Fine Tuned Language Models Generate Stable Inorganic Materials as Text

Ahmet Barış Şimşek

Computer Vision Group

November 26, 2024

Structure of a Crystal

- Element identities
- Latice lengths
- Latice angles
- Translation symmetrie

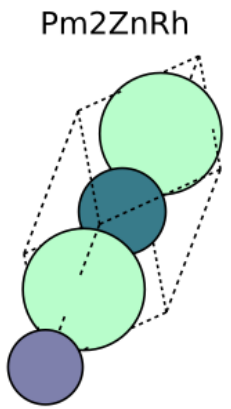


Figure: Crystal Unit Cell

Table of Contents

Crystal Diffusion Variational AutoEncoder

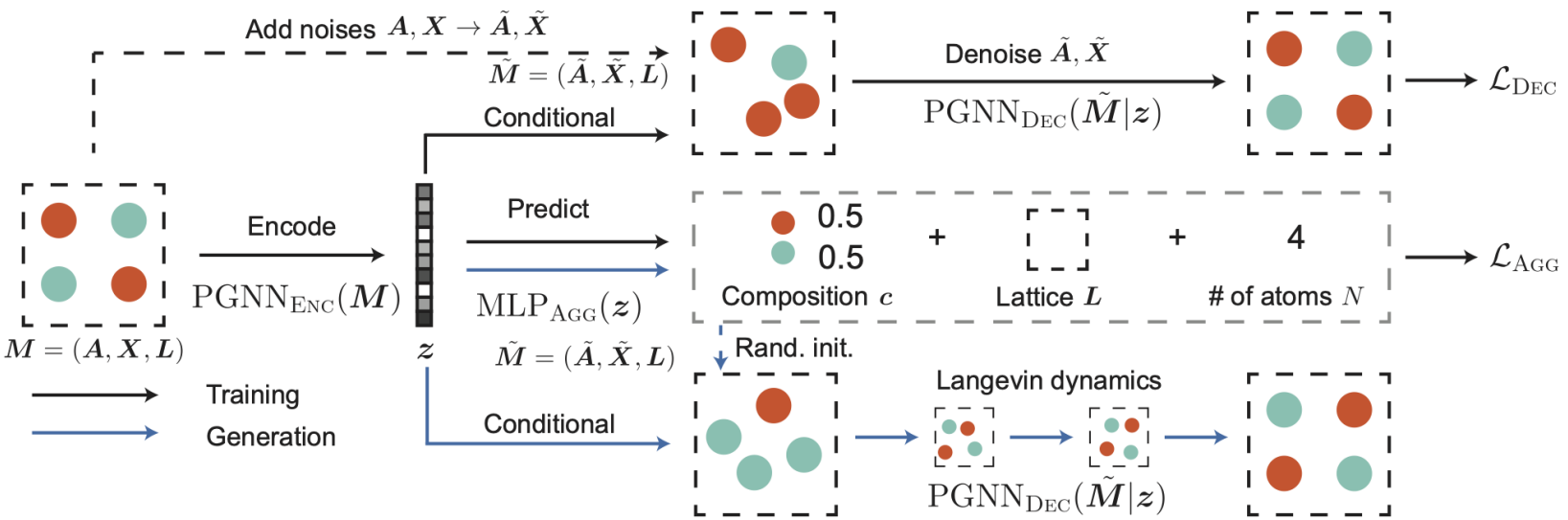


Figure: Overview of the CDVAE

arXiv:2110.06197

Crystal Diffusion Variational AutoEncoder

- Element identities are constructed with VAE
- Atom positions in latice are generated with diffusion

Other Diffusion Aproaches

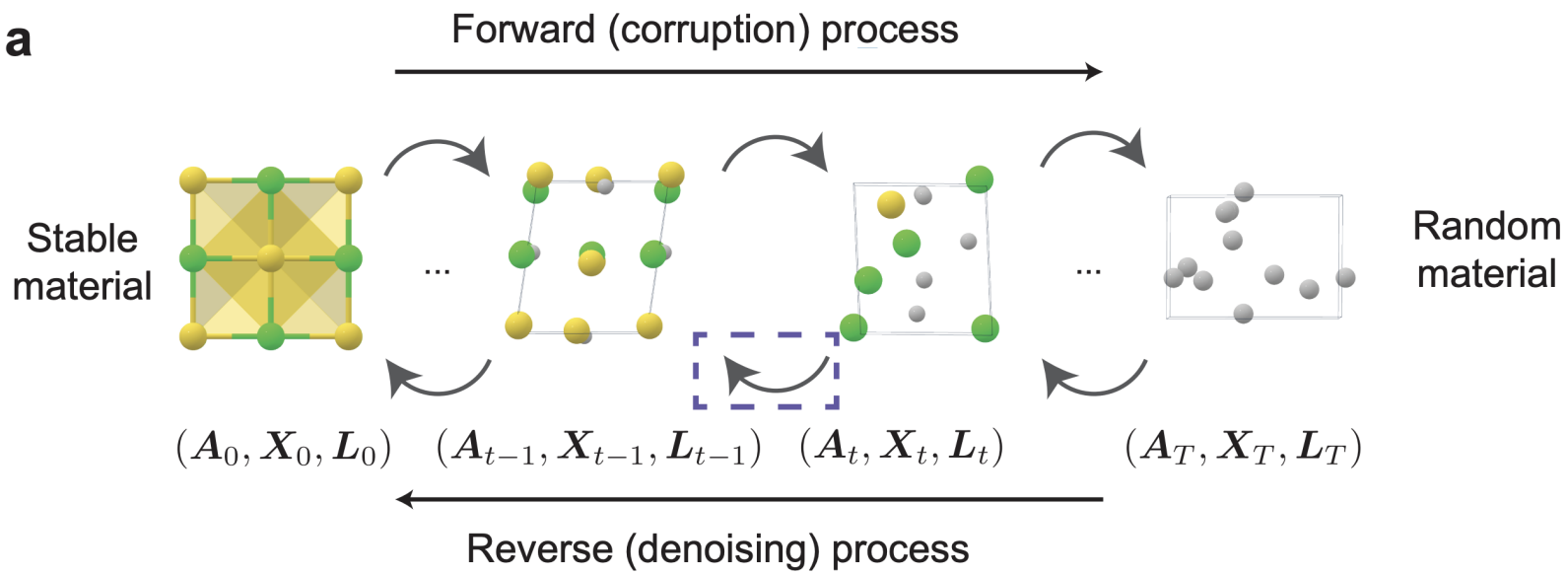


Figure: Overview of MatterGen

arXiv:2312.03687

Can't fine-tuned LLMs be used for generation?

Table of Contents

Transformers

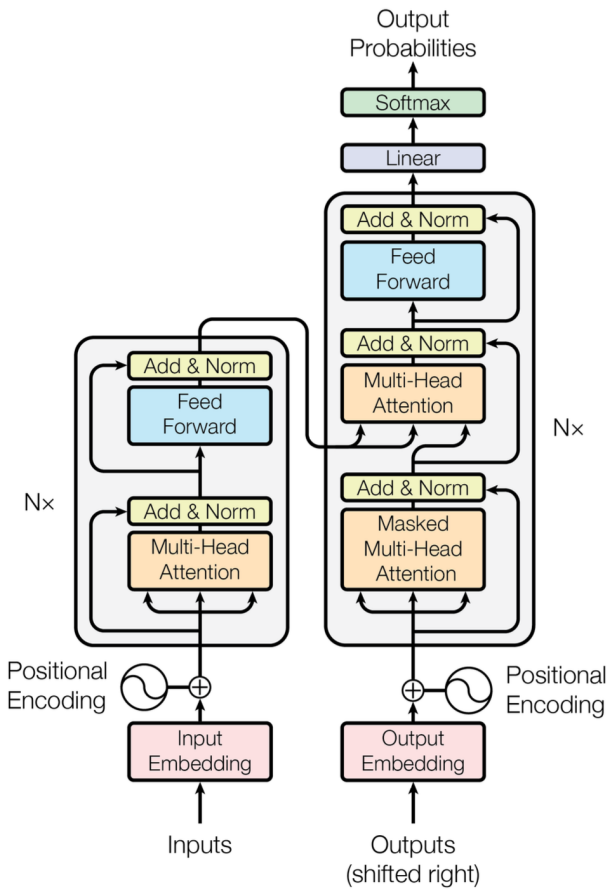


Figure: Architecture of a Transformer

arXiv:1706.03762

Attention

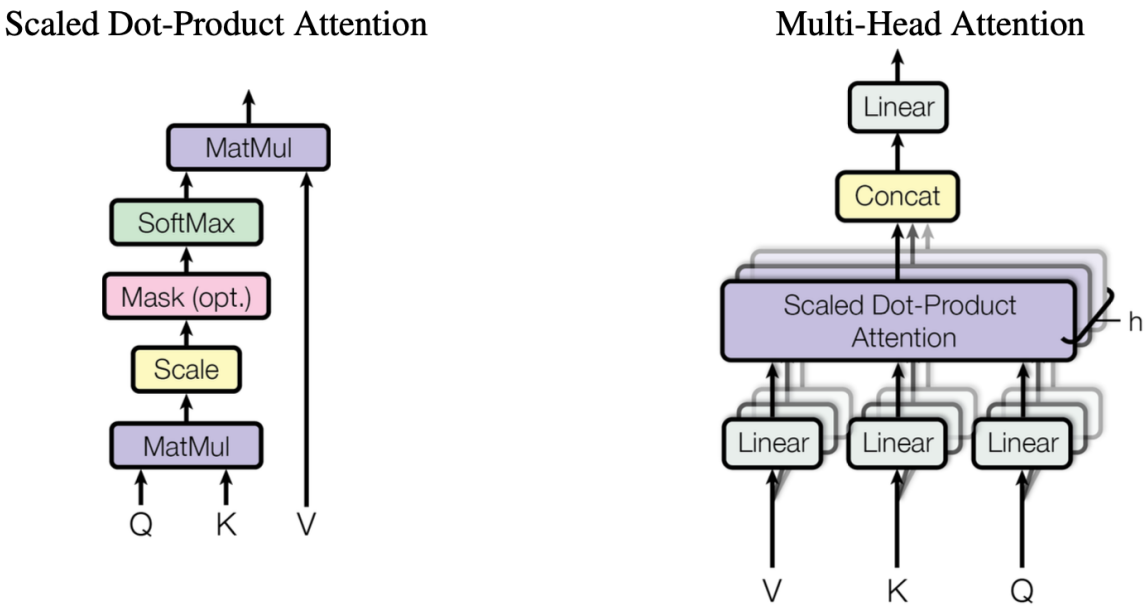


Figure: Attention mechanism

arXiv:1706.03762

Attention

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{\text{QK}^\top}{\sqrt{\text{d}_\text{k}}}\right)\text{V} \tag{1}$$

$$\text{head}_\text{i} = \text{Attention}(\text{QW}_\text{i}^\text{Q}, \text{KW}_\text{i}^\text{K}, \text{VW}_\text{i}^\text{V}) \tag{2}$$

$$\text{MultiHead}(\text{Q}, \text{K}, \text{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})\text{W}^\text{O} \tag{3}$$

Transformers

- Differs from predecessor by ditching recurrence and convolutions
- Uses only attention mechanism

Lama 2

- Built on top of Llama
- Decoder-only optimized transformer architecture for auto-regressive generation
- Models with 7B, 13B and 70B parameters
- Context lenght of 4k
- Tuned versions with SFT and RLHF

Llama 2

	Transformer	Llama
Attention	MHA	GQA
Encodings	Positional Encodings	RoPE
Normalization	LayerNorm	RMSNorm

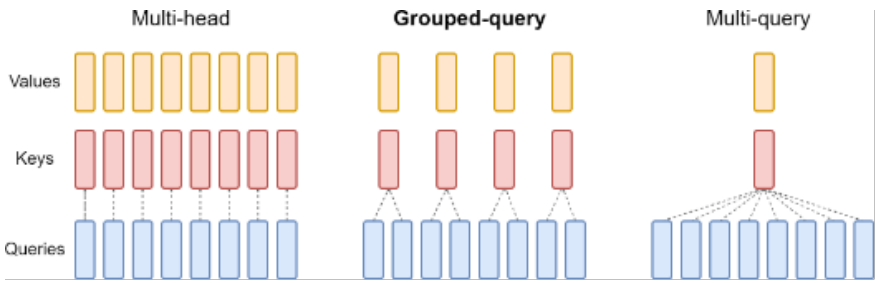


Figure: Grouped-query attention in comparison to multi-head and multi-query attention

arXiv:2305.13245

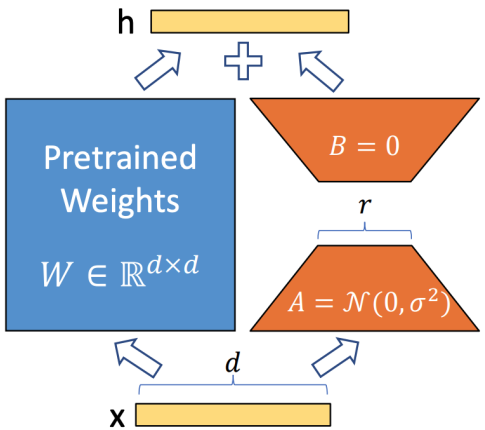
Fine Tuning

- Pre-trained LLMs are sample efficient
- Full fine-tuning vs PEFT
- Low-Rank Adaptation
- Prefix Tuning

LoRA

LoRA is a PEFT approach utilizing adapter matrices while freezing model parameters. r can be adjusted for required expressiveness of the training data.

- Only A and B trained
- Memory efficient
- Very minor inference overhead



arXiv:2106.09685

Table of Contents

Approach

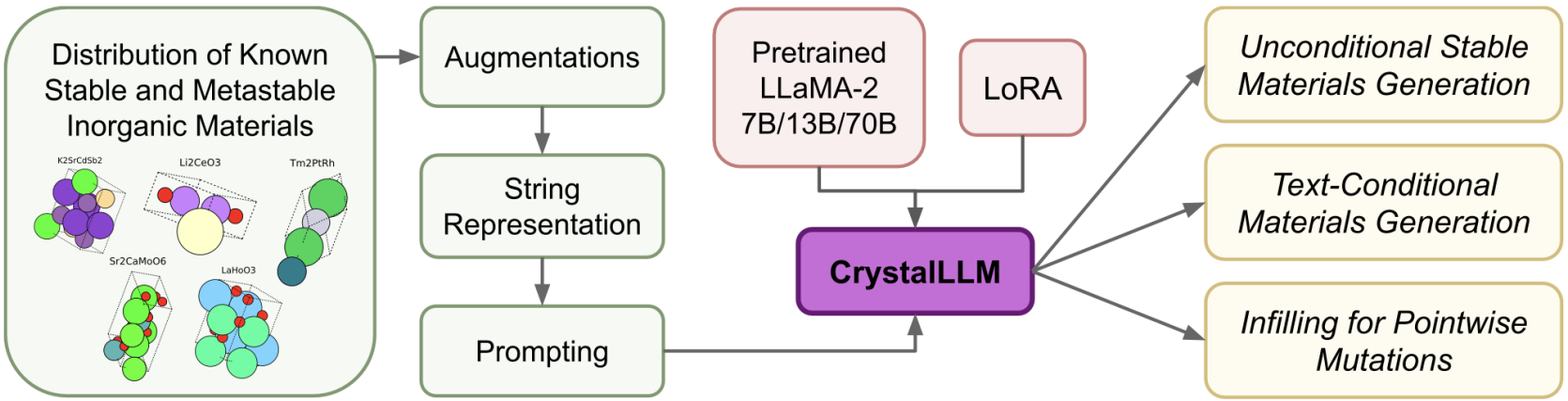


Figure: Proposed architecture

String Representations

An unit cell of crystal can be represented by lattice lengths, angles, element idendities and coordinates

$$\mathrm{C} = (\mathrm{l}_1, \mathrm{l}_2, \mathrm{l}_3, \theta_1, \theta_2, \theta_3, \mathrm{e}_1, \mathrm{x}_1, \mathrm{y}_1, \mathrm{z}_1, \ldots, \mathrm{e}_\mathrm{N}, \mathrm{x}_\mathrm{N}, \mathrm{y}_\mathrm{N}, \mathrm{z}_\mathrm{N}) \quad (4)$$

String Representations

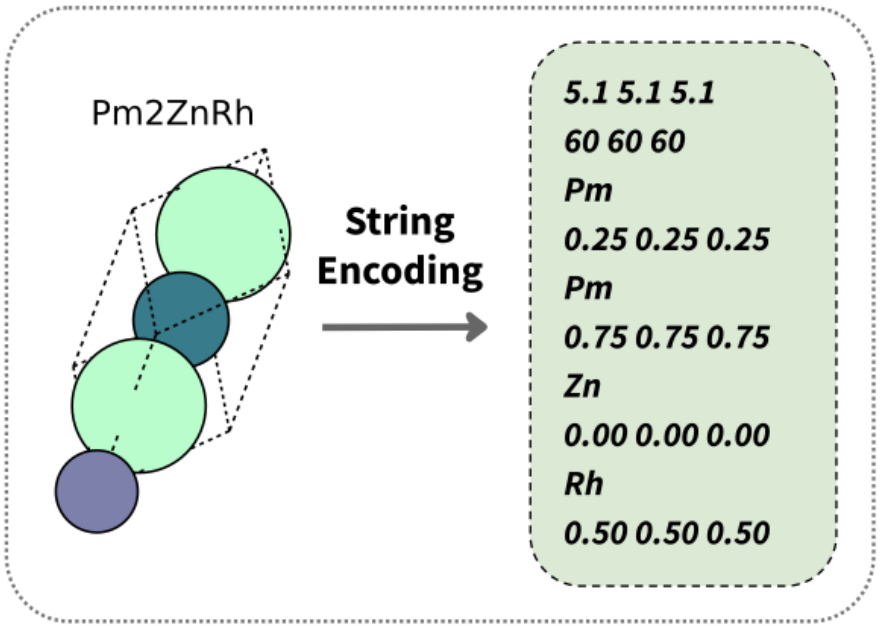


Figure: String encoding of a crystal

Tokenization

Most LLMs use byte pair encodings (BPE). Although useful for compressing singular tokens to more common substrings, it complicates uderstanding numerical values. Luckily Llama 2 breaks numeric values into singular digits.

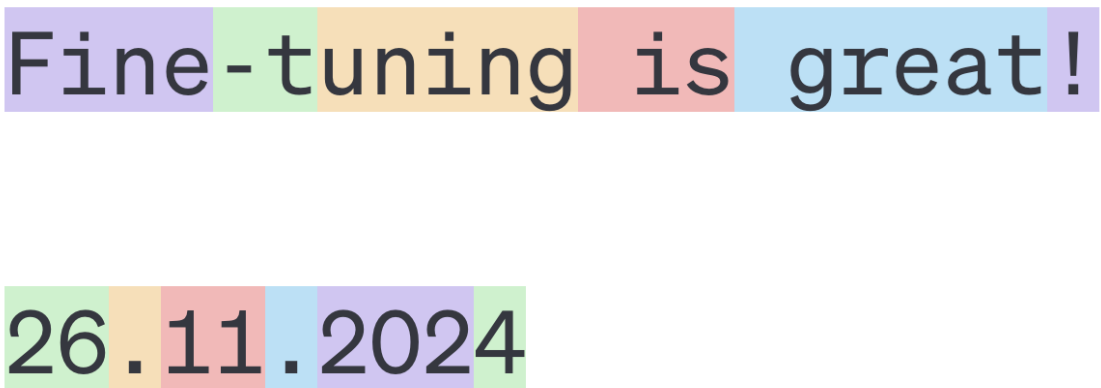


Figure: Tokenization of GPT-4o without single digit tokenization

Prompting

Generation Prompt	Infill Prompt
<s>Below is a description of a bulk material. [The chemical formula is Pm2ZnRh]. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice: [Crystal string]</s>	<s>Below is a partial description of a bulk material where one element has been replaced with the string "[MASK]": [Crystal string with [MASK]s] Generate an element that could replace [MASK] in the bulk material: [Masked element]</s>

Blue text is optional and included to enable conditional generation. Purple text stands in for string encodings of atoms.

Figure: Example prompt templates for text-conditioned generation and infilling

Augmentation

- Unit cells are translation invariant
- Ordering of elements is irrelevant (permutation invariant)
- Random uniform translations are applied to express symetries during fine-tuning (modulo latice boundaries)

Why would LLMs work for such task?

Why would LLMs work for such task?

- LLMs are great compressors
- LLMs are sample efficient
- LLMs have useful biases towards generalizable patterns
- Especially Llama 2 has a great bias towards 3D coordinates

Table of Contents

Energy Prediction

Each crystal configuration has an energy state that defines the likelihood of that configuration for the given environmental conditions. Unlikely atom positions will cause high energy predictions for a configuration.

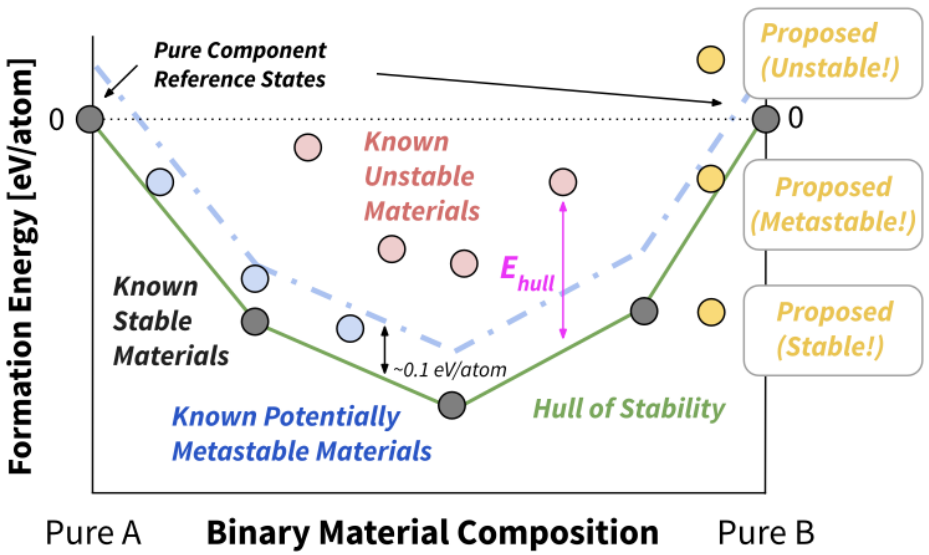
Energy Prediction

Golden standard for energy estimation: density functional theory (DFT). But it has its own downsides

- DFT is slow
- DFT is computationally expensive
- ML based approximation solutions are developed

Energy Prediction

Energy above hull describes the stability of a crystal. Every configuration has different energy. If energy above hull of a crystal smaller than zero, crystal is stable. In case its smaller than 0.1 eV/atom it's metastable instead.



Datasets

MP-20 dataset

- 45231 materials
- All stable configurations

For text-conditioned generation fine-tuned with 120,000 crystal configurations.

Metrics

- Compositional validity: Net charge of the configuration must be neutral for configuration to be valid
- Structural validity: No atomic radii can overlap
- Diversity describes how varied the configurations are

Models

- ML based: M3GNet approximates the energy of the final configuration. M3GNet is trained on VASP calculation, so results are expected to be consistent with VASP
- DFT: Computationally expensive option. Structures are relaxed once with VASP code.

Both results are compatible with MP-20 values

Evaluations

Method	Validity Check		Coverage		Property Distribution		Metastable	Stable
	Structural↑	Composition↑	Recall↑	Precision↑	wdist (ρ)↓	wdist (N_{el})↓	M3GNet ↑	DFT† ↑
CDVAE	**1.00**	0.867	0.991	0.995	0.688	1.43	28.8%	5.4%
LM-CH	0.848	0.835	0.9925	0.9789	0.864	0.13	n/a	n/a
LM-AC	0.958	0.889	0.996	0.9855	0.696	0.09	n/a	n/a
LLaMA-2								
7B ($\tau = 1.0$)	0.918	0.879	0.969	0.960	3.85	0.96	35.1%	6.7%
7B ($\tau = 0.7$)	0.964	0.933	0.911	0.949	3.61	1.06	35.0%	6.2%
13B ($\tau = 1.0$)	0.933	0.900	0.946	0.988	2.20	0.05	33.4%	8.7%
13B ($\tau = 0.7$)	0.955	0.924	0.889	0.979	2.13	0.10	**38.0%**	14.4%
70B ($\tau = 1.0$)	0.965	0.863	0.968	0.983	1.72	0.55	**35.4%**	10.0%
70B ($\tau = 0.7$)	**0.996**	**0.954**	0.858	0.989	0.81	0.44	**49.8%**	10.6%

† Fraction of structures that are first predicted by M3GNet to have $E_{\text{hull}}^{\text{M3GNet}} < 0.1$ eV/atom, and then verified with DFT to have $E_{\text{hull}}^{\text{DFT}} < 0.0$ eV/atom.

Evaluations

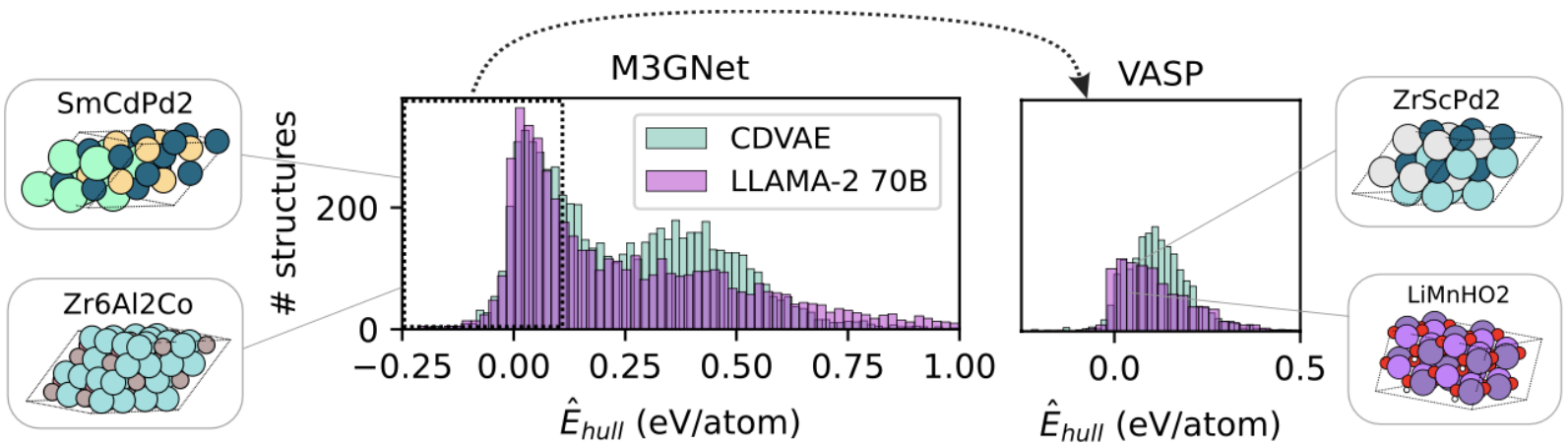


Figure: Distribution of stable and metastable results

Symmetry Metrics

It's assumed that model learned and can express structural symmetries. IPT (Increase in Perplexity under Transformations) is a proposed metric for assessing models invariance to transformations.

$$\mathrm{IPT}(\mathrm{s}) = \mathbb{E}_{\mathrm{g}\in\mathrm{G}}\left[\mathrm{PPL}(\mathrm{t}_{\mathrm{g}}(\mathrm{s})) - \mathrm{PPL}(\mathrm{t}_{\mathrm{g}^*}(\mathrm{s}))\right] \tag{5}$$

$$\mathrm{g}^* = \arg\min \mathrm{PPL}(\mathrm{t}_{\mathrm{g}}(\mathrm{s})) \tag{6}$$

Symmetry Metrics

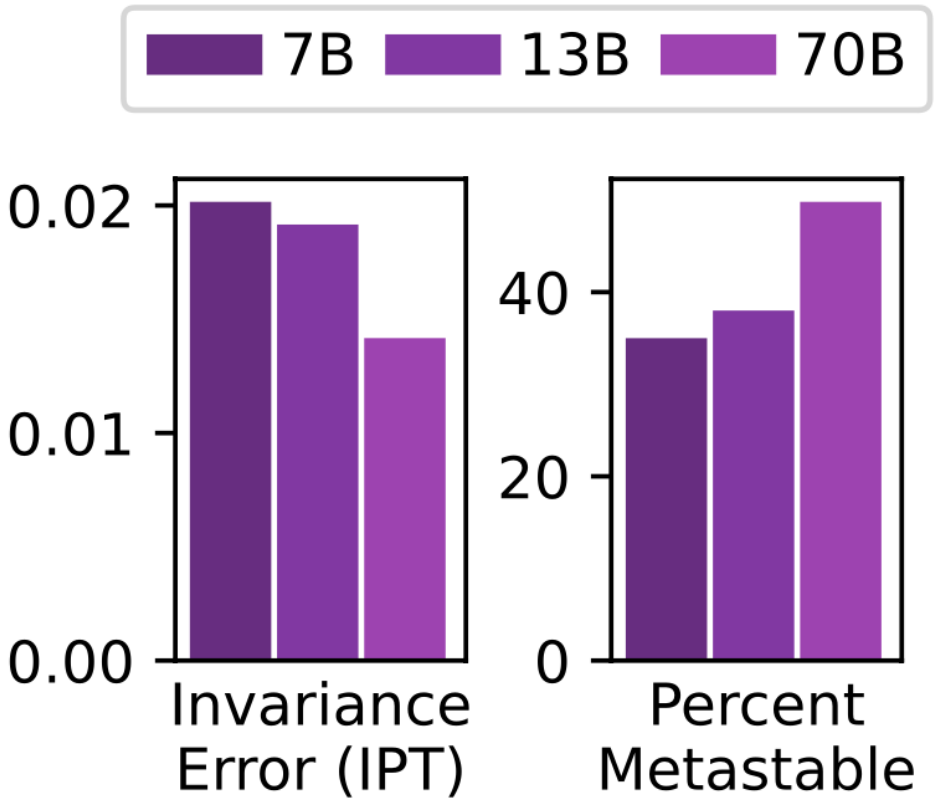
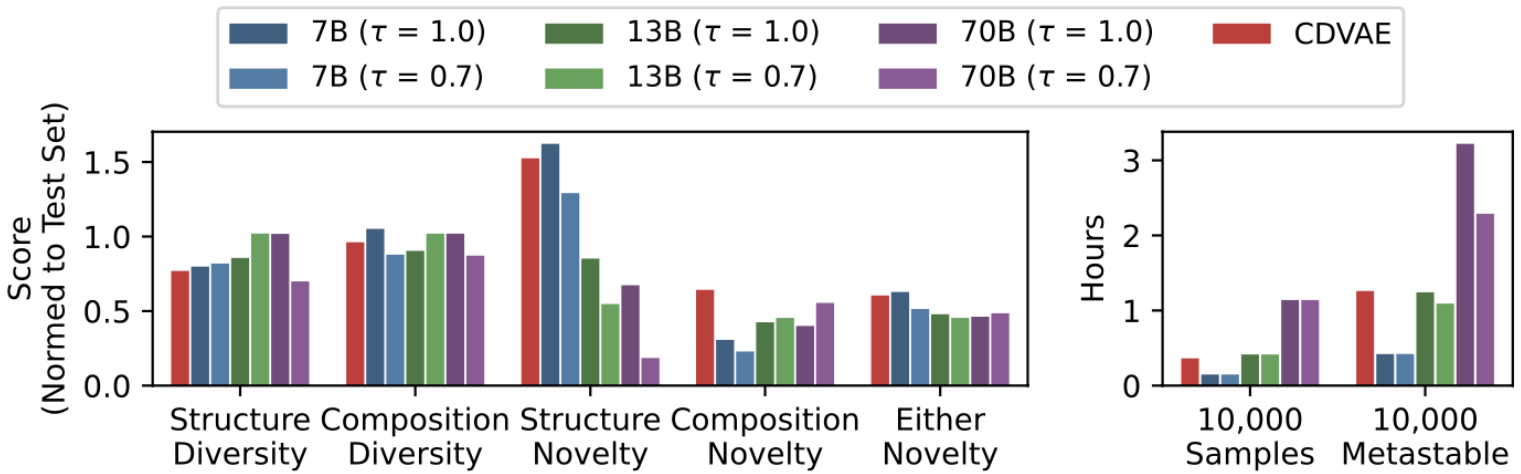


Figure: Translation invariance on test data

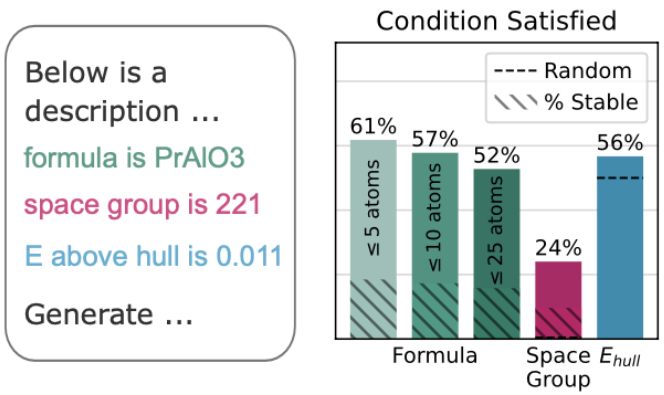
Diversity, Novelty and Sampling Speed

Validity is a prerequisite. Novelty and sampling speed has more practical significance. Novelty is calculated as the distance to the closest neighbor in the training set. Sampling speed is measured as time required to sample 10,000 samples.



Benchmarks on Text-conditioned Generation

For text-conditioned generation starting prompt includes small amount of text, describing conditions for generation. Conditioning on composition, stability and space group number tested. Stability is measured with energy above hull calculations with M3GNet.



Benchmarks on Infilling

In many occasions starting from scratch is unnecessary and slow. It's easier to start with already known materials and mutate over it. Starting materials are called template methods.

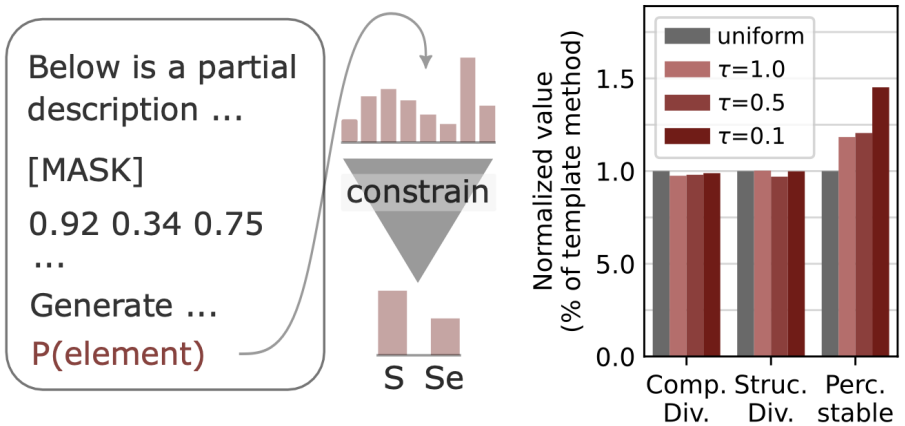


Table of Contents

Discussions

Are LLMs state of the art for crystal generation? Probably not.
What are the limitations?

Thank you for your attention.