

Corresponding author(s): Haohong Song, Qisheng Wu
Last updated by author(s): 17-09-2926

Machine Learning Checklist v1.1

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:
URL
 - ☐ 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/10.6084/m9.figshare.33476008>
 - ☐ A **Readme file** with instructions for installing and running the code is included in the submission or available in a public repository:
URL
 - ☐ The code is made available to reviewers during review.
 - ☒ **Pretrained models** are used in the study and accessible through:
URL<https://doi.org/10.6084/m9.figshare.33476008>
 - ☐ **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 methods and force fileds training
- ☐ No _____
- D. The data cleaning and preprocessing steps are clearly and fully described, either in text or as a code pipeline.
- ☒ Yes methods and force fileds training
- ☐ No _____
- E. Instances of combining data from multiple sources are clearly identified, and potential issues mitigated.
- ☒ Yes methods and force fileds training
- ☐ No _____

3. Model and training

- A. What model architecture is the current model based on? neuroevolution potential (NEP)
- B. A Model Card is provided¹.
- ☐ 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 methods and force fileds training
- ☐ No _____
- E. The data splitting mimics anticipated real-world applications.
- ☒ Yes methods and force fileds training
- ☐ No _____
- F. The data splitting procedure has been chosen to avoid data leakage.
- ☒ Yes methods and force fileds training
- ☐ No _____

¹ <https://huggingface.co/docs/hub/model-cards>

G. The interpretability of the model has been studied and clearly validated.

☒ Yes methods and force fields training

☐ No _____

4. Evaluation

A. The performance metrics used are described and justified in the paper.

☒ Yes methods and force fields training

☐ No _____

B. Cross-validation of the results is included.

☒ Yes

☐ No

C. Community-accepted benchmark datasets/tasks are used for comparisons.

☐ Yes _____

☒ No This work targets a specific materials system and property regime, comparisons are performed on our curated DFT dataset and application-specific validation tasks rather than on community-accepted general benchmark datasets.

D. Baseline comparisons to simple/trivial models (for example, 1-nearest neighbour, random forest, most frequent class) are provided.

☐ Yes _____

☒ No Baseline comparisons to simple/trivial models such as 1-nearest neighbour, random forest, or most frequent class are not provided, as these are not meaningful baselines for MLFF energy/force prediction and downstream atomistic simulations. Instead, we benchmark against DFT-labeled data and physically relevant baseline models.

E. Benchmarks with current state-of-the-art are provided.

☐ Yes _____

☒ No Our evaluation focuses on DFT references and application-relevant baselines for the specific materials system studied in this work.

F. Ablation experiments are included.

☐ Yes _____

☒ No The focus of this work is the development and validation of an MLFF for the target system, with evaluation based on energy/force accuracy and physically relevant downstream tests rather than component-wise architectural ablations.

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