Engineering Point Defects in Transition Metal Dichalcogenides for Tailored Material Properties using LLMs

Aziz Al-Maeeni[3], Denis Derkach[3], and Andrey Ustyuzhanin[2,1]

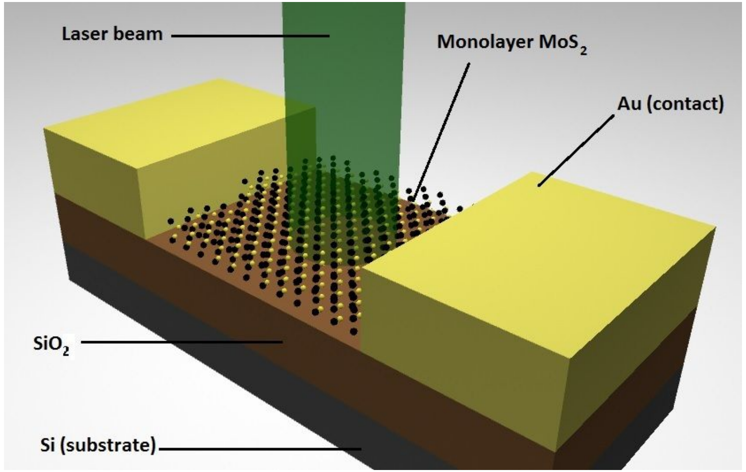
[1]Institute for Functional Intelligent Materials, National University of Singapore, 4 Science Drive 2, Singapore 117544

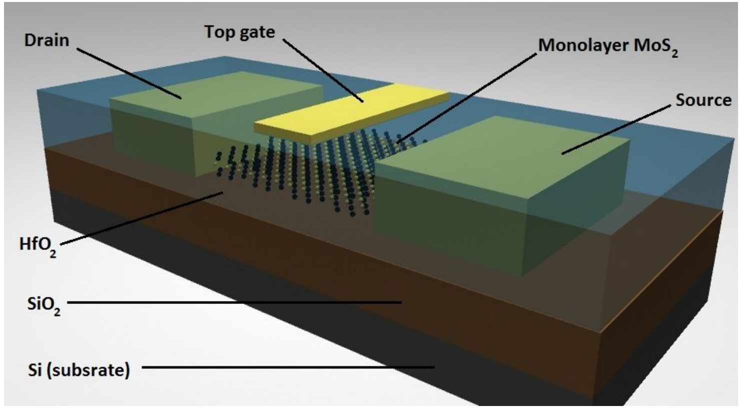
[2]Constructor University Bremen, Campus Ring 1, Bremen, 28759, Germany

[3]HSE University, Myasnitskaya Ulitsa, 20, Moscow, Russia, 101000

Agenda

- What are TMDCs?
- Why do we care about defects?
- Objective
- Data
- Pipeline
- Generation
- Evaluation
- Results
- Future work
- Questions

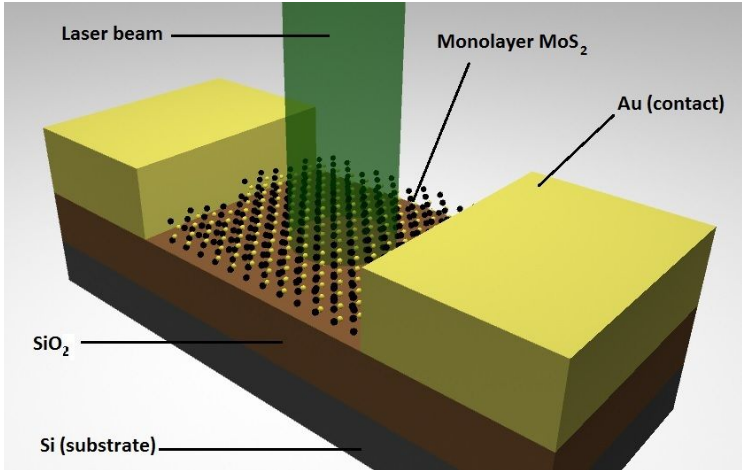


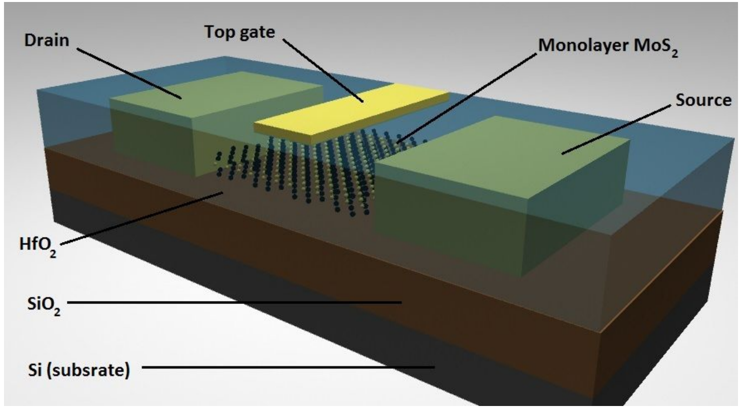


DLCP 2024

Agenda

- What are TMDCs?
- Why do we care about defects?
- Objective
- Data
- Pipeline
- Generation
- Evaluation
- Results
- Questions

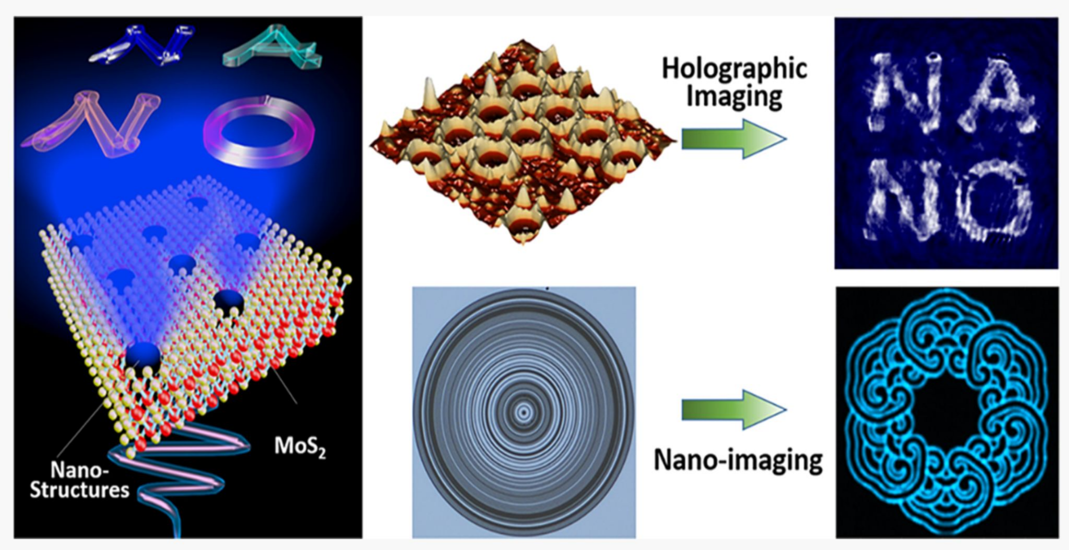




DLCP 2024

Transition Metal Dichalcogenides (TMDCs)

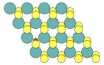
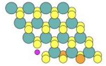
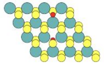
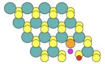
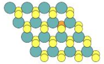
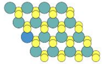
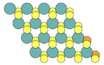
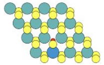
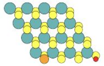
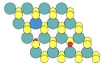
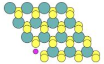
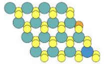
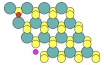
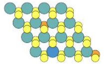
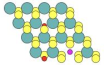
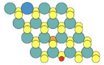
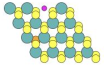
- TMDCs are a class of materials composed of transition metals (like Mo, W) and chalcogens (S, Se, Te).
- They have a layered structure, similar to graphene, with strong in-plane bonds and weak van der Waals forces between layers.
- They exhibit a range of electronic properties, from metallic to semiconducting.
- TMDCs have strong light-matter interactions, making them ideal for optoelectronic applications.
- High flexibility and strength, useful for flexible electronics.

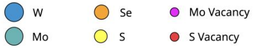


Defects

- **Electronic Properties**: Can introduce localized states in the bandgap, affecting conductivity and carrier mobility.
- **Optical Properties**: Influence photoluminescence and absorption spectra, crucial for optoelectronic applications.
- **Mechanical Properties**: Affect the strength and flexibility of the material.

Material	Substitutions	Vacancies
MoS_2	S $\rightarrow$ Se; Mo $\rightarrow$ W	Mo; S
WSe_2	Se $\rightarrow$ S; W $\rightarrow$ Mo	W; Se
h-BN	B $\rightarrow$ C; N $\rightarrow$ C	B; N
GaSe	Ga $\rightarrow$ Se; Se $\rightarrow$ S	Ga; Se
InSe	In $\rightarrow$ Ga; Se $\rightarrow$ S	In; Se
BP	P $\rightarrow$ N	P

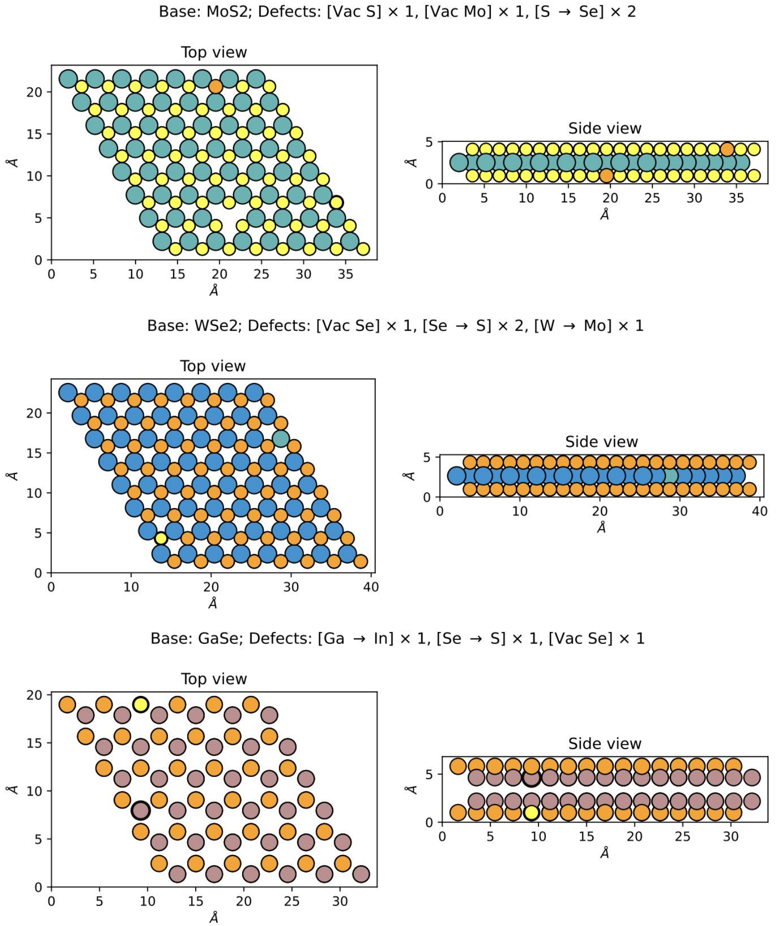
Mo	S	S	Num	Example	Mo	S	S	Num	Example
	vac		1		vac	Se	Se	743	
	vac	vac	19		vac	vac	Se	1415	
	Se		1		W			1	
	Se	Se	19		W	vac		15	
	vac	Se	29		W	vac	vac	743	
vac			1		W	Se		15	
vac	vac		15		W	Se	Se	743	
vac	vac	vac	743		W	vac	Se	1415	
vac	Se		15						



Defect types in low density dataset. The ‘Mo’ and two ‘S’ columns denote the the type of site that is being perturbed either by substituting the listed element, or a vacancy (vac). ‘Num’ column contains the number of structures with defects of the type in the dataset. Finally, ‘Example’ column presents a structure with such defect.

2DMD A 2D Material Defect Dataset

- Established the first comprehensive 2D material defect database (2DMD) for machine learning training and evaluation.
- The dataset enable training models for structure property predictions and various other tasks.
- Includes widely used 2D materials: MoS_2, WSe_2, h-BN, GaSe, InSe, and BP.
- Two parts: low defect concentration (structured configurations) and high defect concentration (random configurations).
- Low defect concentration: 5933 MoS_2 structures and 5933 WSe_2 structures in 8x8 supercell.
- High defect concentration: randomly generated substitution and vacancy defects.
- Dataset contains 14866 structures with 120-192 atoms each.

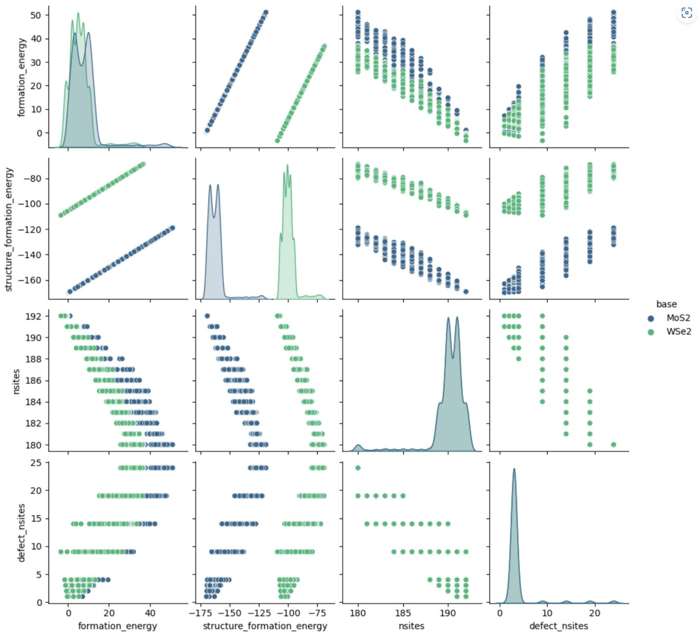


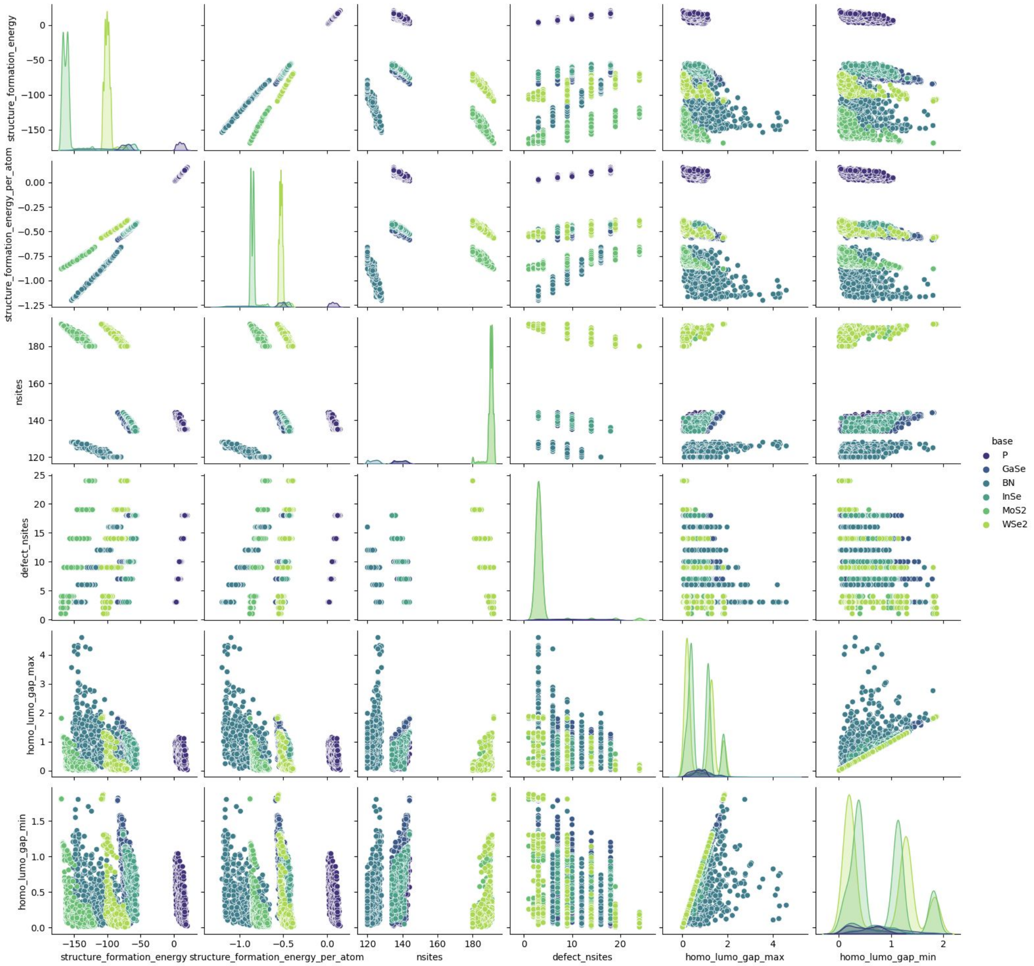
Example structures from [2DMD dataset]

Huang, P., Al-Maeeni et al. Unveiling the complex structure-property correlation of defects in 2D materials based on high throughput datasets. *npj 2D Mater Appl* 7, 6 (2023). https://doi.org/10.1038/s41699-023-00369-1

DLCP 2024

Pair plot of the properties and defects in the dataset for P, GaSe, BN, InSe, MoS2, and WSe2



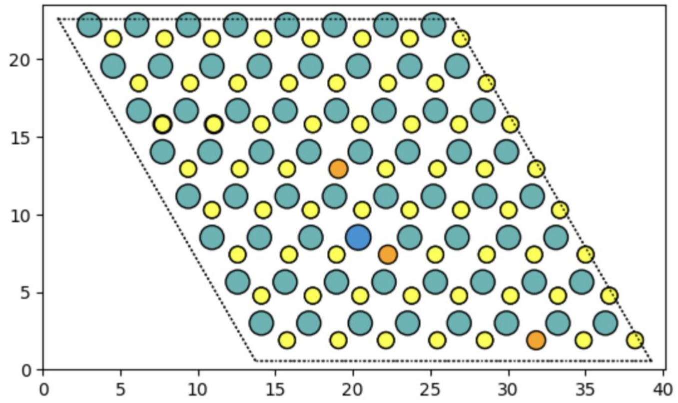


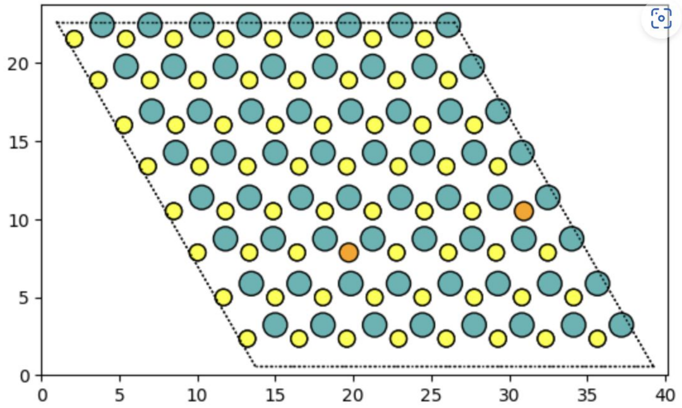
The objective

Tailored defects design in TMDCs is crucial for enhancing their functional properties in various applications. Precise engineered defects enable modulating the electronic, optical, and catalytic characteristics of TMDCs, leading to improved performance in areas such as transistors, sensors, and energy storage devices.

- Inputs:
  - Conditions:
    - Energy,
    - Homo
    - Lumo
    - Bandgap
    - Formation energy
    - Other Physical attributes

- Objective:
  - Generate the crystal with the defects satisfying the properties given



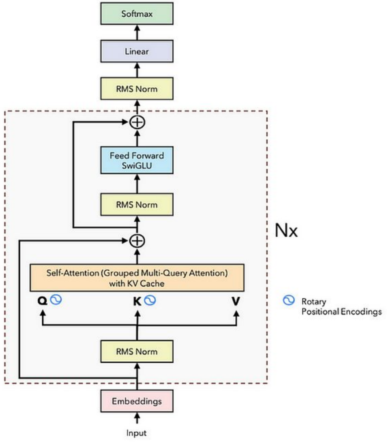


An example of different defects configurations

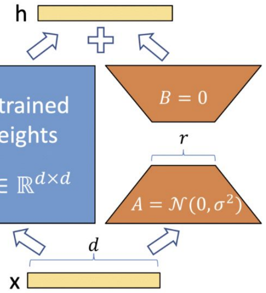
Llama 3

- Large Language Model Meta AI
- Pretrained on over 15T tokens
- Instruction fine-tuning
- **Decoder-only** transformer architecture
- Uses a tokenizer with a vocabulary of 128K tokens that encodes language efficiently
- Tokenizes all numbers into **single digits**
- State-of-the-art performance on NLP tasks

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

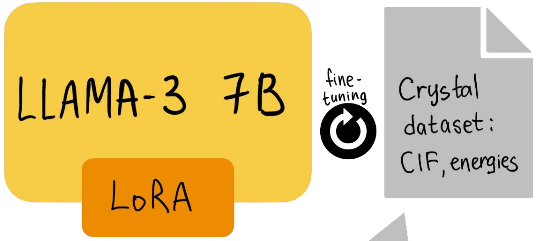


LLaMA



Approach

- Construct an instruction dataset from the 2DMD crystals descriptions:
  - Defect substitution rules
  - Properties
  - Instruction for generation:
    - Types of defects
    - Restriction
- Use a parameter efficient fine tuning method – Low Rank Adaptation of Large Language Models (LoRA).
- Fine tune LLAMA-3 on part of the Crystallographic Information File (CIF).
- Apply random rotations
- Randomly sample physical properties
- Run inference:
  - Instruction + masked coordinates
- Parse the generated coordinates
- Calculate the energies of the generated structure via a surrogate model



Here is a TMDC material (BN), unitcell of size [8, 8, 1].
The defects for this material BN are generated based on following rules:
{'type': 'substitution', 'from': 'N', 'to': 'C'}
{'type': 'substitution', 'from': 'B', 'to': 'C'}
{'type': 'vacancy', 'element': 'N'}.

The material has the following properties:
- The formation energy per atom is 6.1508.
- The energy per atom is -8.6699.
- The Fermi level is -4.1761.

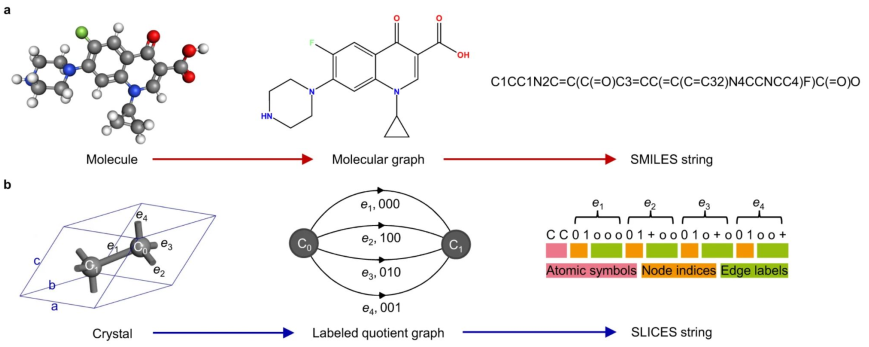
Generate defects in the crystal structure, of the following types (vacancy, substitutional).
You are only allowed to change the atom symbol (in the case of vacancy replace it with VAC)
without changing anything else.

Crystal structure:

<Crystal elements and coords>

Simplified line-input crystal-encoding system (SLICES)

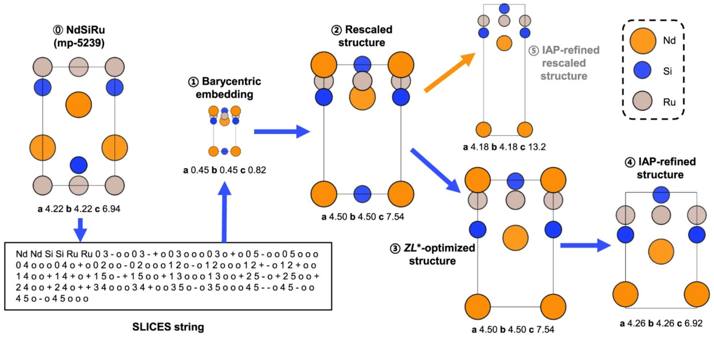
- Based on Quotient graphs to handle periodic crystals
- And Eon's method for reconstruction the structures: [Euclidian embeddings of periodic nets: definition of a topologically induced complete set of geometric descriptors for crystal structures - PubMed (nih.gov)]



a The molecular graph serves as an intermediary to translate between molecules and SMILES strings. **b** Likewise, the labeled quotient graph serves as an intermediary to translate between crystal structures and SLICES strings.

Fig. 2: Intermediate structures generated during reconstructing the crystal structure of NdSiRu (mp-5239) from its SLICES string.

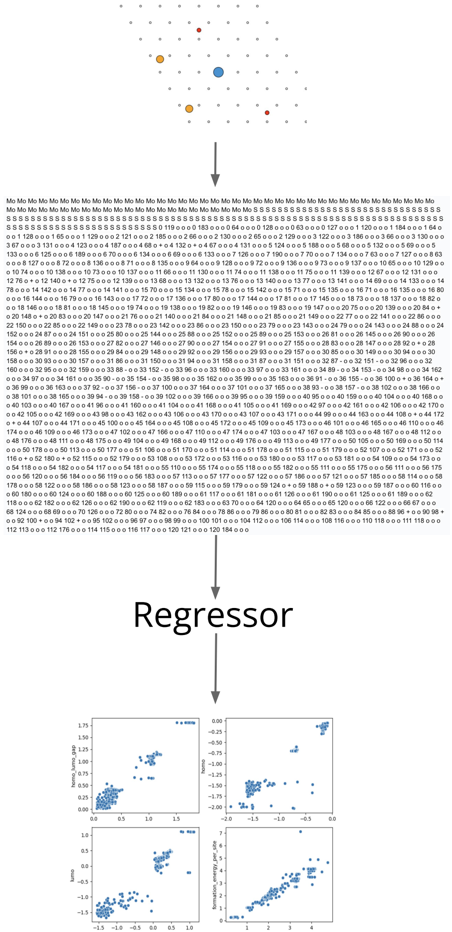
From: [An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning]



[An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning | Nature Communications]

Decision Trees + SLICES

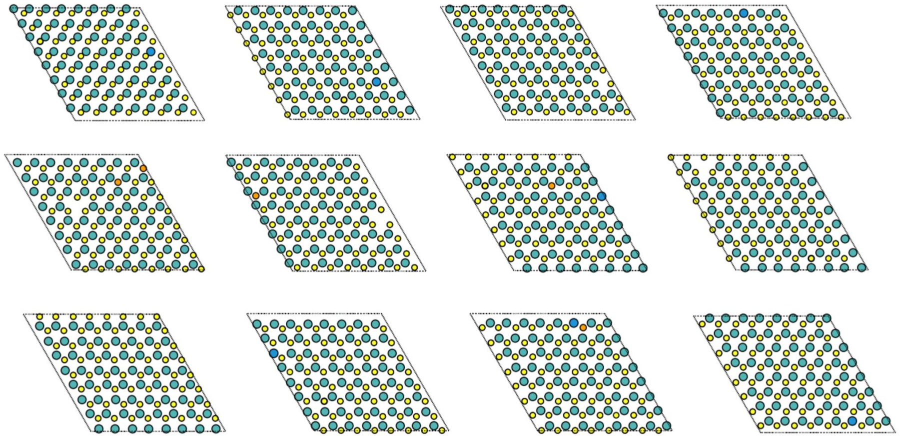
- [Catboost](#) has been used an excellent implementation of gradient boosted decision trees.
- Property predictions are done using Catboost to predict the energies from the SLICES crystal line strings.



Results

- Inference/generation is around ~4-2 crystals per second
- 94% of the generated structures are valid cif files

```
Processed prompts:  96%|██████████| 3577/3717 [1:19:45<00:30,  4.62it/s, est. speed input: 152.65 toks/s, output: 1935.73 toks/s]
Processed prompts:  96%|██████████| 3578/3717 [1:19:45<00:31,  4.41it/s, est. speed input: 152.68 toks/s, output: 1936.16 toks/s]
Processed prompts:  96%|██████████| 3581/3717 [1:19:47<00:49,  2.75it/s, est. speed input: 152.74 toks/s, output: 1936.90 toks/s]
Processed prompts:  96%|██████████| 3582/3717 [1:19:48<00:55,  2.42it/s, est. speed input: 152.76 toks/s, output: 1937.14 toks/s]
Processed prompts:  96%|██████████| 3583/3717 [1:20:13<08:16,  3.70s/it, est. speed input: 152.00 toks/s, output: 1927.52 toks/s]
Processed prompts:  96%|██████████| 3584/3717 [1:20:23<10:17,  4.64s/it, est. speed input: 151.74 toks/s, output: 1924.23 toks/s]
Processed prompts:  96%|██████████| 3585/3717 [1:20:26<09:32,  4.34s/it, est. speed input: 151.68 toks/s, output: 1923.60 toks/s]
Processed prompts:  96%|██████████| 3586/3717 [1:20:30<09:29,  4.34s/it, est. speed input: 151.59 toks/s, output: 1922.41 toks/s]
Processed prompts:  97%|██████████| 3587/3717 [1:20:52<17:52,  8.25s/it, est. speed input: 150.97 toks/s, output: 1914.68 toks/s]
Processed prompts:  97%|██████████| 3588/3717 [1:21:21<29:02, 13.50s/it, est. speed input: 150.10 toks/s, output: 1903.66 toks/s]
Processed prompts:  97%|██████████| 3589/3717 [1:21:22<22:02, 10.33s/it, est. speed input: 150.10 toks/s, output: 1903.69 toks/s]
Processed prompts:  97%|██████████| 3592/3717 [1:21:24<10:36,  5.10s/it, est. speed input: 150.19 toks/s, output: 1904.85 toks/s]
Processed prompts:  97%|██████████| 3597/3717 [1:21:25<04:42,  2.36s/it, est. speed input: 150.37 toks/s, output: 1907.24 toks/s]
Processed prompts:  97%|██████████| 3598/3717 [1:21:25<04:07,  2.08s/it, est. speed input: 150.40 toks/s, output: 1907.65 toks/s]
Processed prompts:  97%|██████████| 3604/3717 [1:21:26<02:00,  1.06s/it, est. speed input: 150.62 toks/s, output: 1910.49 toks/s]
Processed prompts:  97%|██████████| 3607/3717 [1:21:27<01:36,  1.14it/s, est. speed input: 150.71 toks/s, output: 1911.68 toks/s]
Processed prompts:  97%|██████████| 3623/3717 [1:21:28<00:30,  3.06it/s, est. speed input: 151.37 toks/s, output: 1920.06 toks/s]
Processed prompts:  98%|██████████| 3631/3717 [1:21:29<00:22,  3.83it/s, est. speed input: 151.69 toks/s, output: 1924.07 toks/s]
Processed prompts:  98%|██████████| 3639/3717 [1:21:30<00:16,  4.64it/s, est. speed input: 152.00 toks/s, output: 1928.10 toks/s]
Processed prompts:  98%|██████████| 3648/3717 [1:21:32<00:13,  5.00it/s, est. speed input: 152.34 toks/s, output: 1932.44 toks/s]
Processed prompts:  98%|██████████| 3649/3717 [1:21:32<00:13,  4.91it/s, est. speed input: 152.37 toks/s, output: 1932.89 toks/s]
Processed prompts:  98%|██████████| 3651/3717 [1:21:33<00:15,  4.27it/s, est. speed input: 152.43 toks/s, output: 1933.63 toks/s]
Processed prompts:  98%|██████████| 3652/3717 [1:22:01<02:44,  2.53s/it, est. speed input: 151.60 toks/s, output: 1923.25 toks/s]
Processed prompts:  98%|██████████| 3653/3717 [1:22:09<03:20,  3.14s/it, est. speed input: 151.39 toks/s, output: 1920.57 toks/s]
Processed prompts:  98%|██████████| 3655/3717 [1:22:10<02:35,  2.50s/it, est. speed input: 151.45 toks/s, output: 1921.33 toks/s]
Processed prompts:  98%|██████████| 3657/3717 [1:22:11<01:57,  1.95s/it, est. speed input: 151.51 toks/s, output: 1922.20 toks/s]
Processed prompts:  98%|██████████| 3658/3717 [1:22:11<01:39,  1.69s/it, est. speed input: 151.55 toks/s, output: 1922.65 toks/s]
Processed prompts:  98%|██████████| 3659/3717 [1:22:11<01:26,  1.49s/it, est. speed input: 151.57 toks/s, output: 1922.98 toks/s]
Processed prompts:  98%|██████████| 3660/3717 [1:22:12<01:10,  1.24s/it, est. speed input: 151.61 toks/s, output: 1923.43 toks/s]
Processed prompts:  99%|██████████| 3667/3717 [1:22:12<00:24,  2.08it/s, est. speed input: 151.88 toks/s, output: 1926.94 toks/s]
Processed prompts:  99%|██████████| 3669/3717 [1:22:13<00:22,  2.18it/s, est. speed input: 151.94 toks/s, output: 1927.74 toks/s]
Processed prompts:  99%|██████████| 3677/3717 [1:22:14<00:10,  3.93it/s, est. speed input: 152.25 toks/s, output: 1931.81 toks/s]
Processed prompts:  99%|██████████| 3684/3717 [1:22:14<00:05,  5.57it/s, est. speed input: 152.53 toks/s, output: 1935.43 toks/s]
Processed prompts:  99%|██████████| 3685/3717 [1:22:14<00:05,  5.55it/s, est. speed input: 152.57 toks/s, output: 1935.90 toks/s]
Processed prompts:  99%|██████████| 3687/3717 [1:22:15<00:05,  5.37it/s, est. speed input: 152.63 toks/s, output: 1936.82 toks/s]
Processed prompts:  99%|██████████| 3688/3717 [1:22:15<00:05,  5.40it/s, est. speed input: 152.67 toks/s, output: 1937.30 toks/s]
Processed prompts:  99%|██████████| 3692/3717 [1:22:16<00:04,  5.89it/s, est. speed input: 152.82 toks/s, output: 1939.25 toks/s]
Processed prompts: 100%|██████████| 3699/3717 [1:22:17<00:02,  6.21it/s, est. speed input: 153.09 toks/s, output: 1942.65 toks/s]
Processed prompts: 100%|██████████| 3701/3717 [1:22:18<00:04,  3.77it/s, est. speed input: 153.12 toks/s, output: 1943.12 toks/s]
Processed prompts: 100%|██████████| 3702/3717 [1:22:19<00:04,  3.44it/s, est. speed input: 153.15 toks/s, output: 1943.46 toks/s]
Processed prompts: 100%|██████████| 3704/3717 [1:22:19<00:03,  3.57it/s, est. speed input: 153.22 toks/s, output: 1944.35 toks/s]
Processed prompts: 100%|██████████| 3706/3717 [1:22:20<00:02,  3.69it/s, est. speed input: 153.28 toks/s, output: 1945.25 toks/s]
Processed prompts: 100%|██████████| 3710/3717 [1:22:20<00:01,  4.84it/s, est. speed input: 153.44 toks/s, output: 1947.24 toks/s]
Processed prompts: 100%|██████████| 3715/3717 [1:22:21<00:00,  6.94it/s, est. speed input: 153.64 toks/s, output: 1949.84 toks/s]
Processed prompts: 100%|██████████| 3716/3717 [1:22:21<00:00,  7.03it/s, est. speed input: 153.68 toks/s, output: 1950.33 toks/s]
Processed prompts: 100%|██████████| 3717/3717 [1:22:21<00:00,  1.33s/it, est. speed input: 153.72 toks/s, output: 1950.88 toks/s]
```



Conditional on formation energy, bandgap, homo, lumo generation samples

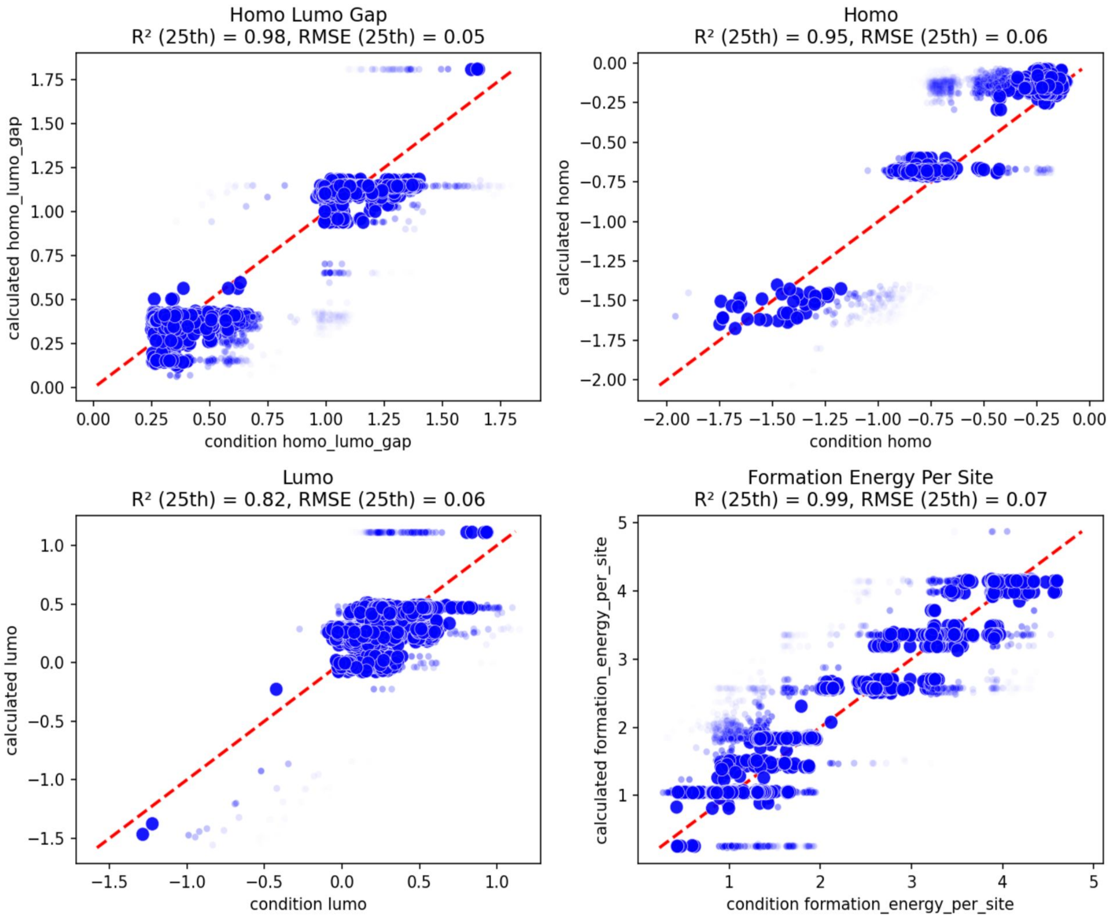
DLCP 2024

Results

- Plot showing 6000 generated MoS2 crystals containing different defect concentrations evaluated with surrogate model.
- Structures with high MSE from the target property have reduced opacity and size.
- R^2 and RMSE are calculated for the 25% lowest MSE values of the generated structures

$$R^2_{25\text{th}} = 1 - \frac{\sum_{i \in \{i | (y_i - \hat{y}_i)^2 \leq Q_{0.25}\}} (y_i - \hat{y}_i)^2}{\sum_{i \in \{i | (y_i - \hat{y}_i)^2 \leq Q_{0.25}\}} (y_i - \bar{y}_{25\text{th}})^2}$$

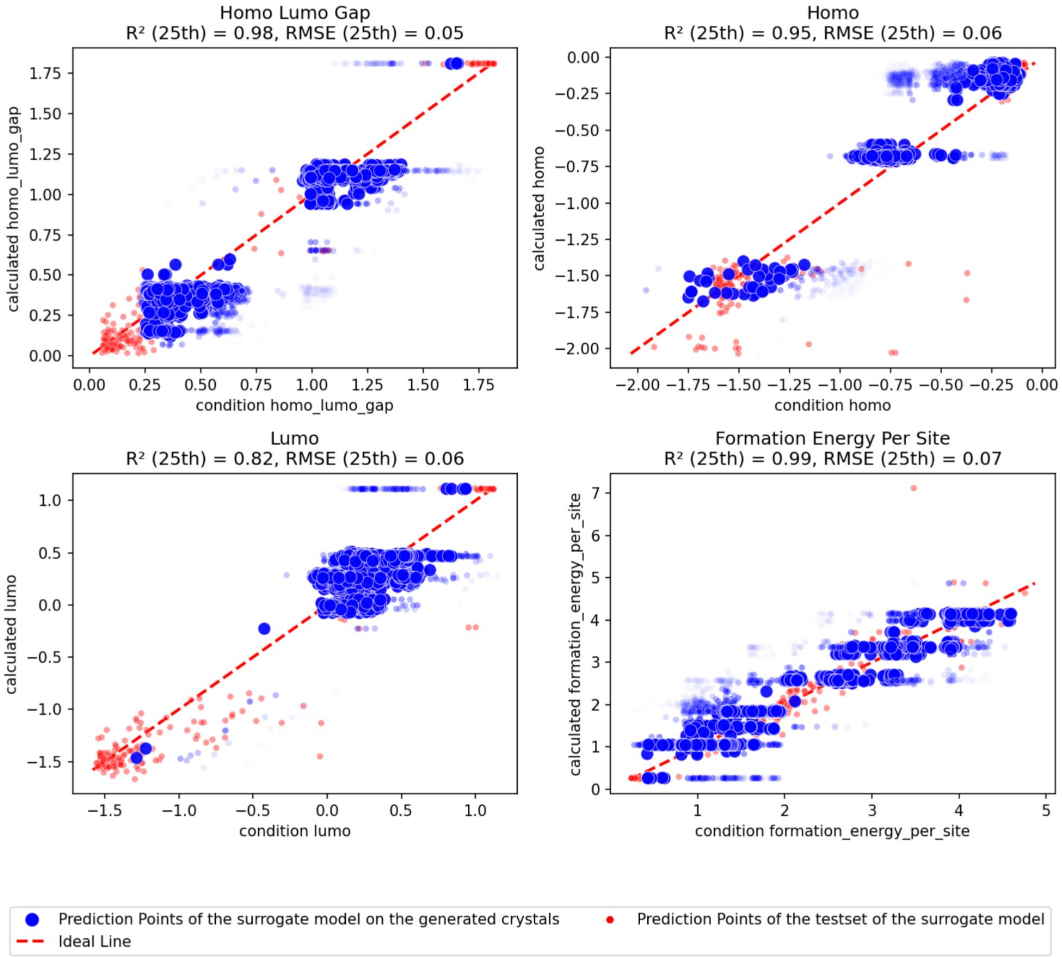
$$\text{RMSE}_{25\text{th}} = \sqrt{\frac{1}{|\{i \mid (y_i - \hat{y}_i)^2 \leq Q_{0.25}\}|} \sum_{i \in \{i | (y_i - \hat{y}_i)^2 \leq Q_{0.25}\}} (y_i - \hat{y}_i)^2}$$



DLCP 2024

Results

- Plot showing 6000 generated MoS2 crystals containing different defect concentrations evaluated with surrogate model, structures with high MSE from the target property have reduced opacity and size.
- R^2 and RMSE are calculated for the 25% lowest MSE values of the generated structures.
- The red points represent the testset of the surrogate model.



Conclusion

- Fine tuning pretrained LLMs are very promising for dealing with crystals structures
- Achieve high quality generation (being valid CIF), only ~6% of the generated crystals are invalid
- Generation/inference (can be fast ~1800 toks/s if optimized properly with:
  - Efficient management of attention key and value memory with PagedAttention
  - Continuous batching of incoming requests
  - Fast model execution with CUDA/HIP graph and
  - Optimized CUDA kernels
- The LLMs learn the physical biases making it useful in generating proposals for Bayesian optimization settings or High throughput screening pipelines.
- The surrogate model has a low fidelity prompting a better model to be used with ab initio simulation

Questions

- Why using LoRA instead of full finetune?
- How LLMs can understand or learn such a fundamental physical thing?
- Why didn't you train on SLICES instead of CIF?
- Why llama?
- Your turn!

GitHub
@abdalazizrashid

@AFDEE1C

DLCP 2024

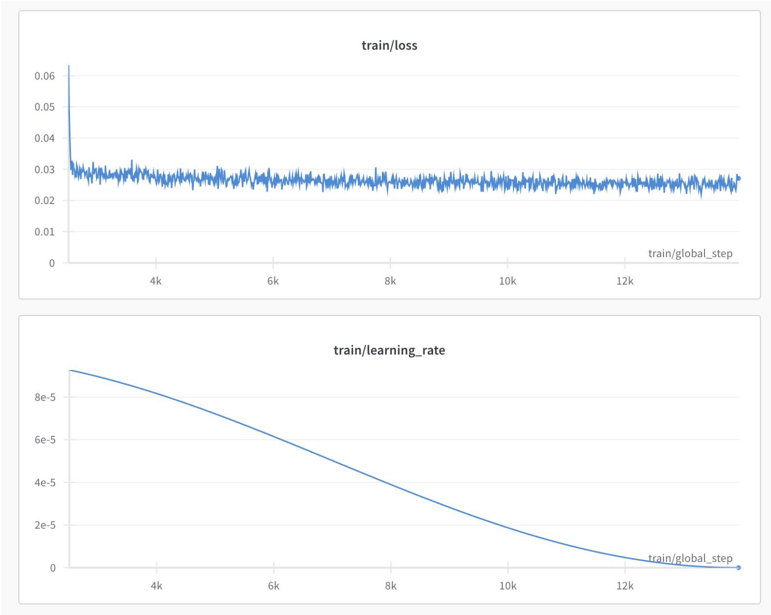
Backup

DLCP 2024

Training Summary Metrics

Training Metrics:

- **Loss**: 0.0271
- **Runtime**: ~20 hours@(128 CPU, 8 NVIDIA A100-SXM4-80GB)
- **Epochs**: 10
- **Gradient Norm**: 0.07@epoch 10
- **Learning Rate**: 0.0001@epoch 0 – 5.15261339e-12@epoch 10
- No quantization



Model & Method	# Trainable Parameters	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	Avg.
RoB$_{\text{base}}$ (FT)*	125.0M	**87.6**	94.8	90.2	**63.6**	92.8	**91.9**	78.7	91.2	86.4
RoB$_{\text{base}}$ (BitFit)*	0.1M	84.7	93.7	**92.7**	62.0	91.8	84.0	81.5	90.8	85.2
RoB$_{\text{base}}$ (Adpt$^{\text{D}}$)*	0.3M	$87.1_{\pm.0}$	$94.2_{\pm.1}$	$88.5_{\pm1.1}$	$60.8_{\pm.4}$	$93.1_{\pm.1}$	$90.2_{\pm.0}$	$71.5_{\pm2.7}$	$89.7_{\pm.3}$	84.4
RoB$_{\text{base}}$ (Adpt$^{\text{D}}$)*	0.9M	$87.3_{\pm.1}$	$94.7_{\pm.3}$	$88.4_{\pm.1}$	$62.6_{\pm.9}$	$93.0_{\pm.2}$	$90.6_{\pm.0}$	$75.9_{\pm2.2}$	$90.3_{\pm.1}$	85.4
RoB$_{\text{base}}$ (LoRA)	0.3M	$87.5_{\pm.3}$	$\mathbf{95.1}_{\pm.2}$	$89.7_{\pm.7}$	$63.4_{\pm1.2}$	$\mathbf{93.3}_{\pm.3}$	$90.8_{\pm.1}$	$\mathbf{86.6}_{\pm.7}$	$\mathbf{91.5}_{\pm.2}$	**87.2**
RoB$_{\text{large}}$ (FT)*	355.0M	90.2	**96.4**	**90.9**	68.0	94.7	**92.2**	86.6	92.4	88.9
RoB$_{\text{large}}$ (LoRA)	0.8M	$\mathbf{90.6}_{\pm.2}$	$96.2_{\pm.5}$	$\mathbf{90.9}_{\pm1.2}$	$\mathbf{68.2}_{\pm1.9}$	$\mathbf{94.9}_{\pm.3}$	$91.6_{\pm.1}$	$\mathbf{87.4}_{\pm2.5}$	$\mathbf{92.6}_{\pm.2}$	**89.0**
RoB$_{\text{large}}$ (Adpt$^{\text{P}}$)†	3.0M	$90.2_{\pm.3}$	$96.1_{\pm.3}$	$90.2_{\pm.7}$	$\mathbf{68.3}_{\pm1.0}$	$\mathbf{94.8}_{\pm.2}$	$\mathbf{91.9}_{\pm.1}$	$83.8_{\pm2.9}$	$92.1_{\pm.7}$	88.4
RoB$_{\text{large}}$ (Adpt$^{\text{P}}$)†	0.8M	$\mathbf{90.5}_{\pm.3}$	$\mathbf{96.6}_{\pm.2}$	$89.7_{\pm1.2}$	$67.8_{\pm2.5}$	$\mathbf{94.8}_{\pm.3}$	$91.7_{\pm.2}$	$80.1_{\pm2.9}$	$91.9_{\pm.4}$	87.9
RoB$_{\text{large}}$ (Adpt$^{\text{H}}$)†	6.0M	$89.9_{\pm.5}$	$96.2_{\pm.3}$	$88.7_{\pm2.9}$	$66.5_{\pm4.4}$	$94.7_{\pm.2}$	$92.1_{\pm.1}$	$83.4_{\pm1.1}$	$91.0_{\pm1.7}$	87.8
RoB$_{\text{large}}$ (Adpt$^{\text{H}}$)†	0.8M	$90.3_{\pm.3}$	$96.3_{\pm.5}$	$87.7_{\pm1.7}$	$66.3_{\pm2.0}$	$94.7_{\pm.2}$	$91.5_{\pm.1}$	$72.9_{\pm2.9}$	$91.5_{\pm.5}$	86.4
RoB$_{\text{large}}$ (LoRA)†	0.8M	$\mathbf{90.6}_{\pm.2}$	$96.2_{\pm.5}$	$\mathbf{90.2}_{\pm1.0}$	$68.2_{\pm1.9}$	$\mathbf{94.8}_{\pm.3}$	$91.6_{\pm.2}$	$\mathbf{85.2}_{\pm1.1}$	$\mathbf{92.3}_{\pm.5}$	**88.6**
DeB$_{\text{XXL}}$ (FT)*	1500.0M	91.8	**97.2**	92.0	72.0	**96.0**	92.7	93.9	92.9	91.1
DeB$_{\text{XXL}}$ (LoRA)	4.7M	$\mathbf{91.9}_{\pm.2}$	$96.9_{\pm.2}$	$\mathbf{92.6}_{\pm.6}$	$\mathbf{72.4}_{\pm1.1}$	$\mathbf{96.0}_{\pm.1}$	$\mathbf{92.9}_{\pm.1}$	$\mathbf{94.9}_{\pm.4}$	$\mathbf{93.0}_{\pm.2}$	**91.3**

Table 2: RoBERTa$_{\text{base}}$, RoBERTa$_{\text{large}}$, and DeBERTa$_{\text{XXL}}$ with different adaptation methods on the GLUE benchmark. We report the overall (matched and mismatched) accuracy for MNLI, Matthew's correlation for CoLA, Pearson correlation for STS-B, and accuracy for other tasks. Higher is better for all metrics. * indicates numbers published in prior works. † indicates runs configured in a setup similar to Houlsby et al. (2019) for a fair comparison.

https://arxiv.org/pdf/2106.09685

Llama-2 vs Llama-3