nature portfolio

Corresponding author(s): Yan Chen, Lei Wang, Hang Xiao
Last updated by author(s): 3-9-2025

Machine Learning Checklist V 1.1

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form is intended to provide structure for consistency and transparency in reporting of works using or developing Machine Learning models. Some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

1. Availability and reproducibility of Code and Data

Please select all that apply regarding the availability of the data and code used in the study.

☐ Code will be included in a CodeOcean capsule.

☑ The **source code** is included in the submission or available in a public repository:
https://github.com/xiaohang007/SLICES/tree/main/MatterGPT/

☐ A **compiled standalone version** of the software is included in the submission or available in a public repository:
URL

☑ A **test dataset** and instructions/scripts for replicating the results are included in the submission or available in a public repository:
https://doi.org/10.5281/zenodo.16879675

☑ A **Readme file** with instructions for installing and running the code is included in the submission or available in a public repository:
https://github.com/xiaohang007/SLICES/tree/main/MatterGPT/

☑ The code is made available to reviewers during review.

☑ **Pretrained models** are used in the study and accessible through:
https://doi.org/10.5281/zenodo.16879675

☐ **Pretrained models** are used in the study and are not accessible.

☑ The paper contains information on how to obtain code and data after publication.

2. Datasets

A. All data sources are listed in the paper.
   ☑ Yes
   ☐ No

B. The train, test and validation datasets are publicly available, and links/accession numbers have been provided in the manuscript or supplementary materials.
   ☑ Yes
   ☐ No

C. We have reported and discussed potential dataset biases in the paper. Where applicable, appropriate mitigation strategies were used.

☑ Yes Dataset curation section discusses filtering criteria and bias toward zero band gap values

☐ No _____

D. The data cleaning and preprocessing steps are clearly and fully described, either in text or as a code pipeline.

☑ Yes Dataset curation section provides detailed preprocessing steps

☐ No _____

E. Instances of combining data from multiple sources are clearly identified, and potential issues mitigated.

☐ Yes _____

☑ No Single dataset used (Alex-20 from Alexandria database)

3. Model and training

A. What model architecture is the current model based on? Transformer (autoregressive Transformer-decoder architecture)

B. A Model Card is provided[1].

☐ Yes

☑ No

C. The model clearly splits data into different sets for training (model selection), validation (hyperparameter optimization), and testing (final evaluation).

☑ Yes

☐ No

D. The method of data splitting (e.g. random, cluster- or time-based splitting, forward cross-validation) is clearly stated.

☑ Yes 9:1 training/testing split mentioned in Methods section

☐ No _____

E. The data splitting mimics anticipated real-world applications.

☑ Yes Standard materials science evaluation approach

☐ No _____

F. The data splitting procedure has been chosen to avoid data leakage.

☑ Yes Standard train/test split prevents data leakage

☐ No _____

[1] https://huggingface.co/docs/hub/model-cards

G. The interpretability of the model has been studied and clearly validated.
☑ Yes Chemical space exploration and interpretability section provides attention analysis
☐ No

4. Evaluation

A. The performance metrics used are described and justified in the paper.
☑ Yes Evaluation criteria section defines validity, uniqueness, novelty, MAE, and TPR metrics
☐ No

B. Cross-validation of the results is included.
☐ Yes
☑ No

C. Community-accepted benchmark datasets/tasks are used for comparisons.
☐ Yes
☑ No The paper uses established evaluation metrics but does not employ standardized benchmark datasets due to their absence in the crystal generation field

D. Baseline comparisons to simple/trivial models (for example, 1-nearest neighbour, random forest, most frequent class) are provided.
☐ Yes
☑ No Simple baselines would perform poorly on conditional crystal generation (low validity, no property control). Comparisons focus on sophisticated generative models which provide more meaningful benchmarks.

E. Benchmarks with current state-of-the-art are provided.
☑ Yes Table S4 (MatterGen efficiency) and CrystaLLM/CDVAE performance comparisons.
☐ No

F. Ablation experiments are included.
☐ Yes
☑ No All components work together as an integrated system for conditional crystal generation.

G. The model has been tested on a fully independent dataset.
☑ Yes
☐ No

5. Computational resources

A. The paper contains information on hardware/computing resources that were used.
☑ Yes
☐ No

B. The paper includes information on the computational costs in terms of computation time, parallelization or carbon footprints estimates.
☑ Yes
☐ No