CLOUD: A Scalable Scientific Foundation Model for Crystal Representation Learning

Changwen Xu[1], **Shang Zhu**[1], **Venkatasubramanian Viswanathan**[* 1,2]
[1]Department of Mechanical Engineering, University of Michigan
[2]Department of Aerospace Engineering, University of Michigan
{changwex, shangzhu, venkvis}@umich.edu

Abstract

Developing machine learning models for crystal property predictions has been hampered by the need for labeled data from costly experiments or Density Functional Theory (DFT), resulting in limited data size and poor generalization to new crystals. Foundation models (FMs) present a potential solution with their self-supervised pre-training on unlabeled datasets and scalable model performance. Yet, applying FMs to crystals is challenging due to the inadequacy of existing string representations to capture critical structural information and the absence of scaling analysis for FMs specialized in materials science. Herein, We propose CrystaL fOUnDation model (CLOUD), a Transformer-based foundation model for crystal representation learning and property prediction. CLOUD utilizes a novel symmetry-aware string representation, eliminating the need for atomic coordinates or equivariant models. Pre-trained on million-scale crystal data, CLOUD is then fine-tuned and assessed on various downstream tasks, significantly outperforming other coordinate-free models on MatBench and MatBench Discovery. In addition, CLOUD achieves state-of-the-art (SOTA) or near-SOTA performance on UnconvBench for unconventional crystal property predictions. Furthermore, the pre-trained CLOUD demonstrates robust scaling with data and model size, which suggests CLOUD's potential as a scalable solution for crystal foundation models.

1 Introduction

The accurate and efficient property prediction is essential to the design of crystalline materials. Historically, materials development has been constrained to a limited portion of the vast chemical space of synthesizable crystalline materials due to the high costs associated with evaluating material properties within specific systems [Hellenbrandt, 2004, Davies et al., 2016]. Machine learning surrogates present a promising solution to significantly broaden the scope of materials performance evaluation across the entire chemical space of crystals [Phuthi et al., 2024b]. Pioneering work has significantly improved the prediction accuracy for crystals [Gilmer et al., 2017, Xie and Grossman, 2018, Goodall and Lee, 2020, Choudhary and DeCost, 2021, Ruff et al., 2023].

However, state-of-the-art (SOTA) crystal property prediction models rely on labeled training data obtained through expensive wet-lab experiments or Density Functional Theory (DFT) calculations. This reliance limits their utility due to the scarcity and variability of labeled materials datasets [Fujinuma et al., 2022]. Moreover, these models exhibit poor out-of-distribution (OOD) generalization, which is crucial for real-world applications where crystals with novel compositions or structures are common [Phuthi et al., 2024a]. Recently, foundation models (FMs) have emerged as a solution to this. The FMs usually take a transformer-based architecture [Vaswani et al., 2017] which enables the model to be effectively pre-trained at scale using unlabeled data which are orders of magnitude more than

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

labeled data [Devlin et al., 2018, Brown, 2020]. Benefiting from the implicit parallelism of attention, the transformer has enabled massive model architecture whose loss can scale with the number of training data and hyperparameters as a power law relationship [Kaplan et al., 2020, Hoffmann et al., 2022]. If this scaling law holds, the FM will achieve SOTA performance with more training data and model parameters. However, questions remain: what is the optimal string representation for crystals and how does the foundation model for crystals scale compared with general foundation models for language?

Herein, we propose CrystaL fOUnDation model (CLOUD), a transformer-based foundation model for accurate and generalizable crystal property predictions. The model consists of a BERT [Devlin et al., 2018] encoder and a multi-layer perceptron (MLP) prediction head. We design a symmetry-aware string representation for crystals that integrates symmetry, equivalent sites, and compositions for an efficient coordinate-free encoding of the crystal structures. Using this representation, CLOUD is pre-trained via masked language modeling (MLM) on $\sim$6.3M crystals from the OPTIMADE API [Evans et al., 2024], then fine-tuned on benchmark datasets to learn task-specific structure-property relationships. CLOUD outperforms other coordinate-free and structure-agnostic models on most MatBench tasks [Dunn et al., 2020], achieving SOTA results on two out of eight datasets. It also exhibits strong OOD performance on MatBench Discovery [Riebesell et al., 2023] and achieves SOTA or near-SOTA performance for unconventional crystals [Wang et al., 2024], highlighting its potential as the surrogate model in the loop of material discovery. Additionally, CLOUD follows a scaling law, suggesting potential performance gains with further scaling.

Our contributions are summarized as follows:

- We propose the symmetry-aware string representation for crystal structures, providing a novel approach to encode the symmetry of crystal structures into a sequence.
- We train the foundation model CLOUD which exhibits solid performance in crystal property predictions, illustrating the benefits of pre-training on representation learning.
- We evaluate the scaling performance of CLOUD, demonstrating the promising perspective of foundation models for material design and discovery.

2 Background

2.1 Crystal Representation Learning

Rational representations that map crystals to continuous vector space are crucial for crystal property prediction with machine learning models. One class of the models relies on the compositions of models, namely the specific combination and proportion of elements that make up the material. The models either build graphs based on the composition (Roost [Goodall and Lee, 2020], Finder [Ihalage and Hao, 2022]) or learn the latent embeddings via attention mechanism (CrabNet [Wang et al., 2021a]). Despite the simplicity and accessibility, composition-based, structure-agnostic models suffer from low expressiveness as one composition may take multiple structures (polymorphs). Therefore, another group of models uses atomic coordinates as input for representation learning. These structure-based models are typically built on graph neural networks (GNNs), which operate on graph structures by iteratively updating node embeddings through message passing along edges and then mapping the resulting feature vectors to properties [Gilmer et al., 2017, Schütt et al., 2018, Satorras et al., 2021]. The crystal graph convolutional neural network (CGCNN) was the first to apply GNNs to crystalline systems by constructing a multi-graph that accurately represents atomic neighbors in a periodic system [Xie and Grossman, 2018]. Progress has been made by integrating more geometric information, as seen in the Atomistic Line Graph Neural Network (ALIGNN), which operates on line graphs, treating edges in the original graph as nodes and learning higher-order geometric features such as bond angles and dihedrals [Choudhary and DeCost, 2021]. Despite all the success, however, systematic analysis points out that GNNs are poor at capturing long-range interactions. [Gong et al., 2023].

In contrast, a third type of models, regarded as coordinate-free models have emerged, using structural information to build crystal representations but do not require any knowledge of atomic coordinates. Therefore, models trained on coordinate-free representations learn more effective features than structure-agnostic models while eliminating the need for atomic positions which are usually

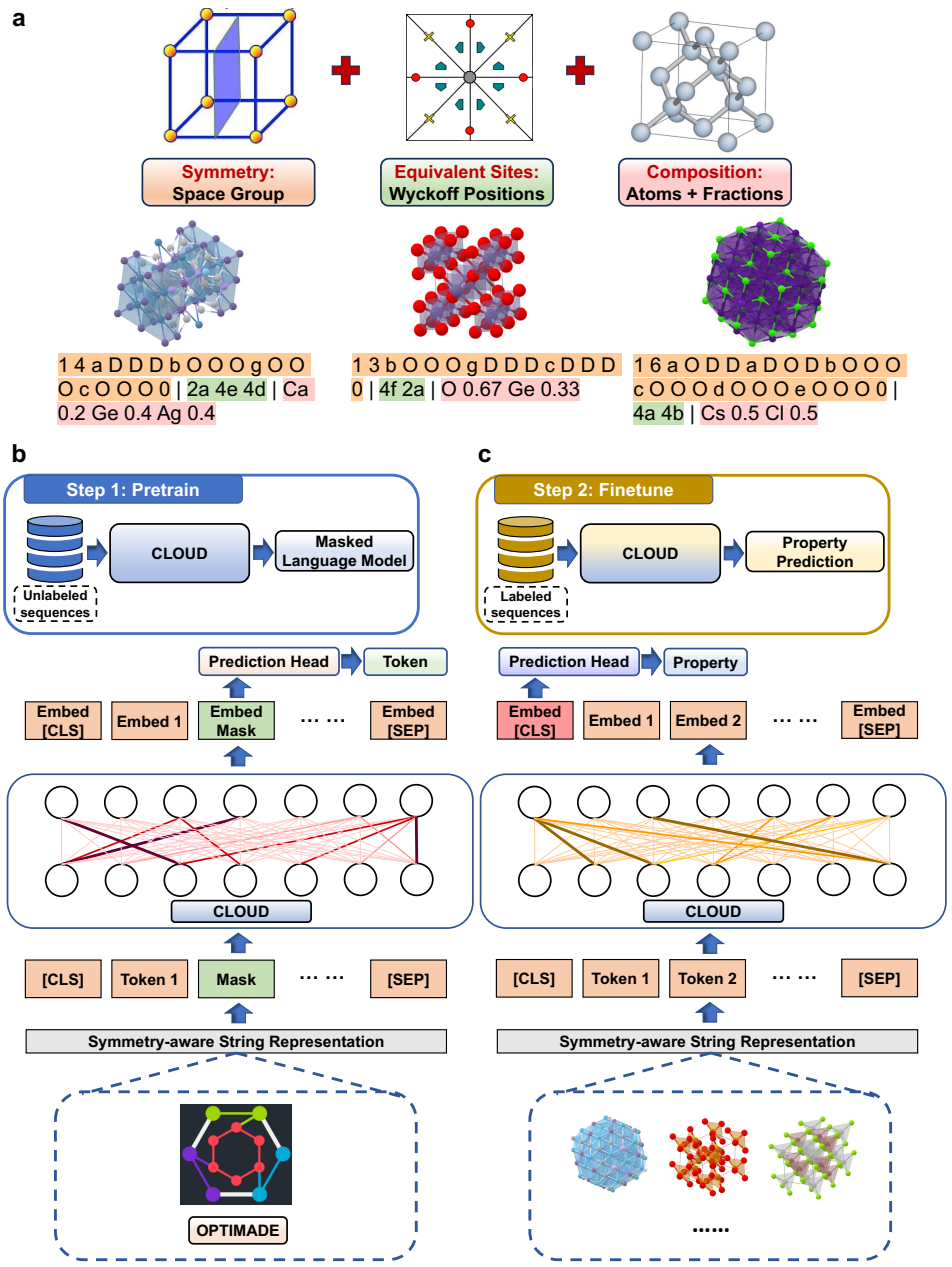

Figure 1: Overview of CLOUD. **a** Symmetry-aware string representation for CLOUD. The representation consists of space group generator strings, Wyckoff position symbols, and material composition. **b**-**c** Illustration of the pre-training and fine-tuning process. The model is pre-trained on million-scale OPTIMADE data with masked language modeling (MLM) to recover the original tokens, while the feature vector corresponding to the special token '[CLS]' of the last hidden layer is used for prediction when fine-tuning on diverse downstream tasks. Within the CLOUD block, lines of deeper color and larger width stand for higher attention scores.

computation-intensive to acquire, significantly reducing computation cost and expanding available data for training [Goodall et al., 2022]. The flexibility of coordinate-free representations has also enabled the design of string representations for crystals so that language models can be leveraged for representation learning. Wrenformer [Riebesell et al., 2023] replaces the message-passing layers in Roost with self-attention blocks and operates on Wyckoff embeddings proposed by Goodall et al. [2022]. Recently, a string representation for crystals called SLICES was proposed [Xiao et al.,

2023]. SLICES encode constituting elements as well as connectivity between atoms, allowing for the efficient representation of material structures as strings, whereas global structural information such as symmetry is not explicitly included. Huang et al. [2023] proposed the first sequence representation that includes space group to train Material Informatics Transformer (MatInFormer). Space group is a mathematical abstraction of the symmetry in the crystal structures, describing how the atomic arrangement in the crystal remains invariant under certain transformations. However, the usage of space group numbers and symbols does not fully convey the symmetry operators indicated by the space group and cannot distinguish the relationship between space groups. This is because space group names and numbers only provide a high-level classification without detailing the specific symmetry operators. Therefore, the optimal approach to encode symmetry in string representations is still unclear.

2.2 Scientific Foundation Model for materials

The scarcity of available labeled data has challenged the application of machine learning models in material representation learning due to the time-consuming evaluation process with wet-lab experiments [Shoemaker et al., 2014, Bugaris and zur Loye, 2012] or ab-initio thermodynamic calculations [Saal et al., 2013, Curtarolo et al., 2012]. Foundation models (FMs) offer a solution to this: large language models based on the transformer architecture benefit from implicit parallelism of attention, and its model performance can be characterized with empirical scaling laws [Kaplan et al., 2020]:

$$L(N, D) = [(\frac{N_c}{N})^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D}]^{\alpha_D} \tag{1}$$

where N and D are the number of tokens and non-embedding parameters in the model, and N_c, D_c, α_N, and α_D are fitting parameters. The equation demonstrates a power law relationship between the cross-entropy loss and the number of tokens, model parameters, and computing budget. Furthermore, Hoffmann et al. [2022] extend the scaling law by considering the intrinsic entropy of data:

$$L(N, D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E \tag{2}$$

FMs use self-supervised pre-training strategies to leverage unlabeled datasets and learn representations of data that can be applied to downstream tasks [Bommasani et al., 2021, Frey et al., 2023] to predict crystal properties [Merchant et al., 2023, Antunes et al., 2023, Gruver et al., 2024]. FMs have successfully been implemented in molecular property prediction: ChemBERTa [Chithrananda et al., 2020] and MoLFormer [Ross et al., 2022] are pioneering work in the field which show that the performance of molecular foundation models improves with increasing data set size on both regression and classification tasks. If this trend is extrapolated, model performance is expected to continue to improve when the pre-training data set size is increased. However, systematic scaling analysis for FMs for materials has not yet been conducted.

3 Method

We design a symmetry-aware string representation for crystal structures that integrates symmetry, equivalent sites, and constituting atoms to efficiently encode the structural and compositional information, eliminating the need for coordinate information or equivariant models. Based on the representation we design, we build up a transformer-based model which is first pre-trained on unlabeled crystal data and then fine-tuned on various downstream datasets.

3.1 Symmetry-aware String Representation

Exiting crystal representations are largely based on either composition or the full structure information which are not without limitations: composition only is not expressive enough given the non-unique composition-structure relationship, while structure-based graph representations require 3D coordinate information which is expensive to acquire. Furthermore, many of the existing representations miss the symmetry information which plays a pivotal role in determining the physical properties of crystals.

Therefore, we design a novel symmetry-aware string representation for coordinate-free representation of crystal structures. The representation can be divided into three parts:

Symmetry Different from the representation proposed by Huang et al. [2023], we use generator strings [De Graef and McHenry, 2012] composed of symbols that represent the basic symmetry operators, called symmetry generators. Given that many symmetry operators can be obtained by multiplication of other operators, we only need the most fundamental symmetry operators to construct the complete space group. According to De Graef and McHenry [2012], there are 14 symmetry generators which are labeled from a to n. Combining the upper-case letters assigned for translation components of symmetry operators, the generator strings form a compact representation of all the allowable basic symmetry of the crystal structure, hence making the generator strings *fingerprints* of symmetries. We provide more information on the generator strings in Appendix A.2.1.

Equivalent sites Wyckoff positions describe sites that map onto equivalent sites under the symmetry transformations of the given space group [Goodall et al., 2022]. Consequently, a single Wyckoff position can encode the positions of multiple atoms. The Wyckoff positions of a given crystal structure are deterministic if the space group is specified because all the equivalent sites can be derived with the symmetry defined for the structure. Wyckoff positions are denoted by a symbol containing a number and a letter, e.g., $4d$. The number denotes the number of equivalent sites, also known as the "multiplicity". The letter distinguishes the different Wyckoff positions by assigning "a" for positions with the lowest multiplicity and letters later in the alphabet to positions of higher multiplicity. We use the symbols of Wyckoff positions that are occupied by atoms in the crystal representation.

Composition We complete the representation by including the elements that make up the material along with their fractions in the material. As a result, the representation is informed of the material's composition, ensuring that the model captures both the identity and proportion of elements.

Combining the three parts above, we obtain the symmetry-aware string representation for crystal structures. Examples of representations can be found in Figure 1**a**. The representation forms a top-down description of the crystal structure – starting from symmetry which is a global information of the structure, followed by the equivalent sites under the symmetry and the stoichiometry of the crystal – without explicitly encoding the structure itself, accomplishing efficient encoding of crystals.

3.2 Training Strategy

We tokenize the string representation so that each chemically meaningful unit is treated as a single token. Specifically, each symbol in the generator string is tokenized individually, each Wyckoff position symbol is treated as a single token, and both element symbols and fractional values are represented as distinct tokens. All the float numbers are rounded to 2 decimal places.

CLOUD takes up the BERT architecture [Devlin et al., 2018] and the training strategy follows the pretrain-finetune paradigm. We first pre-train the model via Masked Language Modeling (MLM) in which the model recovers the masked tokens based on the contexts. To pre-train the model, only the crystal representation itself is required. The self-supervised pre-training paradigm leverages the abundant unlabeled data for effective representation learning. The pre-trained CLOUD is then fine-tuned on various downstream datasets to learn the structure-property relationship with task-specific information. We provide additional information on model implementation in Appendix A.2.3.

4 Experiments

We empirically benchmark CLOUD on various tasks for evaluation, mainly in terms of mean absolute error (MAE) for regression tasks and ROC-AUC for classification tasks. General experimental setups are provided in 4.1, and more information about the implementations can be found in Appendix A.2.3.

4.1 Experimental Setup

Dataset We use the Open Databases Integration for Materials Design (OPTIMADE) [Evans et al., 2024] dataset for pre-training CLOUD. OPTIMADE comprehensively collects crystal data from

Table 1: Results on MatBench regression tasks. All results are shown in the mean of validation MAE in five-fold cross-validation along with the standard deviation in brackets. The best and the second best results among structure-agnostic and coordinate-free models are in bold and underlined, and the best results among all the models are in italics.

Model	jdft2d (↓)	phonons (↓)	dielectric (↓)	gvrh (↓)	kvrh (↓)	perovskites (↓)	band gap (↓)	e form (↓)
coGN	37.1652 (13.6825)	29.7117 (1.9968)	0.3088 (0.0859)	0.0689 (0.0009)	0.0535 (0.0028)	*0.0269 (0.0008)*	*0.1559 (0.0017)*	*0.0170 (0.0003)*
coNGN	36.1698 (11.5973)	*28.8874 (3.2840)*	0.3142 (0.0740)	*0.0670 (0.0006)*	*0.0491 (0.0026)*	0.0290 (0.0011)	0.1697 (0.0035)	0.0178 (0.0004)
ALIGNN	43.4244 (8.9491)	29.5385 (2.1148)	0.3449 (0.0871)	0.0715 (0.0006)	0.0568 (0.0028)	0.0288 (0.0009)	0.1861 (0.0030)	0.0215 (0.0005)
CGCNN	49.2440 (11.5865)	57.7635 (12.3109)	0.5988 (0.0833)	0.0895 (0.0016)	0.0712 (0.0028)	0.0452 (0.0007)	0.2972 (0.0035)	0.0337 (0.0006)
CrabNet	45.6104 (12.2491)	55.1114 (5.7317)	0.3234 (0.0714)	0.1014 (0.0017)	0.0758 (0.0034)	0.4065 (0.0069)	0.2655 (0.0029)	0.0862 (0.0010)
Finder	47.9614 (11.6681)	46.5751 (3.7415)	0.3204 (0.0811)	0.0996 (0.0018)	0.0764 (0.0025)	0.6450 (0.0167)	0.2308 (0.0029)	0.0839 (0.0011)
Roost	44.6405 (11.7353)	54.3893 (4.7283)	0.3252 (0.0780)	0.1034 (0.0020)	0.0797 (0.0042)	0.4025 (0.0077)	0.2571 (0.0055)	0.0847 (0.0016)
Wrenformer	39.6309	92.3349	0.3583	0.1070	0.0811	0.3351	0.2986	0.0694
MatInFormer	45.1309 (12.7339)	44.0134 (5.7265)	0.3046 (0.0765)	0.0873 (0.0019)	0.0703 (0.0030)	0.4339 (0.0105)	0.2541 (0.0020)	0.0646 (0.0003)
SLICES-BERT	37.8586 (10.9928)	44.5470 (4.0192)	0.3417 (0.0918)	0.0932 (0.0015)	0.0698 (0.0021)	**0.0351 (0.0013)**	0.2776 (0.0043)	0.0543 (0.0004)
CLOUD	***35.0057 (11.4550)***	**40.6856 (2.6168)**	***0.3038 (0.0782)***	**0.0873 (0.0028)**	**0.0682 (0.0026)**	0.0969 (0.0010)	**0.2126 (0.0030)**	**0.0542 (0.0007)**

various databases and provides an API for users with easy access to download the data. We download the de-duplicate the initial ∼13M CIF files from OPTIMADE, resulting in ∼6.3M data for pre-training. For fine-tuning the model, we conduct experiments on three benchmarks: MatBench [Dunn et al., 2020], MatBench Discovery [Riebesell et al., 2023], and UnconvBench [Wang et al., 2024] to evaluate CLOUD as well as other baseline models. We use the eight regression datasets out of 13 from MatBench to benchmark the model's capability in predicting a variety of DFT-calculated solid materials' properties. The data split for each task in our experiments is consistent with MatBench. Meanwhile, we benchmark CLOUD on MatBench Discovery which emphasizes designing machine learning models that could be well-generalized to unseen data for the discovery of new stable materials. In addition, to further align with the requirements of material discovery, we test CLOUD's predictive capability for unconventional crystals such as defected crystals, two-dimensional crystals, and crystals of large systems. We follow the data split defined by UnconvBench. More information about the datasets is provided in Appendix A.2.2.

Baselines We compare CLOUD to diverse leading models which are summarized in Appendix A.2.4. The baseline models are primarily divided into three groups: structure-based models which utilize coordinates of atoms as input, structure-agnostic models which do not require structural information, and coordinate-free models which rely on the structures to build the input while do not require the inclusion of atomic coordinates. To ensure a fair comparison, we limit the prediction head for the three pre-trained language models (MatInFormer [Huang et al., 2023], SLICES-BERT (we train on SLICES [Xiao et al., 2023] with the same architecture as CLOUD), and CLOUD) to be a single-layer MLP for tasks from MatBench to ensure consistency with MatInFormer architecture. All the baseline results except for SLICES-BERT are obtained from benchmark leaderboards.

4.2 MatBench

We compare CLOUD with baseline models that show up on the MatBench leaderboard in terms of mean absolute error (MAE). As shown in Table 1, CLOUD surpasses all the other structure-agnostic and coordinate-free models on 7 out of 8 tasks on MatBench. Furthermore, CLOUD achieves state-of-the-art results on jdft2d and dielectric, if excluding MODNet [De Breuck et al., 2021] as it is a feature-based model trained in a multi-task learning strategy. In contrast to structure-based models which tend to struggle on data-limited regimes, the competitive performance of CLOUD on small datasets like jdft2d and dielectric suggests the efficacy of pre-training on a large corpus of unlabeled data for enhanced performance on downstream tasks. Furthermore, the superior performance of CLOUD compared to MatInFormer suggests that encoding symmetry using generator strings enhances the model's predictive accuracy on downstream tasks. Compared to SLICES which encodes bond connectivity while neglecting overall symmetry, CLOUD reaches lower prediction error with much more compact representation and consequently less computing budgets. SLICES-BERT only surpasses CLOUD on perovskites on which none of the structure-agnostic or coordinate-free models perform well while SLICES benefits from connectivity information which conveys more subtle differences between perovskite structures.

Figure 2 presents the comparison between results from CLOUD with/without pre-training. Model performance on downstream tasks is significantly enhanced if the model is pre-trained compared to training from scratch.

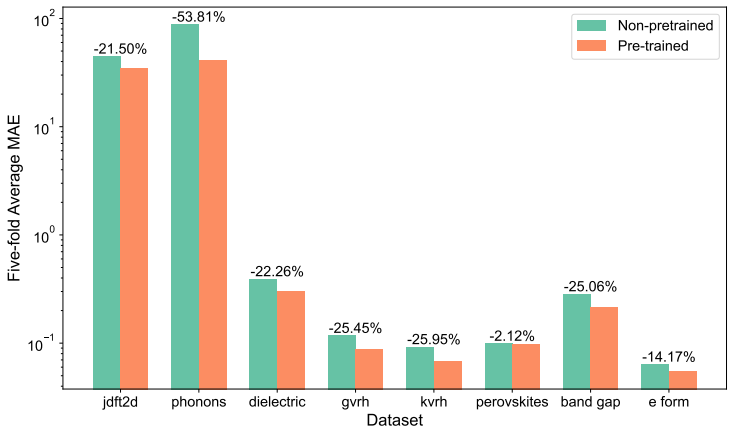

Figure 2: Comparison of five-fold average MAE results on MatBench datasets between CLOUD trained from scratch and pre-trained with MLM. The MAE results are shown in base 10 logarithm. The percentage of reduction in MAE is annotated above each dataset.

4.3 MatBench Discovery

We train the model on Materials Project [Jain et al., 2013] formation energy data from the v2022.10.28 MP release, then make predictions for the unrelated structures in the WBM dataset [Wang et al., 2021b] which are generated via elemental substitution of MP source structures so that the generated structures are not included in the training set. Based on the predictions for formation energies, the energy above the convex hull (E_{hull}), the distance to the convex hull spanned by competing phases in the same chemical system, is readily obtained to determine the stability of the materials. A material will be classified as stable if its E_{hull} is lower than a threshold. The stability threshold for E_{hull}, typically set to 0, should be dynamically adjusted. A fixed threshold fails to account for the systematic prediction shifts unique to each model, making it suboptimal in many cases. Therefore, we record the classification results under varied thresholds and plot the receiver operating characteristic (ROC) curve for CLOUD and the major structure-based models that are benchmarked on MatBench, shown in Figure 3. Area under curve (AUC) scores are calculated consequently and labeled in the figure. In contrast to the results on the MatBench leaderboard on which CLOUD gives a larger prediction error on formation energy than structure-based GNNs, The fine-tuned CLOUD achieves the AUC score of 0.81 which is close to that of ALIGNN and higher than that of CGCNN and MEGNet, suggesting that CLOUD is able to achieve better out-of-distribution performance. We present the ROC plot for more baseline models in Figure 5 in the Appendix. We further examine the other metrics listed in Table 6 in the Appendix. The regression metrics of CLOUD are again second to ALIGNN while outperforming CGCNN and MEGNet. The classification metrics for CLOUD are close to those of Wrenformer under the stability threshold of 0, as shown in Table 6, whereas the results in Table 7 suggests that CLOUD may benefit from a more negative threshold, which is also observed for models that are more optimistic in stability predictions like CHGNet [Riebesell et al., 2023].

4.4 UnconvBench

Table 2 summarizes the model performance in terms of MAE on UnconvBench to evaluate the predictive performance for 'unconventional' crystals, which are in fact the prevailing crystals in the real world. Different from the cases for MatBench in which the sophisticated structure-based GNN models like coGN, coNGN, and ALIGNN dominate the leaderboard, CLOUD demonstrates comparable performance to the SOTA model CrysToGraph and significantly outperforms the other structure-based models. Such evidence reveals the promising potential of CLOUD to be applied for material discovery. Noticeably, the best model on each task of UnconvBench is either CrysToGraph or CLOUD, both of which take account of global information of crystal structures. Different from CrysToGraph which learns the long-range interactions with a graph-wise transformer, CLOUD directly encodes symmetry information via generator strings and Wyckoff positions, providing an alternative for encoding global information.

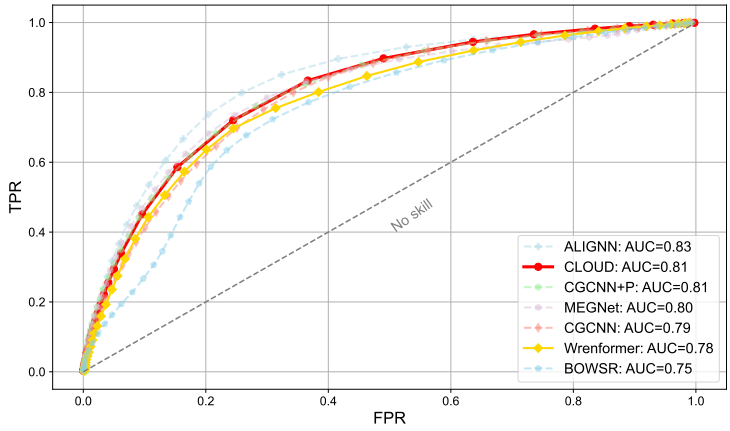

Figure 3: Receiver operating characteristic (ROC) curves for models evaluated on MatBench Discovery. False positive rate (FPR) on the x-axis is the fraction of unstable structures classified as stable. True positive rate (TPR) on the y-axis is the fraction of stable structures classified as stable. The two coordinate-free models CLOUD and Wrenformer are plotted in solid lines while the structure-based models are drawn in dashed lines. Area under curve (AUC) scores are calculated and presented in descending order.

Table 2: Results on UnconvBench general predictive tasks. All results are shown in the mean of validation MAE in five-fold cross-validation. The best results are in bold and the second-best results are underlined.

Type	Model	2d_e_exf (↓)	2d_e_tot (↓)	2d_gap (↓)	qmof (↓)	supercon (↓)	defected (↓)
Structure-based	CrysToGraph	**0.0500**	**0.3623**	0.0986	**121.9662**	**2.6422**	0.8885
	coGN	0.0510	0.5214	0.1168	218.9272	2.8955	1.0615
	coNGN	0.0530	0.4497	0.1432	229.1948	2.9167	1.0441
	ALIGNN	0.0580	0.3705	0.1048	217.2508	2.7372	0.9842
	CGCNN	0.0710	1.2941	0.1499	231.1887	2.9316	1.1321
Coordinate-free	CLOUD	0.0569	0.3723	**0.0919**	159.5723	2.6925	**0.8197**

4.5 Scaling Analysis

The scaling performance of CLOUD is evaluated. We vary the number of pre-training data and the number of non-embedding parameters in the BERT encoder of the model and investigate how the model performance changes. Figure 4**a**-**b** illustrates that the pre-training (MLM) loss scales well with data and model size. We evaluate the scaling-up of CLOUD on the downstream task on one of the datasets from MatBench about the shear modulus of materials (Figure 4**c**-**d**). Besides the promising scaling-up performance on the downstream task, we also discover that the model trained from scratch, though giving a similar performance as the pre-trained one at a small model size, does not scale up as fast as the pre-trained one.

We fit the scaling law for CLOUD following Hoffmann et al. [2022] which takes the form of Equation 2. Specifically, we minimize the following objective function:

$$\min_{A,B,E,\alpha,\beta} \sum_i Huber_\delta(\log \hat{L}(N_i, D_i) - \log L_i) \tag{3}$$

We solve the optimization problem with L-BFGS algorithm [Nocedal, 1980] to obtain the following result:

$$A = 90.7464,\ B = 21.9854,\ E = 0.2296,\ \alpha = 0.4339,\ \beta = 0.3518$$

With the scaling law, we estimate the optimal model size for a given compute budget. Following the derivation in Hoffmann et al. [2022], the optimal model size and data size can be written as:

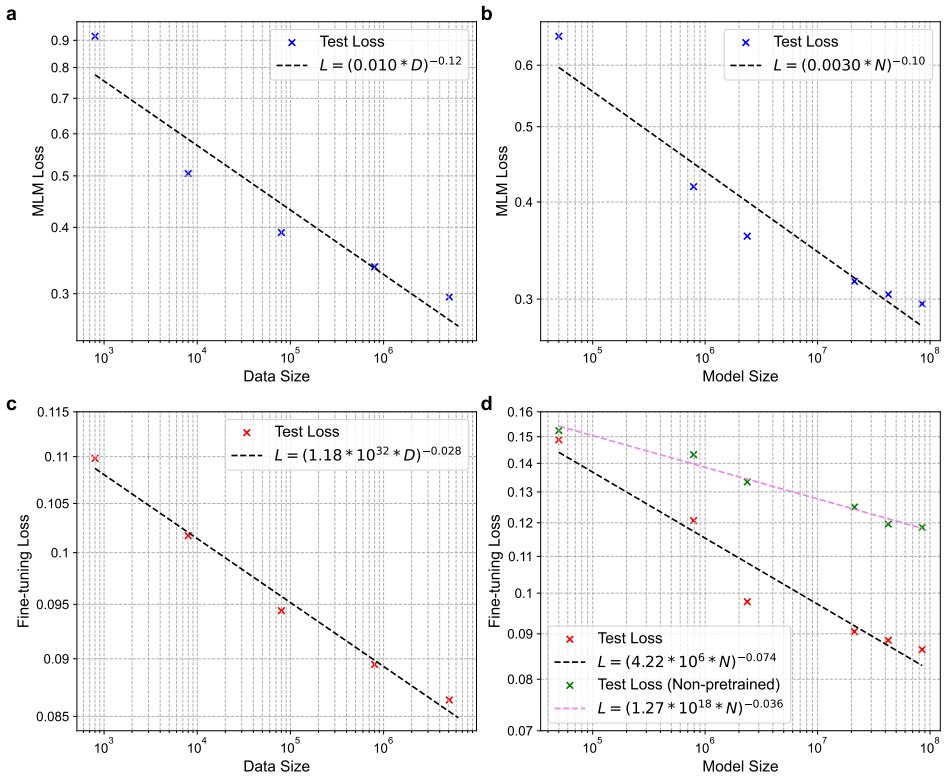

Figure 4: Scaling of model performance in the pre-training task with respect to **a** data size and **b** model size, and model performance in the fine-tuning task **gvrh** from MatBench with respect to **c** data size and **d** model size. Data size refers to the number of data points used for pre-training and model size refers to the number of parameters in the transformer encoder of CLOUD.

$$N_{opt}(C) = G(\frac{C}{6})^{a},\ D_{opt}(C) = G^{-1}(\frac{C}{6})^{b} \tag{4}$$

where

$$G = (\frac{\alpha A}{\beta B})^{\frac{1}{\alpha+\beta}},\ a = \frac{\beta}{\alpha+\beta},\ b = \frac{\alpha}{\alpha+\beta} \tag{5}$$

Given the α and β we fitted, we can readily derive that $a = 0.45$ and $b = 0.55$ for CLOUD. Interestingly, the a and b from the empirical scaling law that we fit for CLOUD are close to the values in Hoffmann's scaling law where $a = b = 0.50$ [Hoffmann et al., 2022]. Given $a \approx b$, the number of parameters and number of tokens should be scaled evenly, unlike the case for Kaplan's scaling law which infers that the model size should scale faster than the number of tokens ($a = 0.73$, $b = 0.27$) [Kaplan et al., 2020].

5 Conclusion

In this paper, we present our CrystaL fOUnDation model (CLOUD) for accurate and generalizable prediction for crystal properties. CLOUD, taking the symmetry-aware string representation as input for encoding structural information to sequences, is pre-trained on the million-scale unlabeled crystal dataset and then fine-tuned on various downstream tasks. CLOUD demonstrates competitive performance on multiple material property prediction benchmarks and manifests robust scaling performance with respect to data and model size. We anticipate that CLOUD will serve as a scalable crystal foundation model that forms a universal solution to materials science tasks with more unlabeled crystal structures and computation resources available.

References

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *arXiv preprint arXiv:2307.04340*, 2023.

Rickard Armiento. Database-driven high-throughput calculations and machine learning models for materials design. *Machine Learning Meets Quantum Physics*, pages 377–395, 2020.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C Lawrence Zitnick, and Zachary W Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436, 2022.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL `https://arxiv.org/abs/2108.07258`.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Daniel E Bugaris and Hans-Conrad zur Loye. Materials discovery by flux crystal growth: quaternary and higher order oxides. *Angewandte chemie international edition*, 51(16):3780–3811, 2012.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885, 2020. URL `https://arxiv.org/abs/2010.09885`.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

Stefano Curtarolo, Wahyu Setyawan, Gus LW Hart, Michal Jahnatek, Roman V Chepulskii, Richard H Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, et al. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.

Daniel W Davies, Keith T Butler, Adam J Jackson, Andrew Morris, Jarvist M Frost, Jonathan M Skelton, and Aron Walsh. Computational screening of all stoichiometric inorganic materials. *Chem*, 1(4):617–627, 2016.

Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rignanese. Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet. *npj computational materials*, 7(1):83, 2021.

Marc De Graef and Michael E McHenry. *Structure of materials: an introduction to crystallography, diffraction and symmetry*. Cambridge University Press, 2012.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Claudia Draxl and Matthias Scheffler. Nomad: The fair concept for big data-driven materials science. *Mrs Bulletin*, 43(9):676–682, 2018.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

Marco Esters, Corey Oses, Simon Divilov, Hagen Eckert, Rico Friedrich, David Hicks, Michael J Mehl, Frisco Rose, Andriy Smolyanyuk, Arrigo Calzolari, et al. aflow. org: A web ecosystem of databases, software and tools. *Computational Materials Science*, 216:111808, 2023.

Matthew L Evans and Andrew J Morris. matador: a python library for analysing, curating and performing high-throughput density-functional theory calculations. *Journal of Open Source Software*, 5(54):2563, 2020.

Matthew L Evans, Johan Bergsma, Andrius Merkys, Casper W Andersen, Oskar B Andersson, Daniel Beltrán, Evgeny Blokhin, Tara M Boland, Rubén Castañeda Balderas, Kamal Choudhary, et al. Developments and applications of the optimade api for materials discovery, design, and data exchange. *Digital Discovery*, 3(8):1509–1533, 2024.

Nathan C Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gomez-Bombarelli, Connor W Coley, and Vijay Gadepally. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, 2023.

Luis E Fuentes-Cobas, Daniel Chateigner, María E Fuentes-Montero, Giancarlo Pepponi, and Saulius Grazulis. The representation of coupling interactions in the material properties open database (mpod). *Advances in Applied Ceramics*, 116(8):428–433, 2017.

Naohiro Fujinuma, Brian DeCost, Jason Hattrick-Simpers, and Samuel E. Lofland. Why big data and compute are not necessarily the path to big materials science. *Communications Materials*, 3(1): 1–9, August 2022. ISSN 2662-4443. doi: 10.1038/s43246-022-00283-x.

Jason Gibson, Ajinkya Hire, and Richard G Hennig. Data-augmentation for graph neural network learning of the relaxed energies of unrelaxed structures. *npj Computational Materials*, 8(1):211, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Morten Niklas Gjerding, Alireza Taghizadeh, Asbjørn Rasmussen, Sajid Ali, Fabian Bertoldo, Thorsten Deilmann, Nikolaj Rørbæk Knøsgaard, Mads Kruse, Ask Hjorth Larsen, Simone Manti, et al. Recent progress of the computational 2d materials database (c2db). *2D Materials*, 8(4): 044002, 2021.

Sheng Gong, Keqiang Yan, Tian Xie, Yang Shao-Horn, Rafael Gomez-Bombarelli, Shuiwang Ji, and Jeffrey C Grossman. Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. *Science Advances*, 9(45):eadi3245, 2023.

Rhys EA Goodall and Alpha A Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature communications*, 11(1):6280, 2020.

Rhys EA Goodall, Abhijith S Parackal, Felix A Faber, Rickard Armiento, and Alpha A Lee. Rapid discovery of stable materials by coordinate-free coarse graining. *Science Advances*, 8(30):eabn4117, 2022.

Saulius Gražulis, Adriana Daškevič, Andrius Merkys, Daniel Chateigner, Luca Lutterotti, Miguel Quiros, Nadezhda R Serebryanaya, Peter Moeck, Robert T Downs, and Armel Le Bail. Crystallography open database (cod): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic acids research*, 40(D1):D420–D427, 2012.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.

Mariette Hellenbrandt. The inorganic crystal structure database (icsd)—present and future. *Crystallography Reviews*, 10(1):17–22, 2004.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hongshuo Huang, Rishikesh Magar, Changwen Xu, and Amir Barati Farimani. Materials informatics transformer: A language model for interpretable materials properties prediction. *arXiv preprint arXiv:2308.16259*, 2023.

Achintha Ihalage and Yang Hao. Formula graph self-attention network for representation-domain independent materials discovery. *Advanced Science*, 9(18):2200164, 2022.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Adam M Krajewski, Jonathan W Siegel, and Zi-Kui Liu. Efficient structure-informed featurization and property prediction of ordered, dilute, and random atomic structures. *arXiv preprint arXiv:2404.02849*, 2024.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, pages 1–6, 2023.

Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 2024.

Mgcini Keith Phuthi, Archie Mingze Yao, Simon Batzner, Albert Musaelian, Pinwen Guan, Boris Kozinsky, Ekin Dogus Cubuk, and Venkatasubramanian Viswanathan. Accurate surface and finite-temperature bulk properties of lithium metal at large scales using machine learning interaction potentials. *ACS Omega*, 9(9):10904–10912, 2024a. doi: 10.1021/acsomega.3c10014. URL `https://doi.org/10.1021/acsomega.3c10014`.

Mgcini Keith Phuthi, Archie Mingze Yao, Simon Batzner, Albert Musaelian, Pinwen Guan, Boris Kozinsky, Ekin Dogus Cubuk, and Venkatasubramanian Viswanathan. Accurate surface and finite-temperature bulk properties of lithium metal at large scales using machine learning interaction potentials. *ACS Omega*, 2024b.

Janosh Riebesell, Rhys EA Goodall, Anubhav Jain, Philipp Benner, Kristin A Persson, and Alpha A Lee. Matbench discovery–an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*, 2023.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Robin Ruff, Patrick Reiser, Jan Stühmer, and Pascal Friederich. Connectivity optimized nested graph networks for crystal structures. *arXiv preprint arXiv:2302.14102*, 2023.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65:1501–1509, 2013.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel AL Marques. Crystal graph attention networks for the prediction of stable materials. *Science advances*, 7(49): eabi7948, 2021.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

Daniel P Shoemaker, Yung-Jin Hu, Duck Young Chung, Gregory J Halder, Peter J Chupas, L Soderholm, JF Mitchell, and Mercouri G Kanatzidis. In situ studies of a platform for metastable inorganic crystal growth and materials discovery. *Proceedings of the National Academy of Sciences*, 111 (30):10922–10927, 2014.

Leopold Talirz, Snehal Kumbhar, Elsa Passaro, Aliaksandr V Yakutovich, Valeria Granata, Fernando Gargiulo, Marco Borelli, Martin Uhrin, Sebastiaan P Huber, Spyros Zoupanos, et al. Materials cloud, a platform for open computational science. *Scientific data*, 7(1):299, 2020.

MPDS Team. Pauling file: Inorganic materials database. `https://paulingfile.com/`, 2024. Accessed: 2024-09-04.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials*, 7(1):77, 2021a.

Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1):12, 2021b.

Hongyi Wang, Ji Sun, Jinzhe Liang, Li Zhai, Zitian Tang, Zijian Li, Wei Zhai, Xusheng Wang, Weihao Gao, Sheng Gong, et al. Crystals with transformers on graphs, for prediction of unconventional crystal material properties and the benchmark. *arXiv preprint arXiv:2407.16131*, 2024.

Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Physical Review B*, 96(2):024104, 2017.

Hang Xiao, Rong Li, Xiaoyang Shi, Yan Chen, Liangliang Zhu, Xi Chen, and Lei Wang. An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nature Communications*, 14(1):7027, 2023.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

Yunxing Zuo, Mingde Qin, Chi Chen, Weike Ye, Xiangguo Li, Jian Luo, and Shyue Ping Ong. Accelerating materials discovery with bayesian optimization and graph deep learning. *Materials Today*, 51:126–135, 2021.

A Appendix

A.1 Limitations and Future Work

While CLOUD shows competitive performance in crystal representation learning and property prediction, it does encounter limitations and potential improvements for future explorations. One limitation is that our model requires a large amount of pre-trained data to achieve SOTA performance in many of the downstream tasks according to the scaling law we fitted, whereas such a dataset has not yet been built. Based on the scaling analysis we performed, CLOUD will get to the top of the leaderboard on gvrh from MatBench if the data size reaches $\sim$10 billion, whereas the largest material database, which is the OPTIMADE dataset we use for pre-training, only includes million-scale unique data points which only covers a small portion of the design space of crystalline materials. Recently, a lot of machine learning models have been proposed for crystal generation [Merchant et al., 2023, Zeni et al., 2023, Antunes et al., 2023, Gruver et al., 2024], however, none of them have pushed the data size to the order of billions. Recently, OMat24 dataset [Barroso-Luque et al., 2024] was released which contains over 100 million DFT calculations for solid-state materials, which is a huge contribution to the research community to advance materials science development with AI. In future work, we plan to comprehensively integrate existing datasets like OMat24 and diverse generation approaches for crystal structure generation, ranging from rule-based methods to data-driven approaches.

Another limitation lies in the representation itself: the representation we design is built on symmetry, whereas symmetry does not capture everything. For instance, while the space group, Wyckoff positions, and stoichiometry remain unchanged during relaxation, the atomic positions and cell parameters undergo significant variations. As a result, CLOUD is unable to leverage the MPtrj dataset, which comprises over 1.5 million crystal structures from relaxation trajectories, unlike many other models on the leaderboard [Deng et al., 2023, Riebesell et al., 2023]. In addition, the representation does not work for amorphous materials which lack the long-range order and symmetry that crystalline materials exhibit, thus limiting the versatility of CLOUD in handling a broader range of materials. Furthermore, some Wyckoff positions include free variables, meaning the representation encodes an ensemble of materials. As a result, it may not distinguish between materials with different physico-chemical properties that share the same prototype but have different atomic positions. In future work, we plan to systematically investigate the design of string representations for crystals which enables effective encoding of more information besides symmetry and composition.

A.2 Implementation Details

A.2.1 Symmetry-aware String Representation

In crystallography, *space groups* are fundamental in describing the symmetrical properties of crystal structures. They encompass both the translational and point symmetries that define how a motif repeats in three-dimensional space. To systematically represent these symmetries, we use *generator strings*, which provide a concise symbolic notation for the generators of a space group [De Graef and McHenry, 2012].

A symmetry operation is a transformation that leaves the crystal structure invariant. These operations include:

- **Rotations** (n-fold rotation axes)
- **Reflections** (mirror planes)
- **Inversions** (inversion centers)
- **Rotoinversions** (combination of rotation and inversion)
- **Translations** (including screw axes and glide planes)

A *generator* is a fundamental symmetry operation from which all other operations of the space group can be derived through combination. By selecting a minimal set of generators, we can fully describe the symmetry content of a space group. Therefore, each generator is represented with a symbolic notation to form the generator string for space groups. There are 14 generators for all the symmetry operations besides translational symmetry, and these matrices are represented by letters ranging from

a to n. Combining the generator matrices with the translation components which are represented with 10 upper-case letters, we can obtain all the symmetry operators. With the generators, it is possible to compile all 230 space groups in a short ASCII file that is only 4104 bytes long. More information can be found in Section 10.3.6 in De Graef and McHenry [2012].

Take the No. 35 space group as an example. Its generator string is given by '03aDDDbOOOjOOO0'. The '0' at the beginning of the generator string indicates the inversion operator is not a generator. After that, '3' indicates that there are three generator matrices. Every four letters correspond to a generator: the first lower-case letter determines which of the 14 matrices is to be used. and the subsequent 3 letters stand for the translation components in three dimensions. For instance, 'aDDO' corresponds to the symmetry operator:

$$\begin{pmatrix} 1 & 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The last symbol in the generator string is '0' which indicates that there is not any alternative choice of the origin for this space group.

We argue that generator strings are more informative than space group symbols, as space group symbols only provide a high-level classification of symmetry, whereas generator strings offer detailed insight into the specific symmetry operations that define the crystal's structure. For example, $Cmc2_1$ (No. 36) and $Amm2$ (No. 38) only have one character in common in the space group symbols. However, their generator strings are '03aDDObOODjOOD0' and '03aODDbOOOjOOO0', respectively, suggesting that their symmetry generators are quite similar and only differ in the translation components. In fact, In fact, both space groups belong to the orthorhombic crystal system and exhibit mirror planes and two-fold rotation axes, but the subtle difference in the translation components (specifically in glide operations) reflects the underlying structural variation. This distinction is not apparent from the space group symbols alone, highlighting the value of generator strings in capturing more nuanced symmetry information. In contrast, $Fmm2$ (No. 42) has a face-centered lattice while $Imm2$ (No. 44) has a body-centered lattice, and the difference in lattice type significantly affects the symmetry operations and translations within the unit cell even though they both belong to Orthorhombic crystal system. Their space group names have 75% of the characters in common. Instead, their generator strings, '04aODDaDODbOOOjOOO0' and '03aDDDbOOOjOOO0', reflect more difference in the symmetry operators than their names.

A.2.2 Datasets

Open Databases Integration for Materials Design (OPTIMADE) [Evans et al., 2024] dataset is used to pre-train CLOUD. OPTIMADE integrates data from various major materials databases and provides an application programming interface (API) to make the data accessible to users. The contributing databases include: AFLOW [Esters et al., 2023], Alexandria [Schmidt et al., 2021], Computational Two-Dimensional Materials Database (C2DB) [Gjerding et al., 2021], Crystallography Open Database (COD) [Gražulis et al., 2012], Joint Automated Repository for Various Integrated Simulations (JARVIS) [Choudhary et al., 2020], Materials Cloud [Talirz et al., 2020], Materials Platform for Data Science (MPDS) [Team, 2024], Materials Project [Jain et al., 2013], Material-Property-Descriptor Database (MPDD) [Krajewski et al., 2024], Material Properties Open Database (MPOD) [Fuentes-Cobas et al., 2017], NOMAD [Draxl and Scheffler, 2018], Open Database of Xtals (odbx) [Evans and Morris, 2020], Open Materials Database (omdb) [Armiento, 2020], and Open Quantum Materials Database (OQMD) [Saal et al., 2013]. Therefore, such comprehensive datasets enable effective pre-training on crystal representations. The original data we downloaded using the OPTIMADE API consists of $\sim$13M CIF files. We de-duplicate the data by only keeping the structure with the smallest volume per volume when several CIF files have exactly the same chemical formula and space group. Finally, we use the de-duplicated $\sim$6.3M data as the pre-training set.

MatBench [Dunn et al., 2020] is a benchmark test suite that contains 13 tasks using data from density functional theory-derived and experimental sources. We use 8 out of 13 tasks for benchmarking as they are regression tasks and provide the crystal structures to build the symmetry-aware string representation. The information about these 8 tasks is summarized in Table 3.

Table 3: Summary of the 8 regression tasks from MatBench and the 6 general predictive tasks from UnconvBench used in this paper.

Benchmark	Dataset	Size	Property	Unit
MatBench	jdft2d	636	Exfoliation energy of 2D materials	meV/atom
	phonons	1265	Highest frequency optical phonon mode peak	cm^{-1}
	dielectric	4764	Refractive index	unitless
	gvrh	10987	Logarithm of VRH average shear moduli	$\log GPa$
	kvrh	10987	Logarithm of VRH average bulk moduli	$\log GPa$
	perovskites	18928	Formation energy of perovskites	eV/unit cell
	band gap	106113	Band gap	eV
	e form	132752	Formation energy	eV/atom
UnconvBench	2d_e_exf	4527	Exfoliation energy of 2D materials	eV/atom
	2d_e_tot	3520	Total formation energy of 2D crystals	eV
	2d_gap	3520	Band gap of 2D materials	eV
	qmof	5106	Formation energy of MOFs	eV
	supercon	1058	Curie temperature	K
	defected	530	Formation energy of defects	eV/atom

MatBench Discovery [Riebesell et al., 2023] is a comprehensive resource designed to evaluate machine learning models for materials discovery, particularly for predicting the thermodynamic stability of materials. The dataset aims to reflect practical challenges in the discovery process by requiring predictions based on unrelaxed crystal structures, which avoid reliance on expensive DFT calculations. The models are trained on customized training data and tested on WBM dataset [Wang et al., 2021b] which contains $\sim$257K OOD crystal structures generated by systematically substituting elements in pre-existing structures from the Materials Project (MP). This dataset is heavily composed of ternary phases, with a significant fraction of transition metals and metalloids, offering a broader chemical diversity than the MP dataset. The WBM is designed to challenge models to perform OOD predictions, making it a demanding benchmark for stability.

UnconvBench [Wang et al., 2024] consists of unconventional crystal structures including 2D crystals, metal-organic frameworks (MOFs), defected crystals. The benchmark contains various crystals with irregular and complex long-range, whereas such crystalline systems are rarely observed in highly ordered traditional crystals. Hence, UnconvBench provides complementary insights to existing benchmark datasets, offering a broader perspective on evaluating model performance on unconventional materials (which are in fact conventional in the real world). We include the information about the 6 general predictive tasks from UnconvBench which are used in this work for benchmarking in Table 3.

A.2.3 Training Details

The hyperparameters used for training CLOUD are summarized in Table 4. A list of values is shown in the table if the hyperparameter takes different values for different tasks. The model architecture is fixed to 12 hidden layers and 12 attention heads in each layer for the BERT part in CLOUD when used for benchmarking, while we experiment with smaller models in order to fit the scaling law. The number of hidden layers is fixed to 1 for experiments on MatBench for a fair comparison, while we experiment with up to 3 hidden layers in the prediction head on MatBench Discovery and UnconvBench for optimal performance as we do not need to compare with MatInFormer or SLICES-BERT on those two benchmarks. The hyperparameters that are used in experiments for pre-training and fine-tuning are also listed in Table 4.

A.2.4 Baseline

We summarize the baseline models used for benchmarking in Table 5. The baselines cover a variety of models:

- Voronoi RF directly uses chemical descriptors as features for input.
- Roost, Finder, and CrabNet use composition only as input.
- coGN, coNGN, ALIGNN, MEGNet, CGCNN, CGCNN+P, and BOWSR build graphs from crystal structures and learn the mapping from the structure to target properties.

Table 4: Hyperparameters of CLOUD.

Hyperparameter	Value
# of hidden layers in BERT	{1,3,6,12}
# of attention heads in BERT	{1,4,12}
# of hidden layers in MLP	{1,2,3}
max position embeddings	64
block size	64
embedding size	768
# of hidden layers in MLP	{1,2,3}
hidden layer width in MLP	768
activation function in MLP	SiLU
pre-train epochs	50
pre-train learning rate	1e-4
pre-train batch size	2048
pre-train weight decay	0.0
pre-train optimizer	AdamW
pre-train scheduler	linear warmup and cosine decay
pre-train warm-up ratio	0.05
fine-tune epochs	{50,100,200}
fine-tune learning rate	{1e-4,5e-5}
fine-tune batch size	{32,64,128,512}
fine-tune weight decay	{0.0, 0.0001, 0.01}
fine-tune optimizer	AdamW
fine-tune scheduler	linear warmup and cosine decay
fine-tune warm-up ratio	{0.05,0.1}
fine-tune # of freezing-encoder epochs	{0,5,10}

- SevenNet, MACE, CHGNet, and M3GNet are GNNs trained to serve as universal machine learning-based interatomic potentials (MLIPs).
- CrysToGraph is a transformer-based geographic model for learning both local and long-range information from graphs.
- MatInFormer, Wrenformer, and SLICES-BERT are built on a transformer architecture and take sequences as input, same as our model CLOUD.

A.2.5 Evaluation Metrics

Our evaluation strategies follow the ones adopted by the benchmark leaderboards. We use MAE as the metric for regression tasks in MatBench and UnconvBench. MatBench Discovery leaderboard provides results for multiple metrics: MAE, RMSE, and R^2 for energy above convex hull prediction, and F1, accuracy, precision, true positive rate (TPR), and true negative rate (TNR) for stability classification. In particular, Riebesell et al. [2023] propose discovery acceleration factors (DAF) quantifies how much faster a machine learning model can identify stable materials compared to random selection. The expression of DAF is given as follows:

$$DAF = \frac{Precision}{N_{stable,true}/N_{total}} \tag{6}$$

where $N_{stable,true}$ and N_{total} are the number of stable materials and the total number of materials in the dataset, respectively.

We also provide the results for ROC-AUC in order to remove the sensitivity of the classification metrics to the choice of thresholds.

Table 5: Summary of baseline models used in this work.

Model	Type	Architecture	Source
coGN	Structure-based	GNN	Ruff et al. [2023]
coNGN	Structure-based	GNN	Ruff et al. [2023]
ALIGNN	Structure-based	GNN	Choudhary and DeCost [2021]
CGCNN	Structure-based	GNN	Xie and Grossman [2018]
CrabNet	Structure-agnostic	Transformer	Wang et al. [2021a]
Finder	Structure-agnostic	GNN	Ihalage and Hao [2022]
Roost	Structure-agnostic	GNN	Goodall and Lee [2020]
Wrenformer	Coordinate-free	Transformer	Riebesell et al. [2023]
MatInFormer	Coordinate-free	Transformer	Huang et al. [2023]
SLICES-BERT	Coordinate-free	Transformer	Xiao et al. [2023]
SevenNet	Structure-based (MLIP)	GNN	Park et al. [2024]
MACE	Structure-based (MLIP)	GNN	Batatia et al. [2022]
CHGNet	Structure-based (MLIP)	GNN	Deng et al. [2023]
M3GNet	Structure-based (MLIP)	GNN	Chen and Ong [2022]
MEGNet	Structure-based	GNN	Chen et al. [2019]
CGCNN+P	Structure-based	GNN	Gibson et al. [2022]
BOWSR	Structure-based	GNN	Zuo et al. [2021]
Voronoi RF	Feature-based	Random Forest	Ward et al. [2017]
CrysToGraph	Structure-based	Graph Transformer	Wang et al. [2024]

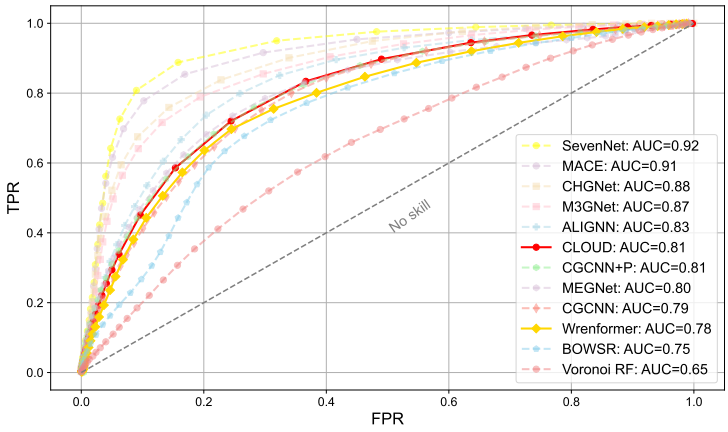

Figure 5: Receiver operating characteristic (ROC) curves for each model evaluated on MatBench Discovery. False positive rate (FPR) on the x-axis is the fraction of unstable structures classified as stable. True positive rate (TPR) on the y-axis is the fraction of stable structures classified as stable. Area under curve (AUC) scores are calculated and presented in descending order. More models are included in the plot compared to Figure 3.

A.3 Additional Results

A.3.1 MatBench Discovery

We plot the ROC curves with more models included in Figure 5 compared to Figure 3. Though not comparable to machine-learned interatomic potentials (MLIPs) at this stage, CLOUD still shows solid performance compared with CGCNN, MEGNet, Wrenformer, etc.

We provide the classification and regression metrics on MatBench Discovery in Table 6. The stability threshold is set to 0 for classification metrics. CLOUD exhibits better regression performance than CGCNN while its classification falls behind Wrenformer under this threshold.

We list the classification results by CLOUD under different stability thresholds in Table 7. Note that the true labels for the test data are derived with the threshold of 0, consistent with the benchmark

Table 6: Classification and regression metrics for models tested on MatBench Discovery ranked by F1 score. The stability threshold for model predictions are set to 0.

Model	F1 (↑)	DAF (↑)	Prec (↑)	Acc (↑)	TPR (↑)	TNR (↑)	MAE (↓)	RMSE (↓)	R^2 (↓)
SevenNet	0.719	3.804	0.653	0.893	0.800	0.912	0.046	0.090	0.750
MACE	0.668	3.400	0.583	0.867	0.781	0.885	0.055	0.099	0.698
CHGNet	0.612	3.038	0.521	0.839	0.740	0.859	0.061	0.100	0.690
M3GNet	0.576	2.647	0.454	0.802	0.788	0.804	0.072	0.115	0.588
ALIGNN	0.565	2.921	0.501	0.829	0.649	0.866	0.092	0.154	0.274
MEGNet	0.513	2.699	0.463	0.813	0.574	0.862	0.128	0.204	-0.277
CGCNN	0.510	2.631	0.451	0.807	0.587	0.852	0.135	0.229	-0.624
CGCNN+P	0.510	2.398	0.411	0.779	0.670	0.801	0.108	0.178	0.027
Wrenformer	0.479	2.130	0.365	0.741	0.693	0.751	0.105	0.182	-0.020
CLOUD	0.474	2.043	0.340	0.711	0.781	0.697	0.104	0.167	0.147
BOWSR	0.437	1.836	0.315	0.702	0.711	0.680	0.114	0.164	0.142
Voronoi RF	0.344	1.509	0.259	0.665	0.511	0.697	0.141	0.206	-0.316
Dummy	0.194	1	0.168	0.680	0.231	0.770	0.120	0.181	0

Table 7: Classification metrics for CLOUD tested on MatBench Discovery under varying stability thresholds.

Threshold	F1 (↑)	DAF (↑)	Prec (↑)	Acc (↑)	TPR (↑)	TNR (↑)
-0.05	0.467	2.903	0.484	0.828	0.451	0.904
0.0	0.474	2.043	0.340	0.711	0.781	0.697
0.05	0.380	1.434	0.239	0.494	0.932	0.407

setting [Riebesell et al., 2023], while the dynamic threshold applies to the model prediction. More negative thresholds will result in higher precision for CLOUD and subsequently higher DAF. However, the trade-off across metrics leads to decreased F1 score and TPR when a negative threshold is used.

A.3.2 Scaling Law

We plotted the fitted scaling law of CLOUD in Figure 6. The plot reflects the cross-entropy loss under the model size and the compute budget in terms of floating point operations per second (FLOPs). The compute optimal frontier is plotted by connecting the points with the lowest compute budget on each iso-loss contour.

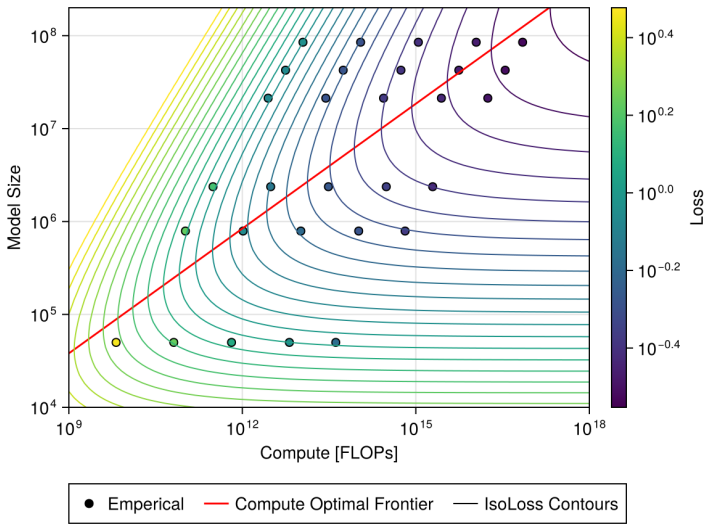

Figure 6: Fitted scaling law for CLOUD. The efficient frontier is shown in red. The line goes through each iso-loss contour at the point with the fewest FLOPs.