Text to text (T2) crystal relaxation

Team: T2-Relax

Daniel T. Speckhard[1], Aniket Phutane[2], Alexander Kister[3], Heike Quosdorf[3], Claudia Draxl[1], José A. Márquez[1]

[1] *Humboldt-Universität zu Berlin, Zum Großen Windkanal 2, 12489 Berlin, Germany*
[2] *Helmholtz-Zentrum Berlin für Materialien und Energie*
[3] *Bundesanstalt für Materialforschung und -prüfung (BAM)*

Introduction

Crystal structure relaxation is an integral part of high throughput computational workflows, which often optimize the geometry of thousands of structures, selecting a DFT code and the corresponding computational parameters [97, 98]. In this project, we use the LeMatTraj dataset [99] to train an LLM (T5-Transformer) [100] to perform crystal structure relaxation. In order to feed the model textual input data, we use the ASE library [101] to transform the crystal structure into a crystallographic input file (CIF), which is the standard in the Inorganic Crystallographic Database (ICSD) [102], and aim at predicting the relaxed structure in the CIF format via Δ-learning [103]. This work was largely inspired by the MD-LLM-1 work [104], which predicted relaxed protein structures using an LLM. To improve the LLM's performance, we fine-tune the model on the relaxation task.

Results

The T5-Transformer is an LLM that was trained on the C4 text corpus for natural language processing (NLP) tasks (e.g. translation, summarization, question answering, etc.). The model is fine-tuned to optimize text-based metrics (ROUGE [105] and BLEU [106]) to ensure that much of the CIF file remains unchanged after relaxation (i.e., the atomic elements in the structure). Using a weighted sum, the model is also fine-tuned to minimize the RMSE of the lattice parameters and atomic positions, which are givenrelative to the basis vectors. Since the crystal structures in the LeMatTraj dataset vary widely ... We apply Δ-learning [103] to predict the differences in the lattice parameters (lengths and angles between them) from the initial structures to the final structures as obtained by DFT. In this way, the target distribution is more Gaussian. The general framework of the approach is shown in Figure 24.

The results are shown in Table 2. The MAE and RMSE of the predicted change in the atomic positions for each structure in the dataset are shown in Table 3. In general, we see small errors, which show that the LLM manages to generally predict the relaxed structured with a good degree of accuracy. Without the ROUGE and BLEU metrics to keep the elements the same from the input to the predicted CIF output, the LLM often removes atoms or changes them. More thorough testing is required to benchmark the method, but the initial results shown here are promising.

Parameter	MAE	MSE	RMSE
Δa (Å)	0.053	0.006	0.080
Δb (Å)	0.053	0.006	0.080
Δc (Å)	0.074	0.020	0.140
$\Delta\alpha$ (°)	0.086	0.036	0.189
$\Delta\beta$ (°)	0.101	0.070	0.265
$\Delta\gamma$ (°)	0.160	0.104	0.322

Table 2: Errors in predicting the lattice cell parameters.

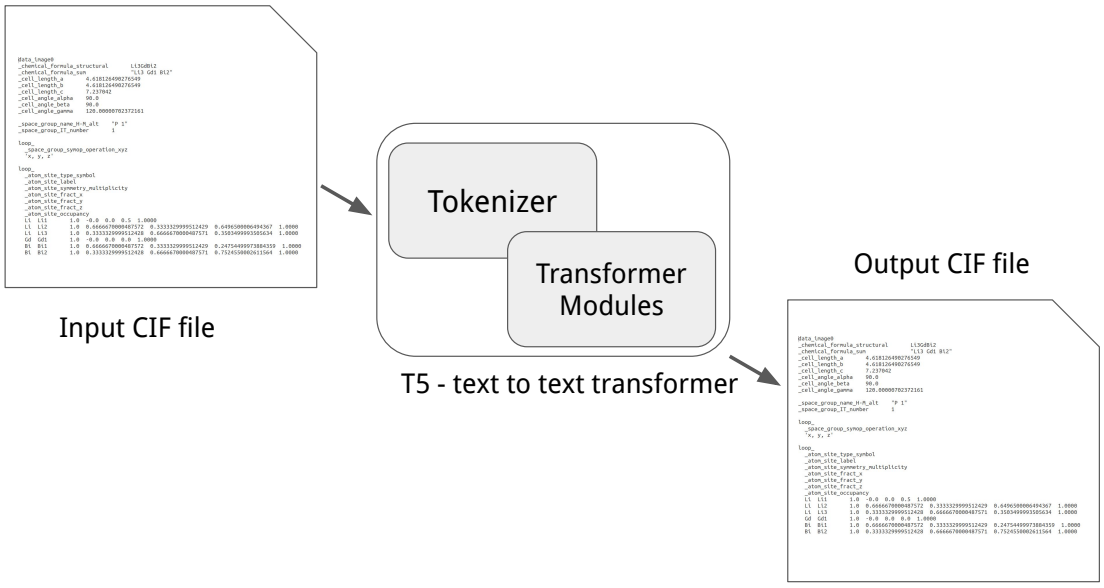

Figure 24: Overview of the relaxation training pipeline. The input structure from the LeMatTraj dataset is transformed to a CIF file using ASE. It then is fed into the T5 tokenizer and the fine-tuned transformer block layers to output a relaxed structure CIF file.

Metric	Value
MAE	0.066 Å
MSE	0.252 Å^2
RMSE	0.502 Å

Table 3: Errors in predicting the relaxed atom positions

Future Work

Future work should seek to use larger LLM models, use a larger combined dataset (e.g., data from NOMAD [107]), create a more systematic comparison with state-of-the art graph neural network methods used in materials science [108, 109] and look into how DFT parameters affect the calculated results [103] that are used as targets.

Open-source Materials

Code and tutorial available on GitHub: A demonstration video of running the model inference to get relaxed structures and a tutorial explaining the work can also be found on the GitHub.

Author Contributions

D.S.: conceptualization, code, writing. A.P.: conceptualization, feature engineering, code. A.K.: conceptualization, code, reviewing. H.Q.: conceptualization, code, reviewing. C.D.: writing, reviewing and editing. J.M.: supplying data, reviewing.

Acknowledgements

This work was supported by the NFDI consortium FAIRmat - Deutsche Forschungsgemeinschaft (DFG) - Project 460197019.