Fine-Tuned Language Models Generate Stable Inorganic Materials as Text (ICLR)

Nate Gruver[1] Anuroop Sriram[2] Andrea Madotto[2]
Andrew Gordon Wilson[1] C. Lawrence Zitnick[2] Zachary Ulissi[2]
[1]NYU [2]Meta FAIR

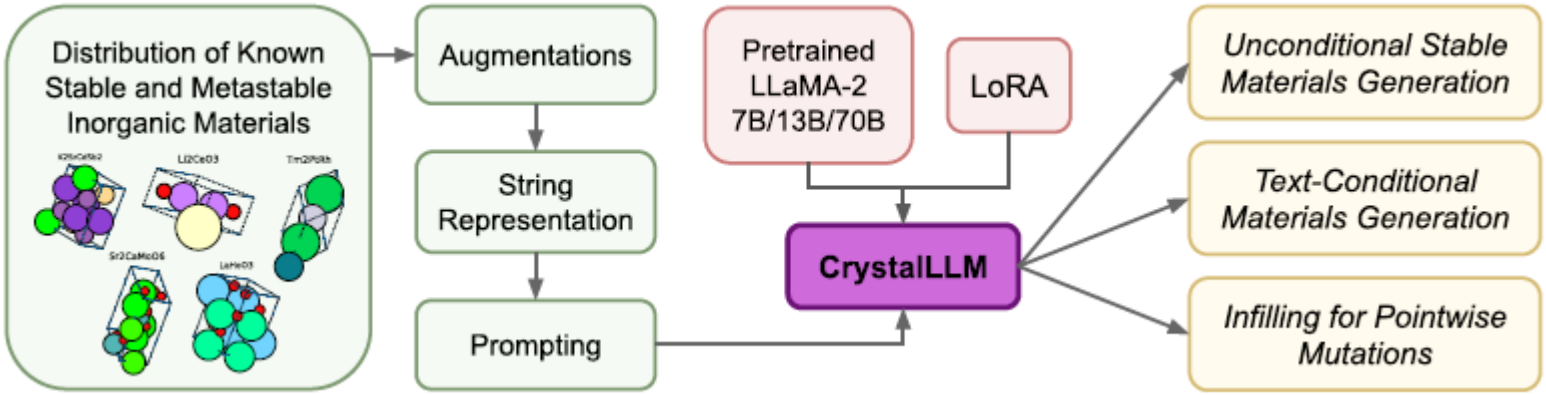
By *Dr. Jaewoong Choi*

(Dr. B. Lee group @KIST)

2024.04.04

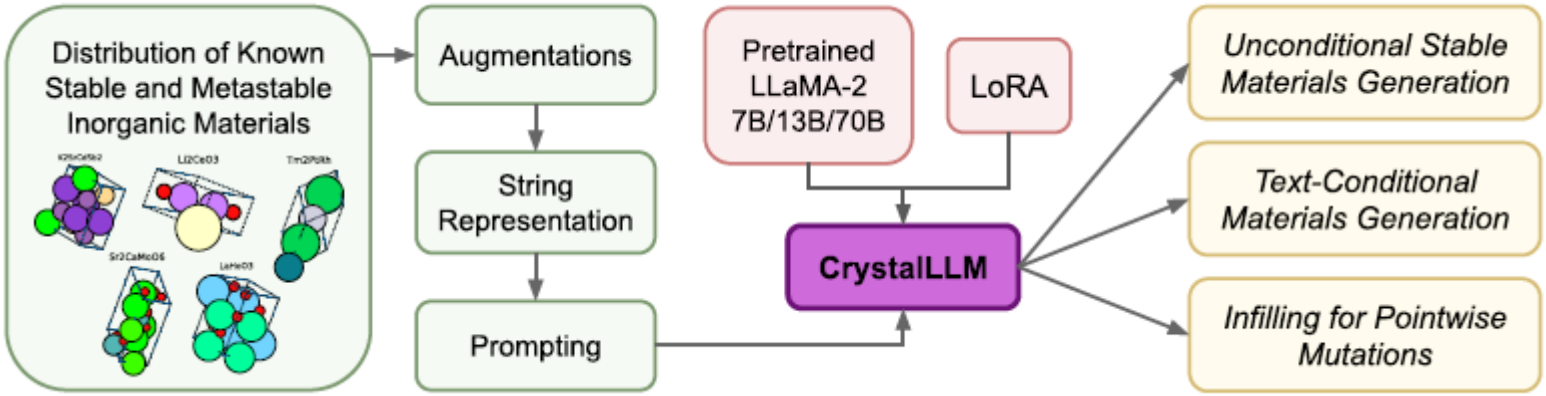
0. Summary

- Fine-tuning of large language models for the generation of stable materials
  - ✓ Using energy above hull calculations from both learned ML potentials and gold-standard DFT calculations, they show that their strongest model (fine-tuned LLaMA-2 70B) can generate materials predicted to be metastable at about twice the rate (49% vs 28%) of CDVAE.



1. Introduction

- The proposed approach proceeds as follows:
  - ✓ (1) Encode crystals as new-line separated strings and combine with text instructions.
  - ✓ (2) Perform parameter efficient fine tuning (PEFT) on LLaMA-2 with a multi-task curriculum and translation augmentations.
  - ✓ (3) Evaluate our method with Materials Project data comparing against an invariant diffusion model and a sequence model trained from scratch.

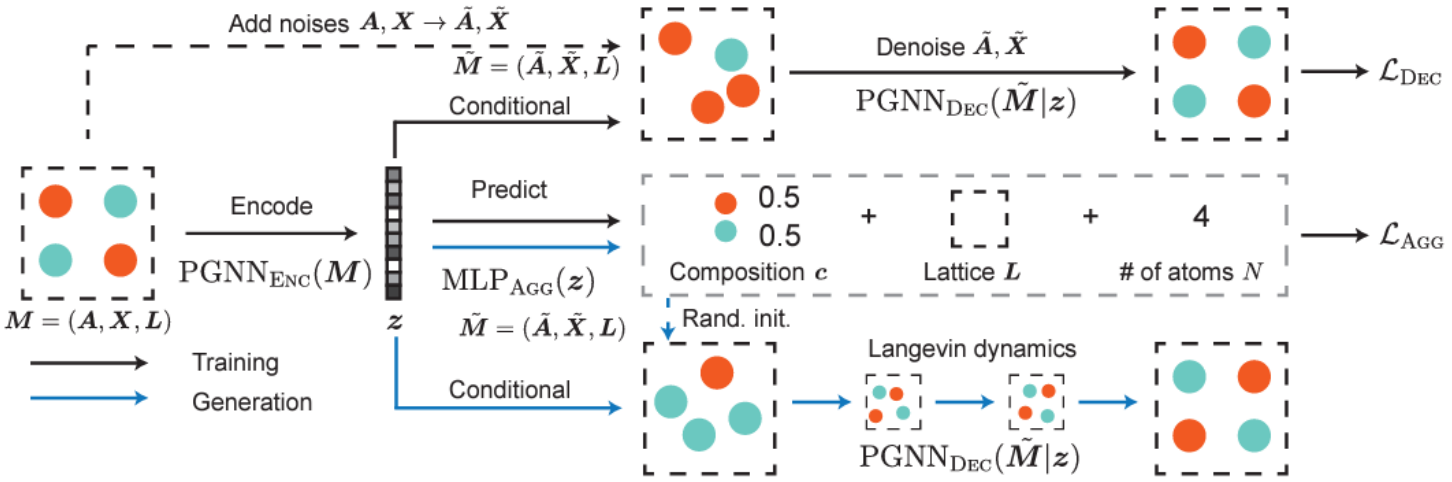


2. Related works

- Two central challenges in applying generative models to crystals and related atomistic data.
  - ✓ (1) Discrete and continuous nature of atoms.
    - Each atom has both an element identity and a position in three dimensional space.
    - But, generative modelling approaches are different for discrete or continuous data.
  - ✓ (2) Prevalence of symmetries in atomistic data.
    - The unit cell is common representation of crystals, capturing translation invariance underlying structure.
    - Symmetries can pose challenges to deep learning models because they entail constraints on the functions that neural networks can learn.

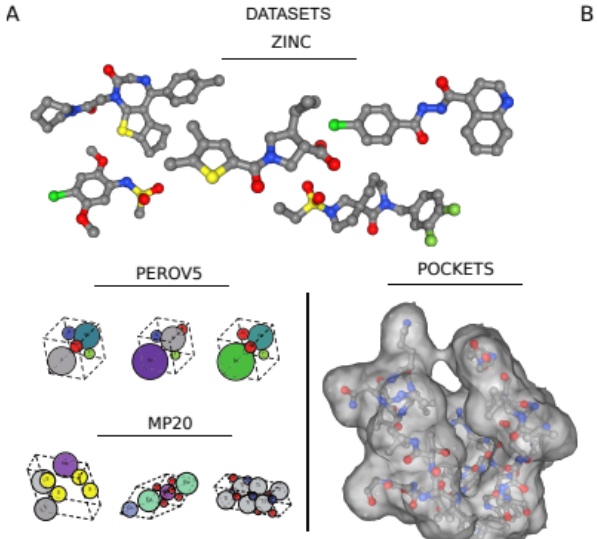
2. Related works

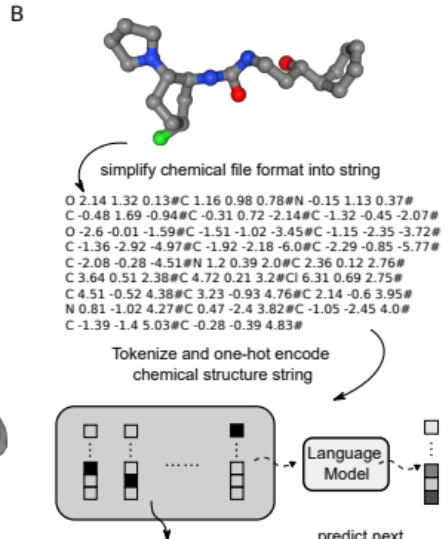
- Diffusion models
  - ✓ Xie et al. (2021) introduced crystal diffusion variational autoencoder (CDVAE) to directly deal with both of these challenges.
    - The decoder generates materials in a diffusion process that moves atomic coordinates towards a lower energy state and updates atom types to satisfy bonding preferences between neighbors.
    - Explicitly encodes interactions across periodic boundaries and respects permutation, translation, rotation, and periodic invariances.



2. Related works

- Language models: Shepherd & Guzik (2023)
  - ✓ LM-AC: Atom+coordinate-level models use a larger vocabulary of 100-10K tokens consisting of atom types tokens C, N, or atom-residue tokens 'HIS-C', 'HIS-N', and coordinate tokens like '-1.9', '-1.98' or '-1.987'.
    - LM-CH: Character-level models use a small vocabulary of 30 tokens consisting of atom type tokens 'C', 'N', digit characters and minus sign and other file symbols like newline character.

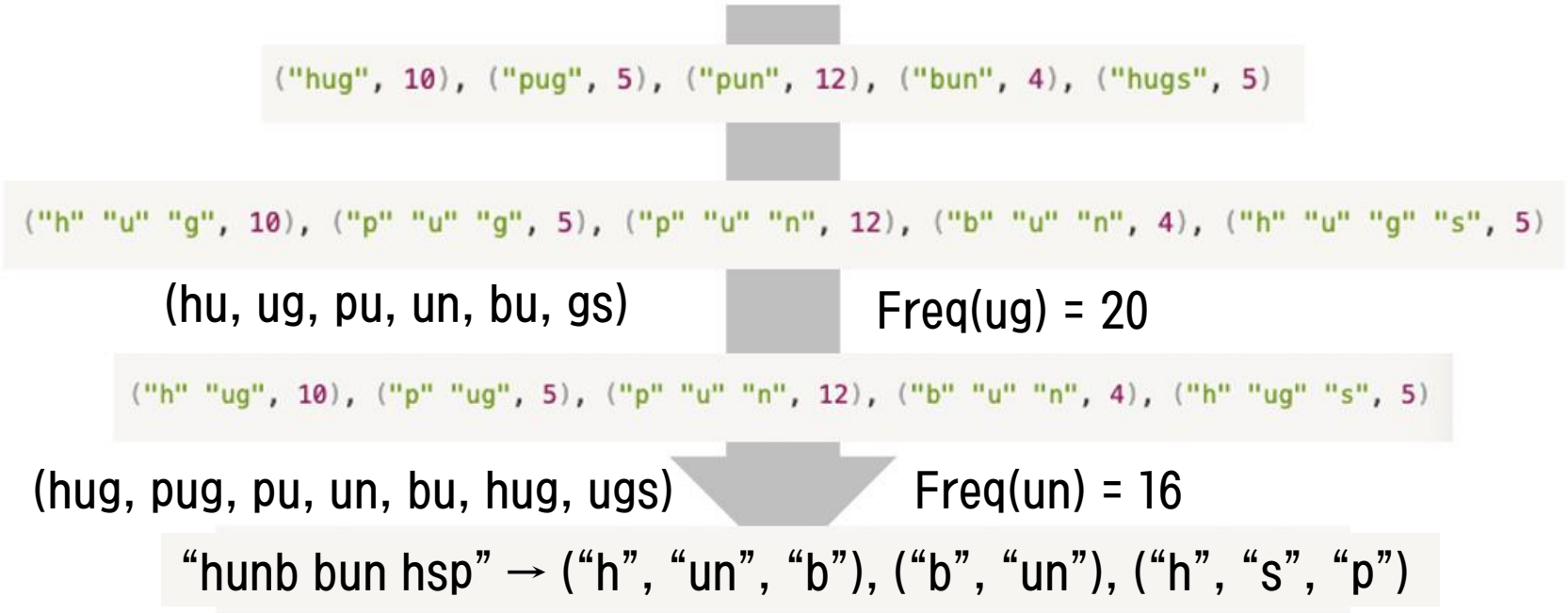




Data	Model	Valid (%) ↑ Struc.	Comp.	COV (%) ↑ R.	P.	WA ↓ ρ	#
Perov5	Train	100.0	98.60	100.0	100.0	0.010	0.008
	FTCP	0.24	54.24	0.00	0.00	10.27	0.630
	GSchNet	99.92	**98.79**	0.18	0.23	1.625	0.037
	PGSchNet	79.63	**99.13**	0.37	0.25	0.276	0.455
	CDVAE	**100.0**	**98.59**	**99.45**	98.46	0.126	0.063
	LM-CH	**100.0**	**98.51**	**99.60**	**99.42**	**0.071**	**0.036**
	LM-AC	**100.0**	**98.79**	**98.78**	**99.36**	**0.089**	**0.028**
MP20	Train	100.0	91.13	100.0	100.0	0.051	0.016
	FTCP	1.55	48.37	4.72	0.09	23.71	0.736
	GSchNet	99.65	75.96	38.33	**99.57**	3.034	0.641
	PGSchNet	77.51	76.40	41.93	**99.74**	4.04	0.623
	CDVAE	**100.0**	86.70	99.15	**99.49**	**0.688**	1.432
	LM-CH	84.81	83.55	**99.25**	97.89	0.864	**0.132**
	LM-AC	95.81	**88.87**	**99.60**	98.55	0.696	**0.092**

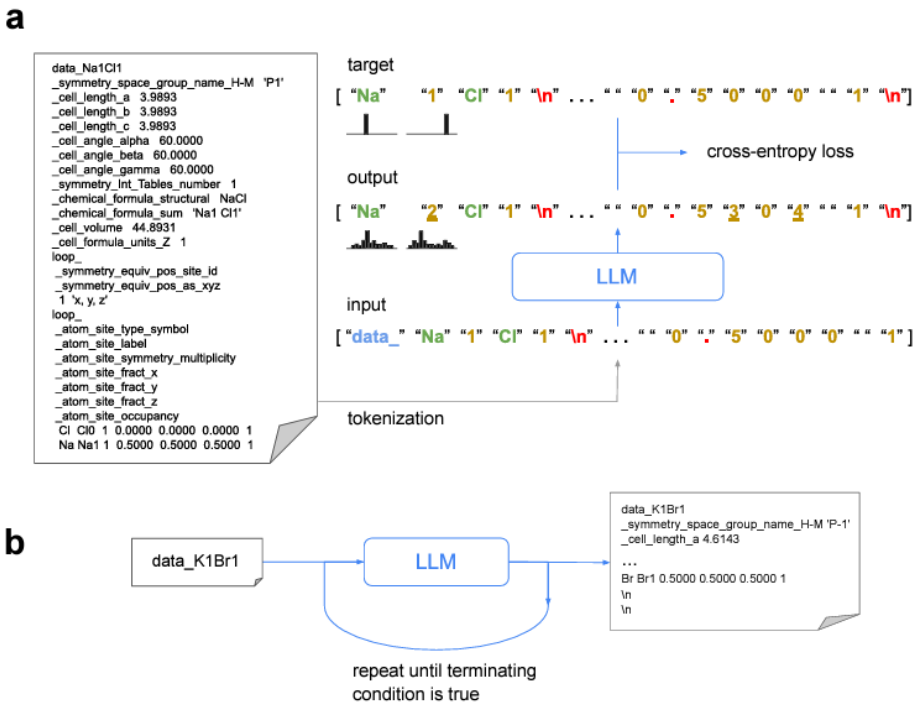
2. Related works

❖ Tokenization – BPE (Byte Pair Encoding)



2. Related works

- Language models: Antunes et al (2023)
  - ✓ LMs are used to generate crystal structures as discrete sequences by training from scratch on millions of CIF strings, from Materials Project, OQMD, NOMAD,
    - Consists of compounds containing anywhere from 1 to 10 elements and contains roughly 800,000 unique formulas, and 1.2 million unique cell compositions.



2. Related works

- Language models: Antunes et al (2023)

Pre-processed CIF file for PbTe (Z=2, Pmma)

```
data_Te2Pb2
loop_
_atom_type_symbol
_atom_type_electronegativity
_atom_type_radius
_atom_type_ionic_radius
Te 2.1000 1.4000 1.2933
Pb 2.3300 1.8000 1.1225
_symmetry_space_group_name_H-M Pmma
_cell_length_a 5.6440
_cell_length_b 4.0012
_cell_length_c 5.6807
_cell_angle_alpha 90.0000
_cell_angle_beta 90.0000
_cell_angle_gamma 90.0000
_symmetry_Int_Tables_number 51
_chemical_formula_structural TePb
_chemical_formula_sum 'Te2 Pb2'
_cell_volume 128.2864
_cell_formula_units_Z 2
loop_
_symmetry_equiv_pos_site_id
_symmetry_equiv_pos_as_xyz
1 'x, y, z'
loop_
_atom_site_type_symbol
_atom_site_label
_atom_site_symmetry_multiplicity
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_occupancy
Te Te0 2 0.2500 0.5000 0.7357 1
Pb Pb1 2 0.2500 0.0000 0.2691 1
```

Supported atom tokens.

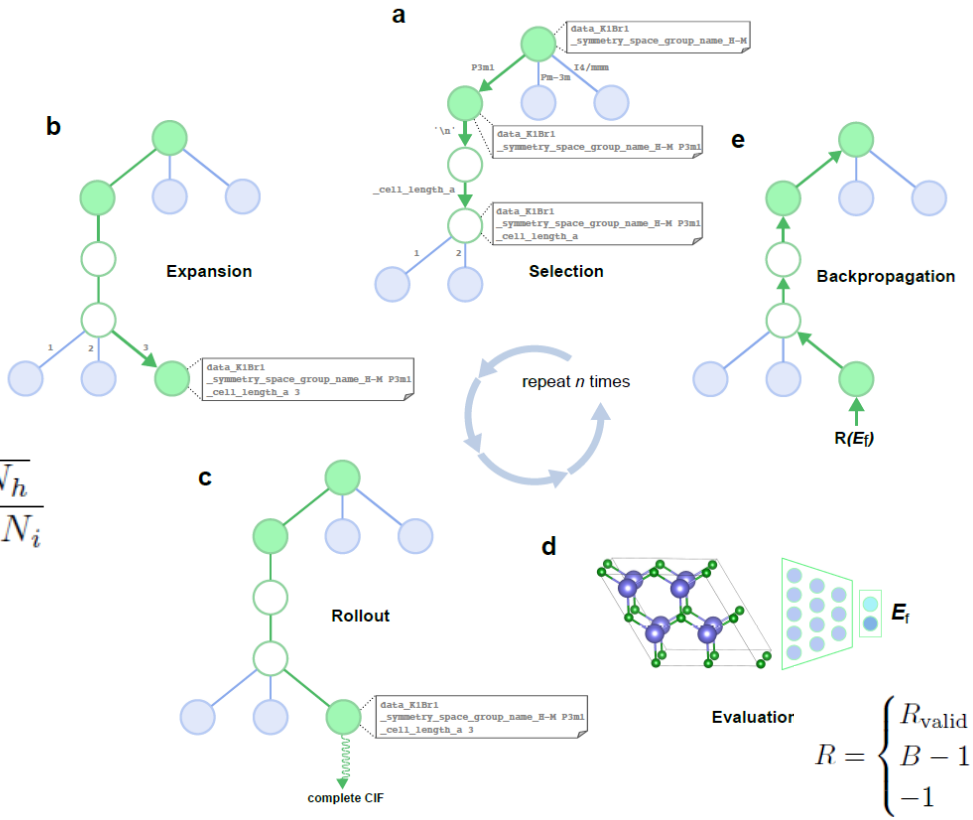
```
Ac Ag Al Ar As Au B Ba Be Bi Br C Ca Cd Ce Cl Co Cr Cs Cu Dy Er Eu F Fe Ga Gd Ge
H He Hf Hg Ho I In Ir K Kr La Li Lu Mg Mn Mo N Na Nb Nd Ne Ni Np O Os P Pa Pb Pd
Pm Pr Pt Pu Rb Re Rh Ru S Sb Sc Se Si Sm Sn Sr Ta Tb Tc Te Th Ti Tl Tm U V W Xe
Y Yb Zn Zr
```

Supported CIF tag tokens.

```
_cell_length_b                  _atom_site_occupancy
_atom_site_attached_hydrogens   _cell_length_a
_cell_angle_beta                _symmetry_equiv_pos_as_xyz
_cell_angle_gamma               _atom_site_fract_x
_symmetry_space_group_name_H-M  _symmetry_Int_Tables_number
_chemical_formula_structural    _chemical_name_systematic
_atom_site_fract_y              _atom_site_symmetry_multiplicity
_chemical_formula_sum           _atom_site_label
_atom_site_type_symbol          _cell_length_c
_atom_site_B_iso_or_equiv       _symmetry_equiv_pos_site_id
_cell_volume                    _atom_site_fract_z
_cell_angle_alpha               _cell_formula_units_Z
loop_                           data_
_atom_type_symbol               _atom_type_electronegativity *
_atom_type_radius *             _atom_type_ionic_radius *
_atom_type_oxidation_number
```

2. Related works

- Language models: Antunes et al (2023)
  - ✓ The integration with predictors of formation energy permits the use of a Monte Carlo Tree Search algorithm to improve the generation of meaningful structures.



$$\mathrm{PUCT}(i) = \frac{w_i}{N_i} + c_{\mathrm{puct}} P_i \frac{\sqrt{N_h}}{1+N_i}$$

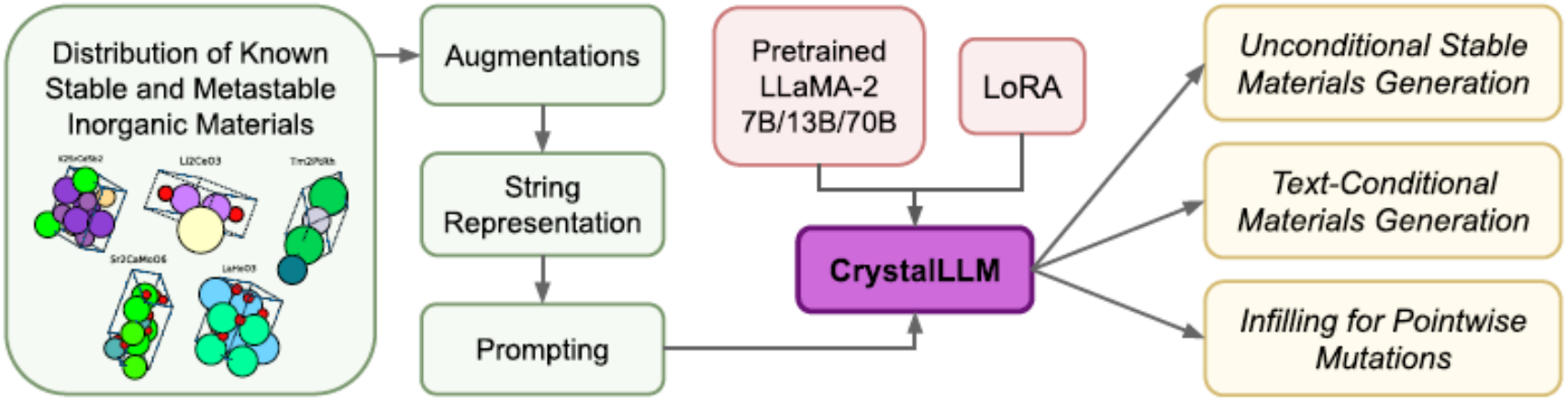
$$R_{\mathrm{valid}} = \frac{1}{1+e^{\lambda((E_{\mathrm{f}}-\mu)/\sigma)}}$$

$$R = \begin{cases} R_{\mathrm{valid}} & \text{if valid,} \\ B-1 & \text{if bond lengths unreasonable,} \\ -1 & \text{otherwise} \end{cases}$$

2. Related works

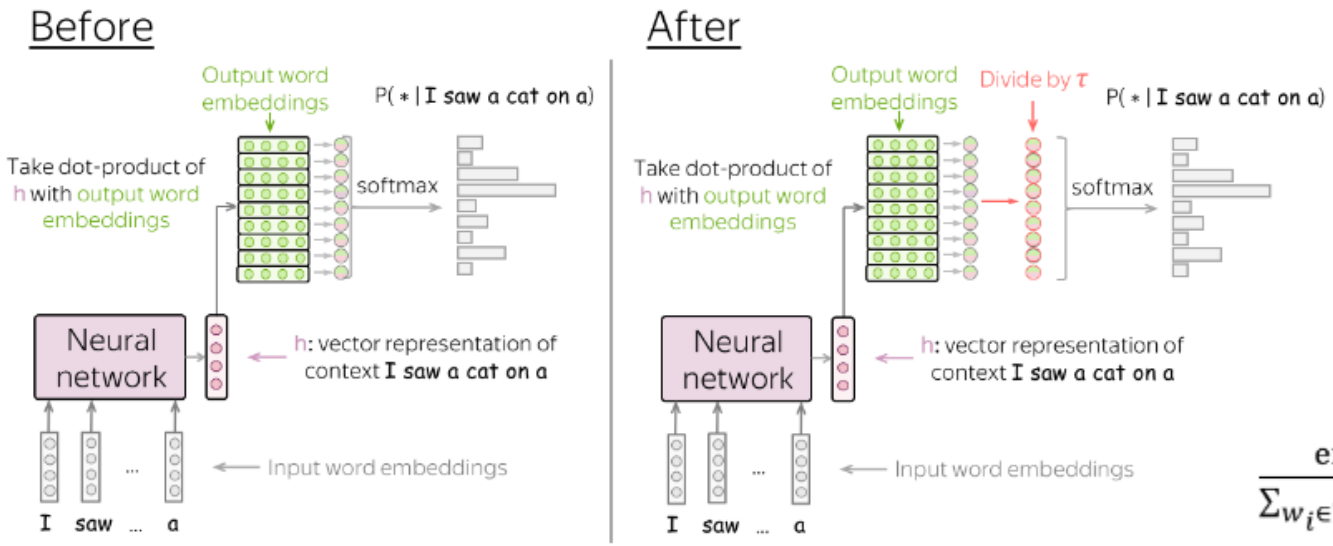
❖ Proposed approach

- ✓ By using a pre-trained LLM, we can achieve high rates of validity without crystal-specific tokenization (Shepherd & Guzik, 2023) or millions of auxiliary structures (Antunes et al, 2023).
  - Aligned with prior findings, they show that larger models, which are often more effective compressors of data, demonstrate improved ability to learn symmetries from the training data and augmentation.



3. Background

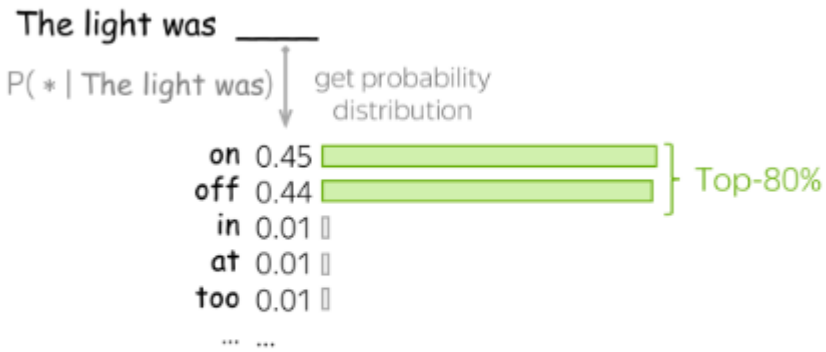
- Language Modelling: Temperature 지표
  - ✓ LLMs perform next-token prediction over sequences: $p(w_{t+1} \mid w_{0:t})$ and the sampling procedure is typically modulated with two hyperparameters
    - Temperature: flatten the conditional distributions to uniform (high temperature) or collapse them around their maximal probabilities (low temperature).



$$\frac{\exp(h^T w)}{\sum_{w_i \in V} \exp(h^T w_i)} \rightarrow \frac{\exp\left(\frac{h^T w}{\tau}\right)}{\sum_{w_i \in V} \exp\left(\frac{h^T w_i}{\tau}\right)}$$

3. Background

- ❖ Language Modelling: Nucleus 지표
  - ✓ LLMs perform next-token prediction over sequences: $p(w_{t+1} \mid w_{0:t})$ and the sampling procedure is typically modulated with two hyperparameters
    - Nucleus size: limits which tokens can be sampled based on the cumulative distribution function, clipping out values that contribute very little mass.



3. Background

❖ BPE tokenization의 단점은 숫자에 취약한 점.

✓ BPE typically breaks numbers into irregular substrings instead of individual digits.

- While breaking numbers into multi-digit tokens creates shorter sequences, it also complicates learning basic arithmetic operations, which typically operate at the level of individual digits.

✓ LLaMA-2 models break numbers into a sequence of digits, which has been shown to dramatically improve performance on arithmetic tasks.

Model	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
GPT-2 (from scratch)	0.946	0.878	0.807	0.757	0.740
LLaMA-13B (LoRA)	0.457	0.432	0.424	0.401	0.385
LLaMA-70B (LoRA)	0.402	0.344	0.325	0.305	0.296

3. Background

- LoRA(Low-Rank Adaptation)
  - ✓ There exists a low dimension reparameterization that is as effective for fine-tuning as the full parameter space.
    - LORA is designed to fine-tune large-scale models efficiently by targeting a small subset of the model's weights that have the most significant impact on the task at hand.

	Weight Type	$r=1$	$r=2$	$r=4$	$r=8$	$r=64$
WikiSQL(±0.5%)	W_q	68.8	69.6	70.5	70.4	70.0
	W_q, W_v	73.4	73.3	73.7	73.8	73.5
	W_q, W_k, W_v, W_o	74.1	73.7	74.0	74.0	73.9
MultiNLI (±0.1%)	W_q	90.7	90.9	91.1	90.7	90.7
	W_q, W_v	91.3	91.4	91.3	91.6	91.4
	W_q, W_k, W_v, W_o	91.2	91.7	91.7	91.5	91.4

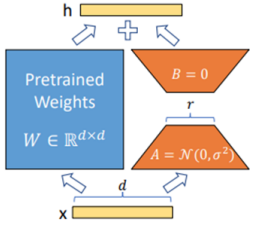


Figure 1

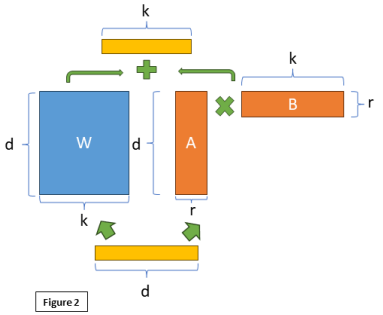
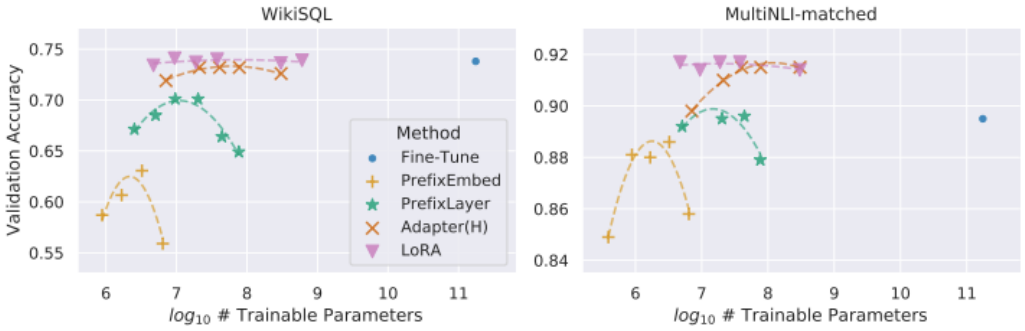
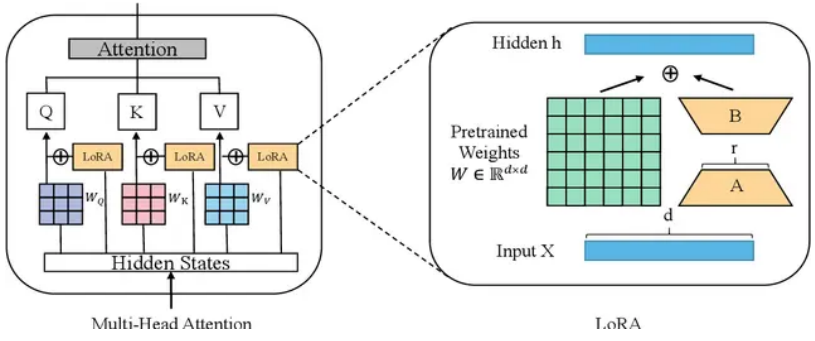


Figure 2

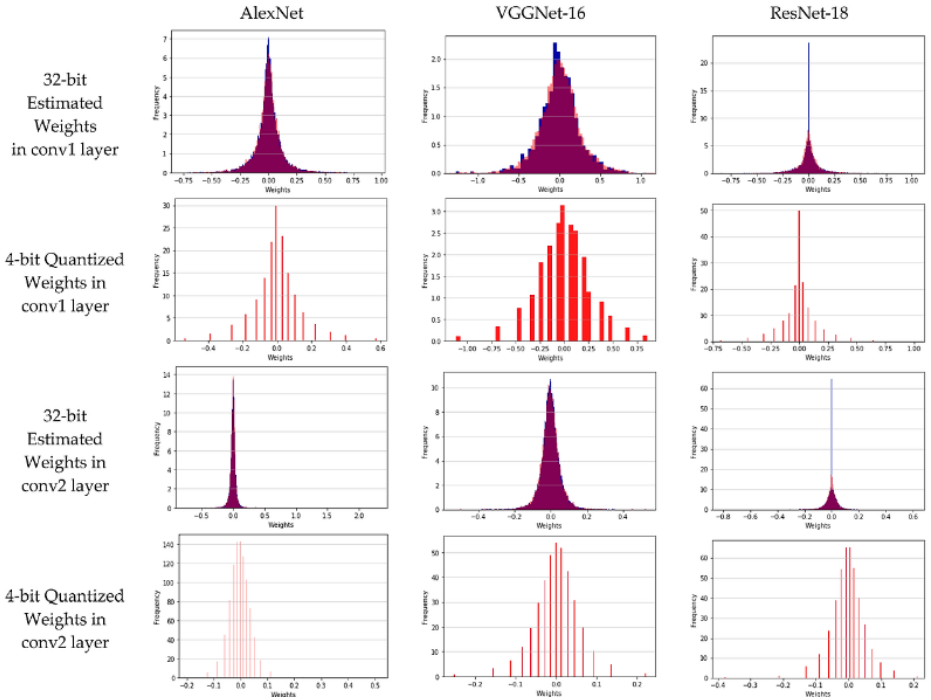




3. Background

- Quantization.
  - ✓ Model Quantization is a technique used to reduce the size of large neural networks, including large language models (LLMs) by modifying the precision of their weights.
    - The term quantization refers to the process of mapping continuous infinite values to a smaller set of discrete finite values.

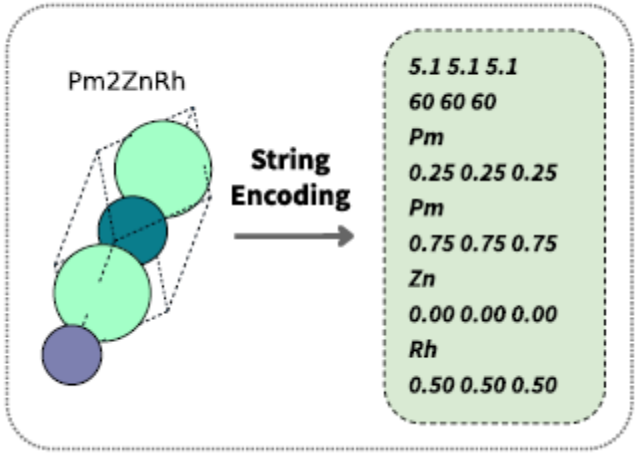
Model	Original Size (FP16)	Quantized Size (INT4)
Llama2-7B	13.5 GB	3.9 GB
Llama2-13B	26.1 GB	7.3 GB
Llama2-70B	138 GB	40.7 GB



3. Background

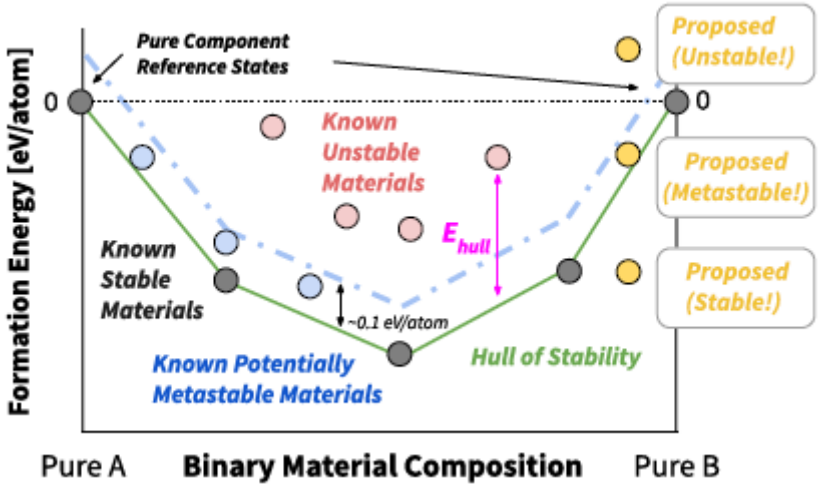
- Crystal 구조와 에너지 예측
  - ✓ Periodic materials are defined by a unit cell repeated infinitely along all three dimensions.
  - ✓ The unit cell comprises a lattice (parallelepiped) with side lengths and angles and a bulk material can be defined as follows:

$$C = (l_1, l_2, l_3, \theta_1, \theta_2, \theta_3, e_1, x_1, y_1, z_1, ..., e_N, x_N, y_N, z_N)\,.$$



3. Background

- ❖ hypothetical materials의 안정성
  - ✓ The energy of a crystal is influenced by its composition, with certain element ratios favored.
  - ✓ A crystal's stability is assessed by its energy above hull (Ehull), where Ehull < 0 signifies stability and Ehull < 0.1 eV/atom indicates practical utility as a metastable material.

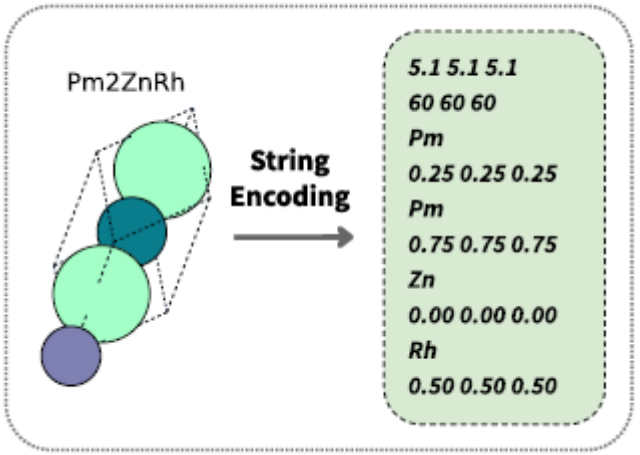


4. Method

❖ 1. String 포맷팅과 토큰화

✓ Convert the crystal tuple C (Equation 1) using fixed precision numbers

$$C = (l_1, l_2, l_3, \theta_1, \theta_2, \theta_3, e_1, x_1, y_1, z_1, ..., e_N, x_N, y_N, z_N)\,.$$



✓ Chose LLaMA-2 models because they are both state-of-the-art in overall performance among open-source models and because they tokenize numbers as individual digits by default.

	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
Special Crystal Tokens	0.783	0.693	0.623	0.611	0.588
Shared Tokenization	0.457	0.432	0.424	0.401	0.385

4. Method

❖ 2. 프롬프트 설계

- ✓ To train a model that can be used for many tasks, they use task-specific prompts: (1) unconditional generation, (2) text-conditional generation, and (3) infilling.
  - During training all three tasks are included through random sampling, with two thirds generation and one third infilling.

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.

4. Method

❖ 3. 데이터 증강

- ✓ Crystals structures are symmetric under translational.
  - All atomic coordinates can be shifted modulo the lattice boundaries without changing the resulting material structure.
  - Similarly, the ordering of atoms within the lattice is irrelevant to the underlying material (permutation invariance).
- ✓ Apply random uniform translations to the fractional coordinates.
  - Chose not to augment the ordering of atoms because these variables often contained valuable information.

Setting	Structural Validity	Compositional Validity
Fractional coords	91.4%	83.2%
Absolute coords	90.8%	80.5%
No permutations	92.5%	82.9%
With permutations	89.2%	81.7%

5. Experiments

❖ 데이터셋과 모델

✓ For consistency with prior works, they used MP-20, a dataset of 45231 materials and 127609 crystals, when training for unconditional generation.

- All structures in MP-20 are stable, and therefore an effective generative model trained on MP-20 should tend to propose new crystals that are at least metastable.
- For text-conditioned generation, they train with all forms of prompting with basic property information, such as the space group number, band gap, E_hull and the chemical formula.

- **LLaMA-2 7B**: Batch size of 256 for 65 epochs with a cosine annealed learning rate of 0.0005. LoRA rank 8 and alpha 32.
- **LLaMA-2 13B**: Batch size of 256 for 44 epochs with a cosine annealed learning rate of 0.0005. LoRA rank 8 and alpha 32.
- **LLaMA-2 70B**: Batch size of 32 for 21 epochs with a cosine annealed learning rate of 0.0005. LoRA rank 8 and alpha 32.

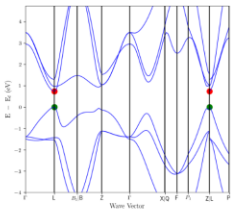
5. Experiments

❖ 평가지표 from Xie et al. (2021)

- ✓ Structural validity: non-overlapping atomic radii(overlapping taken to be both atoms within half a radius of each other).
- ✓ Compositional validity captures the net charge of the structure (only structures with net neutral total charge are valid).
- ✓ Diversity is computed as pairwise distance between samples under featurization of the structure and composition from matminer.

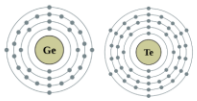
matminer

Bandstructure

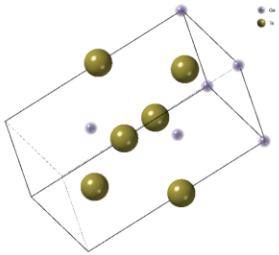


Composition

GeTe



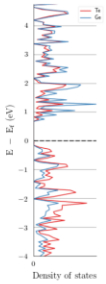
Site and Structure



Function

$f(x)$

DOS



pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Structure		Composition		Function	
Feature 1	Feature 2	Feature 3	Feature 4	Feature 1, 1/x	Feature 1, logx
0.1	1e-14	.003	223	10	-1
4.2	1.2e-12	.002	14	.238	.62
1.1	1e-6	.0031	101	.91	.041

5. Experiments

❖ 성능 평가

- ✓ Fine-tuned LLaMA-2 models were evaluated using validity as well as coverage and property metrics, which capture alignment between the ground truth and sampling distribution.
  - They added stability checks, which count the percentage of samples estimated to be stable by M3GNet and DFT.
  - They only run VASP calculations on materials that have already been predicted as metastable by M3GNet (<0.1 eV/atom $\hat{E}_{hull}$).

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.

5. Experiments

❖ 성능 평가 (Unconditional generation)

✓ They sampled 10,000 structures from each fine tuned LLaMA model, parsing a CIF from the generated string.

- They rejected the sample and draw another if a CIF cannot be parsed from the sampled string, which guarantees all samples can be interpreted as crystals but does not guarantee validity of the resulting crystal.

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%

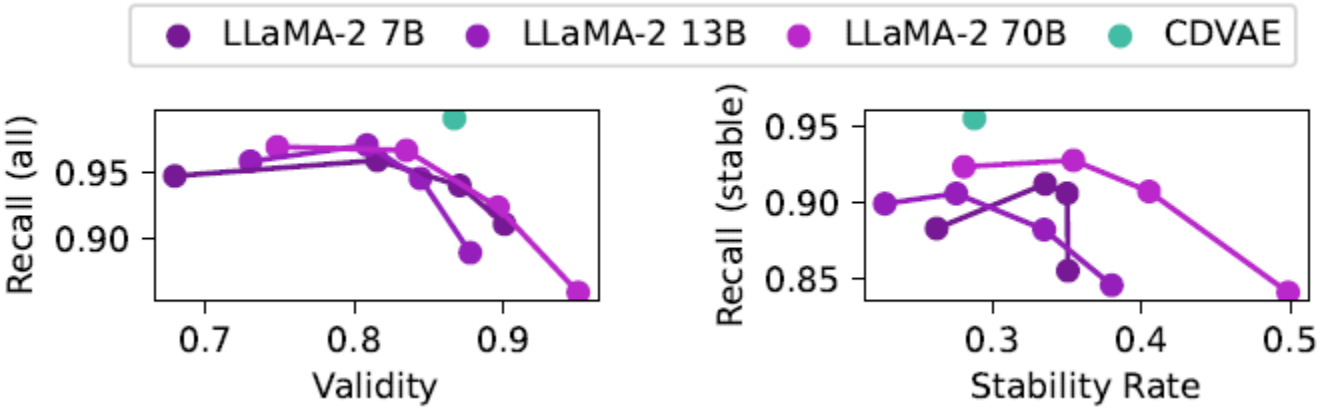
† Fraction of structures that are first predicted by M3GNet to have $E_{\text{hull}}^{\text{M3GNet}} < 0.1$ eV/atom, and then verified with DFT to have $E_{\text{hull}}^{\text{DFT}} < 0.0$ eV/atom.

5. Experiments

❖ (Unconditional generation) 파라미터에 따른 validity와 stability 비율

✓ Hyper-parameters like temperature and nucleus size can be used to trade-off validity and stability of samples with their coverage.

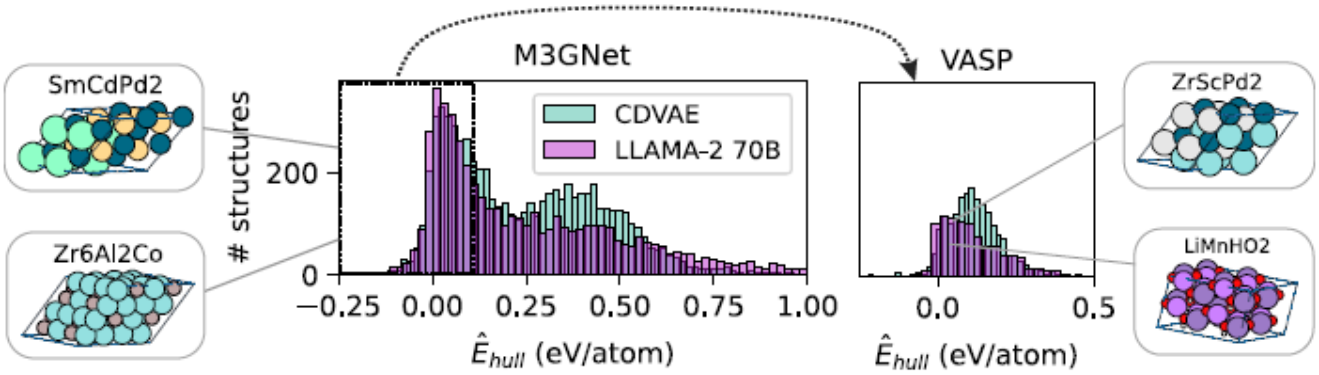
- Coverage most likely decreases because nucleus size and temperature collapse the distribution around samples with high likelihood, which are also more likely to be valid or stable.
- Lowering the temperature or restricting the nucleus size leads to significant improvements in validity/stability but incurs a cost to coverage of a held-out test set (recall).



5. Experiments

❖ (Unconditional generation) 환각 해결

- ✓ LLaMA-2 70B strikes an effective balance, generating high rates of stable materials with good coverage and diversity.
- ✓ Generation is completely unconstrained and therefore the model can hallucinates imaginary elements, but the problem can be easily avoided by constraining the tokens for element identities.
  - For example, 'Ln' is an imaginary element.




```
6.7 6.7 6.7
138.0 138.0 59.0
Ln
0.88 0.05 0.96
Ln
0.63 0.30 0.46
In
0.38 0.55 0.96
In
0.13 0.80 0.46
```

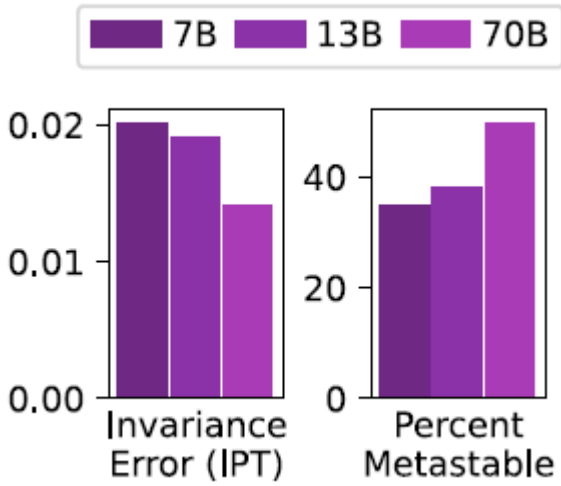
5. Experiments

- ❖ Symmetric learning: translation한 구조에 대해 일관되게 헷갈려하는지
  - ✓ As crystal structures have translational symmetry, the model's likelihood should be invariant to translations.
    - Increase in Perplexity under Transformation (IPT) as metric for assessing the invariance of language models to continuous group transformations.
      - PPL is the perplexity of the sequence, the exponent of the length-normalized cross entropy loss, PPL(s) = 2 CE(s)/n.
    - Larger models learn invariances from augmentations more effectively during training.

$$\mathrm{IPT}(s) = \mathbb{E}_{g \in G}[\mathrm{PPL}(t_g(s)) - \mathrm{PPL}(t_{g^*}(s))]$$

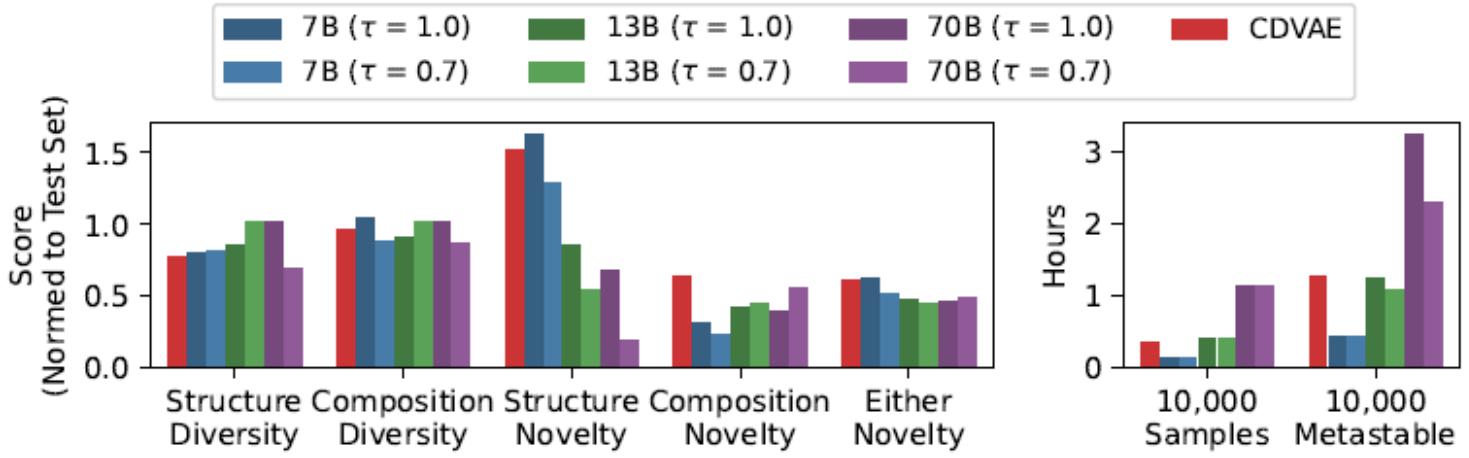
where

$$g^* = \arg\min \mathrm{PPL}(t_{g^*}(s))$$



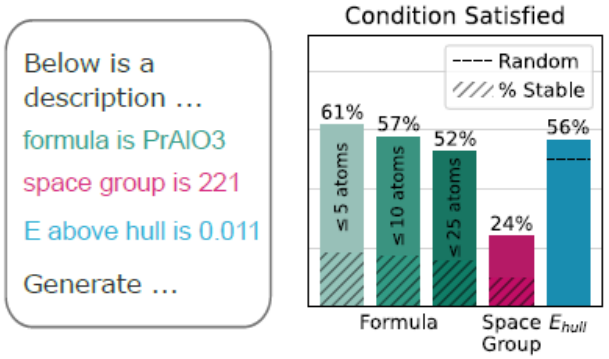
5. Experiments

- ❖ Diversity, novelty, and sampling speed
  - ✓ Interestingly, larger LLaMA models display less novel structures but more novel compositions.
  - ✓ Calculating pairwise distances for diversity and distance to the closest neighbor in the training set for novelty.
    - Diversity는 학습데이터 각각과의 거리 평균, Novelty는 학습데이터에서 제일 유사한 물질과의 유사도가 특정 임계값을 넘는지 평가 (a structural distance cutoff of 0.1 and composition distance cutoff of 2).



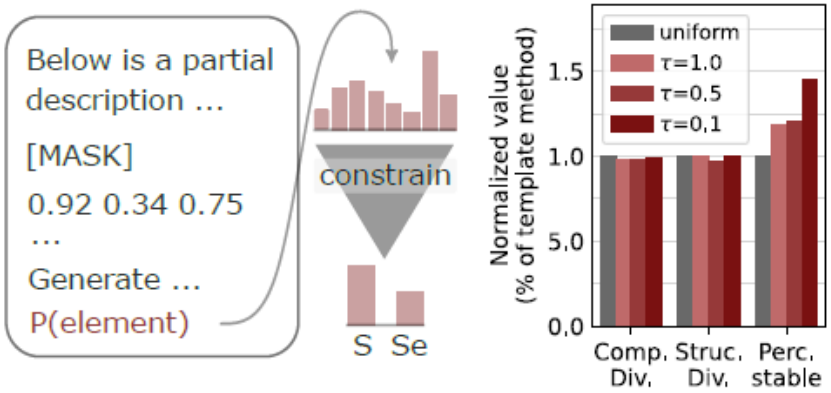
5. Experiments

- 2번째 task: Text-conditioned generation
  - ✓ Including additional information (space group number, composition, and E_hull) in the prompt.
    - For space group determination, use pymatgen's SpacegroupAnalyzer with a precision of 0.2 angstroms
  - ✓ Using the M3GNet's labels, compute the percentage of cases in which the condition was properly met.
    - It becomes less reliable as the number of atoms in the chemical formula increases.
    - Space group conditioning is more challenging, as it requires precise control and understanding of 3D structure.



5. Experiments

- ❖ 3번째 task: Infilling Existing Materials
  - ✓ Good starting materials' manufacturing processes are easier to adapt to related compositions, by making small edits to their composition
    - Uniform: construct a lookup table that maps each element to elements that have a similar atom radius when in the same oxidation state and choose an element uniformly at random and swap it with a random element chosen from the table.
    - Proposed: Using the fine-tuned LLM, they used the infilling prompt to obtain a distribution over elements (modulated with temperature), and constrain the elements using knowledge of atom radii and charge interactions.



Infill Prompt

<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>

onditional generation. Purple text stands in for string encodings of atoms.

6. Discussion

❖ Summary

- ✓ By generating a high rate of plausible stable, they demonstrated LLMs can be state-of-the-art generative models for atomistic domains with direct application of parameter-efficient instruction tuning and minimal task-specific modeling choices.
  - This approach to generative modeling opens the door to multi-task capabilities within a single sampling paradigm and multi-modal training on atoms and text.

Model	Batch size	Seconds / batch	Samples / hour	Hours / 10,000 crystals
CDVAE	512	n/a	n/a	1.260
LLaMA-2 7B	512	27.18	67814	0.147
LLaMA-2 13B	256	38.24	24100	0.414
LLaMA-2 70B	128	52.52	8774	1.139

Model	Hours / 10,000 crystals	Hours / 10,000 metastable (M3GNet) crystals
CDVAE	0.363	1.260
LLaMA-2 13B	0.416	1.094
LLaMA-2 70B	0.864	1.728

Thank you

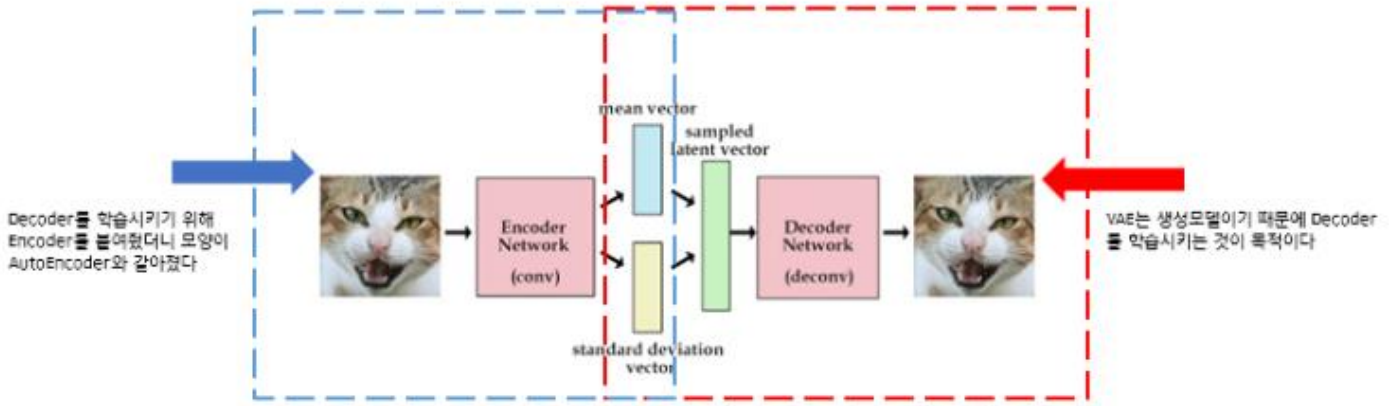
By *Dr. **Jaewoong Choi***

(Dr. B. Lee group @KIST)

2024.04.04

Appendix

❖ VAE



$$p(x) \approx \sum_i p(x|g_\theta(z_i))p(z_i)$$

If $p(x|g_\theta(z)) = \mathcal{N}(x|g_\theta(z), \sigma^2 * I)$, the negative log probability of X is proportional squared Euclidean distance between $g_\theta(z)$ and x.

x : Figure 3(a)
$z_{bad} \rightarrow g_\theta(z_{bad})$: Figure 3(b)
$z_{bad} \rightarrow g_\theta(z_{good})$: Figure 3(c) (identical to x but shifted down and to the right by half a pixel)

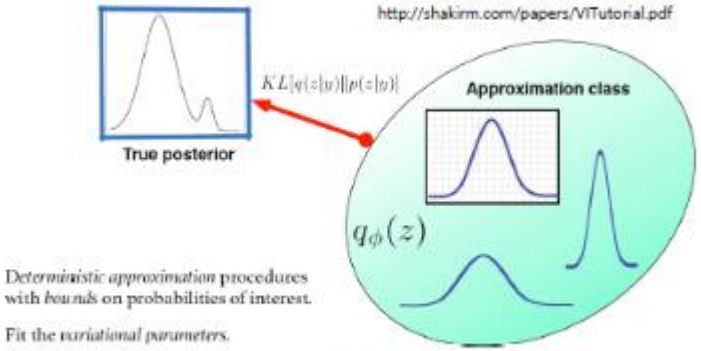
$$\|x - z_{bad}\|^2 < \|x - z_{good}\|^2 \rightarrow p(x|g_\theta(z_{bad})) > p(x|g_\theta(z_{good}))$$

Solution 1: we should set the σ hyperparameter of our Gaussian distribution such that this kind of erroroneous digit does not contribute to p(X) → hard..
Solution 2: we would likely need to sample many thousands of digits from z_{good} → hard..

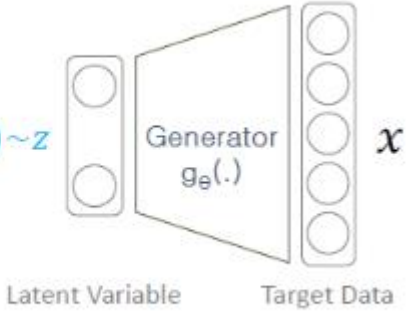
Figure 3: It's hard to measure the likelihood of images under a model using only sampling. Given an image X (a), the middle sample (b) is much closer in Euclidean distance than the one on the right (c). Because pixel distance is so different from perceptual distance, a sample needs to be extremely close in pixel distance to a datapoint X before it can be considered evidence that X is likely under the model.

Appendix

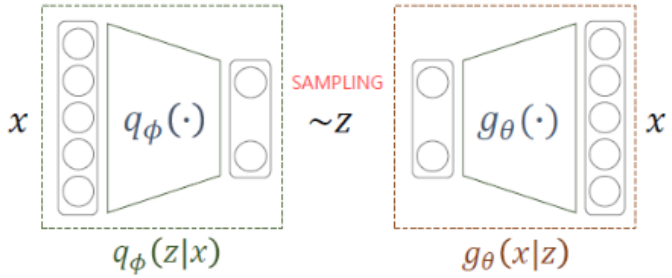
- VAE



[Variational Inference]



$$\arg\min_{\phi,\theta} \sum_i \underbrace{-\mathbb{E}_{q_\phi(z|x_i)}[\log(p(x_i|g_\theta(z)))] + KL(q_\phi(z|x_i)||p(z))}_{L_i(\phi,\theta,x_i)}$$

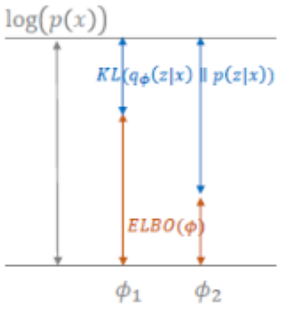


$$\log(p(x)) = \int \log(p(x)) q_\phi(z|x) dz \quad \leftarrow \int q_\phi(z|x)dz = 1$$

$$= \int \log\left(\frac{p(x,z)}{p(z|x)}\right) q_\phi(z|x) dz \quad \leftarrow p(x) = \frac{p(x,z)}{p(z|x)}$$

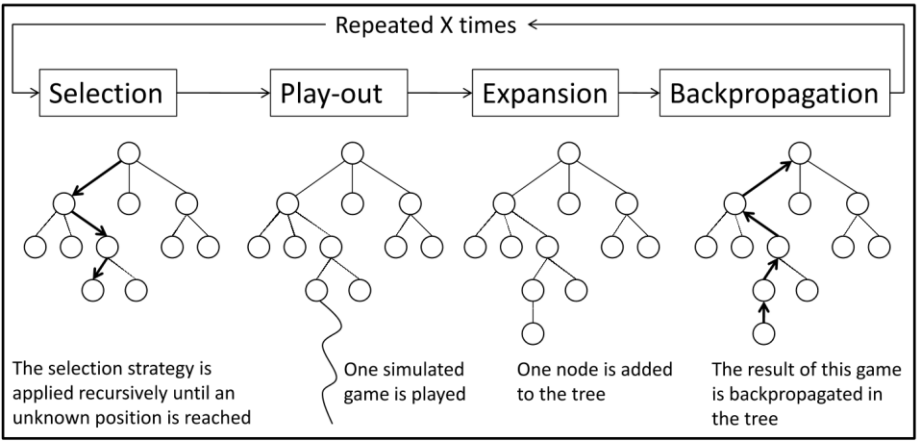
$$= \int \log\left(\frac{p(x,z)}{q_\phi(z|x)} \cdot \frac{q_\phi(z|x)}{p(z|x)}\right) q_\phi(z|x) dz$$

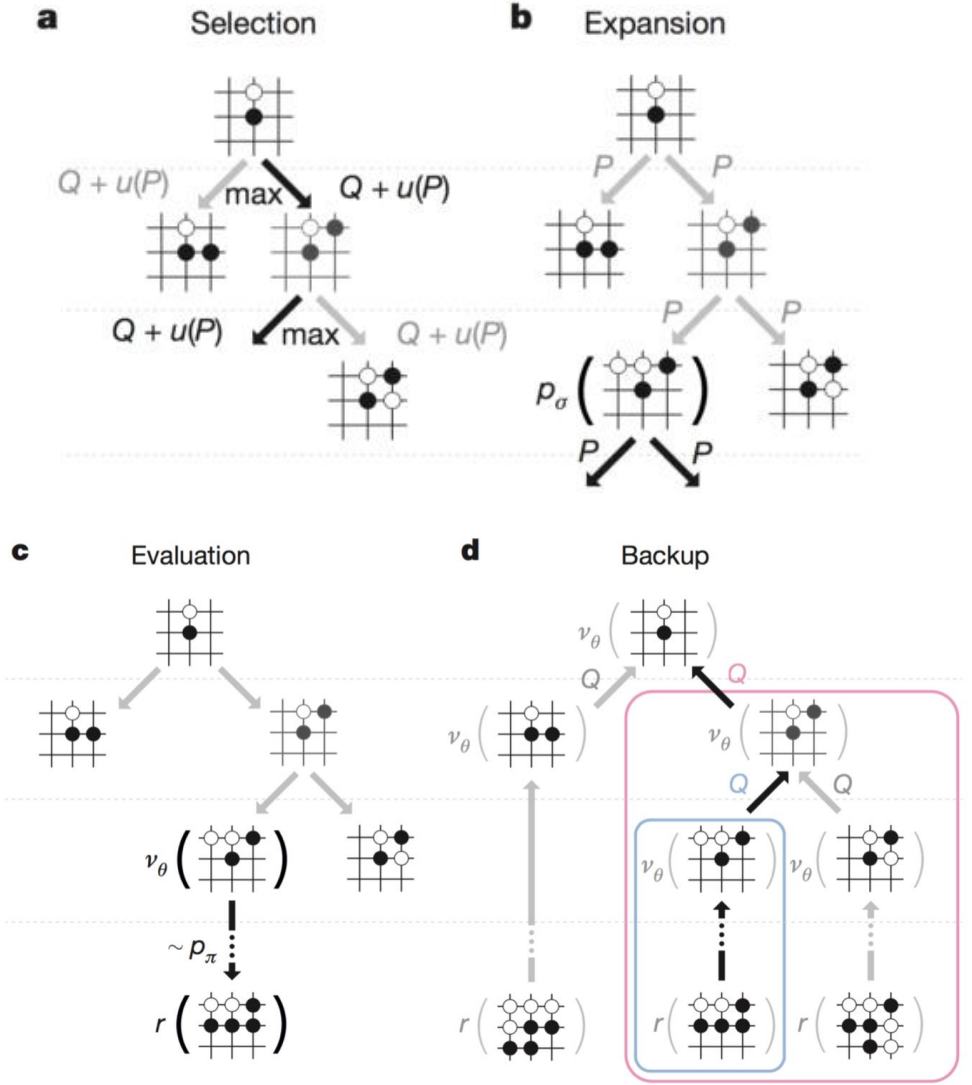
$$= \underbrace{\int \log\left(\frac{p(x,z)}{q_\phi(z|x)}\right) q_\phi(z|x) dz}_{ELBO(\phi)} + \underbrace{\int \log\left(\frac{q_\phi(z|x)}{p(z|x)}\right) q_\phi(z|x) dz}_{KL\left(q_\phi(z|x) \,\|\, p(z|x)\right)}$$



Appendix

❖ 몬테카를로 트리 탐색





Appendix

- LoRA:
  - ✓ For GPT-3 (175B), VRAM can be reduced from 1.2TB to 350GB.
  - ✓ LoRA is applied to attention weight matrixes such as W_q, W_k, W_v, W_o, not MLP module.

$$h = W_0 x + \Delta W x = W_0 x + BAx$$