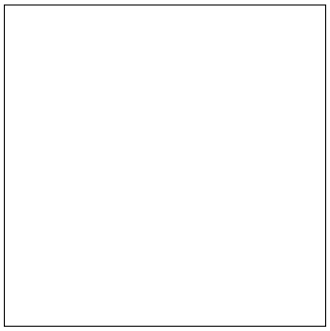

Figure 1: Sample figure caption.

5.2 Ablation Study (Footnotes)

For our novelly modified part of the model structure /algorithm lines, we should give an ablation study[1] in the paragraph.

Also, we can here give the reason that why we choose such strategy/sub-structure for the model. For example, we apply `GRU` rather than `LSTM` in *ChemXTree* as the feature "amplifier", so that we gave the comparison in **Runtime** and **AUROC scores** on a subset of benchmark.[2]

5.3 Results We Expect

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.
You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

5.4 Plausible Extensions

5.5 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

6 Discussion

The shortcomes (but not fatal ones) of the current version of work. However, it is not avoidable for now because of the limitations of current technology or cost badget or dataset systematic error or some other things.

7 Conclusion (Preparing PDF files)

Our Work is plausible and meaningful!!!

[1] Namely, give the comparison w/o our modified structures/components

[2] Acknoledgement: Thanks Dr. Yuzhi Xu `url`{https://scholar.google.com/citations?user=jiUlHrUAAAAJ&hl} from NYU teaching me how to do research at my very starting point!!

References

[1] Y. Xu *et al.*, "ChemXTree: A Feature-Enhanced Graph Neural Network Decision Tree Framework for ADMET Prediction," *Journal of Chemical Information and Modeling*, vol. 64, 2024, doi: 10.1021/acs.jcim.4c01186.

[2] X. Liu, K. Fan, X. Huang, J. Ge, Y. Liu, and H. Kang, "Recent advances in artificial intelligence boosting materials design for electrochemical energy storage," *Chemical Engineering Journal*, vol. 490, p. 151625, 2024, doi: https://doi.org/10.1016/j.cej.2024.151625.

A Appendix: Full Results for Downstream Evaluations

No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal. No SI for Proposal.

B Appendix: Balabala

No SI for Proposal. No SI for Proposal.