A APPENDIX

A.1 DFTB SIMULATIONS

The structural analysis and electronic structure of anatase, brookite, and rutile phase TiO_2 nanoparticles (NPs) have been investigated using the DFTB method and molecular dynamics simulations implemented in the DFTB+ code (DFT, 2020). The calculations employ the tiorg-0-1 (Dolgonos et al., 2010) set of Slater-Koster parameters.

The initial structures of anatase, brookite, and rutile phase TiO_2 NPs are illustrated in Fig. 4. All three TiO_2 NP models were derived from a bulk $60 \times 60 \times 60$ supercell. The nanoparticle radius was set to the desired value of 0.9 nm, with only atoms within this sphere considered, while those outside were removed. All simulations were conducted at constant volume conditions.

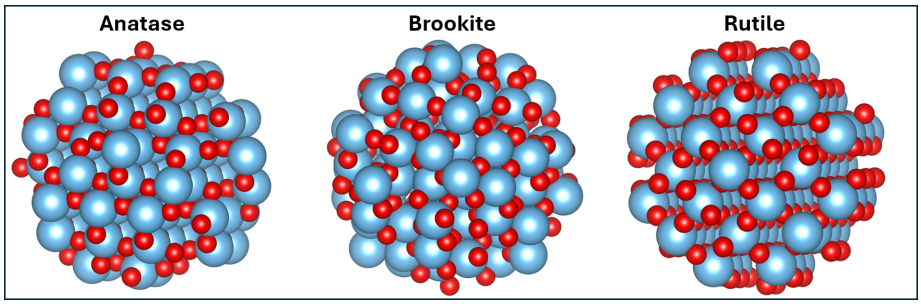

Figure 4: Structural models of anatase, brookite, and rutile phase TiO_2 NPs.

$$H_{DFTB} = H_0 + H_{SCC} + H_{REP} \ , \tag{2}$$

where H_0 represents the non-self-consistent part of the Hamiltonian, H_{SCC} accounts for self-consistent charge corrections, and H_{REP} corresponds to the repulsive potential between atoms. The total energy of the system is then obtained as:

$$E_{DFTB} = \sum_i f_i \epsilon_i + E_{SCC} + E_{REP} \ , \tag{3}$$

where f_i are the occupation numbers, ϵ_i stands for the orbital energies, E_{SCC} denotes the self-consistent charge energy, and E_{REP} represents the repulsive energy.

A.2 CLASSIFICATION RESULTS

Detailed classification results for the experiments are presented in Table 2. The reported values represent the average outcomes from three independent runs.

A.3 PROPERTY PREDICTION EXTENDED RESULTS

The results of the property prediction experiments are expanded with additional STD values in Table 3.

Table 2: Classification results with extended STD values in parenthesis.

Model	Accuracy (STD) (%)		Loss (STD)	
	ID	OOD	ID	OOD
ResNet	65.55 (11.89)	61.76 (17.82)	5.62 (0.5139)	6.59 (0.8840)
SchNet	60.00 (11.55)	66.67 (0.000)	0.82 (0.2960)	0.84 (0.2831)
DimeNet++	66.67 (33.33)	66.67 (33.33)	0.65 (0.4470)	0.66 (0.4453)

Table 3: Extended property prediction MAE results with STD values in parenthesis for ID and OOD settings. Averaged over 3 runs. All values are in eV.

Model	E_H		E_L		E_G	
	ID (STD)	OOD (STD)	ID (STD)	OOD (STD)	ID (STD)	OOD (STD)
ViT	0.2130 (0.0036)	0.2711 (0.0037)	0.2161 (0.0085)	0.2317 (0.0090)	0.3514 (0.0156)	0.3791 (0.0145)
	E_F : 0.2175 (0.0054), 0.2234 (0.0056)			E_T : 0.6620 (0.0253), 0.7047 (0.0321)		
Equiformer	0.3843 (0.0069)	0.3794 (0.0080)	0.1995 (0.0070)	0.2015 (0.0085)	0.6288 (0.0264)	0.6426 (0.0244)
	E_F: 0.5014 (0.0140), 0.5110 (0.0158)			E_T: 0.7651 (0.0344), 0.7340 (0.0272)		
FAENet	0.4843 (0.0097)	0.4967 (0.0124)	0.1670 (0.0063)	0.1755 (0.0072)	0.4825 (0.0217)	0.5087 (0.0254)
	E_F: 0.3268 (0.0098), 0.3294 (0.0089)			E_T: 0.6584 (0.0250), 0.6590 (0.0277)		

A.4 Extended Explainability Results

This subsection presents used prompts and detailed results for the LLM tasks, including BLEU and ROUGE metrics, temperature prediction accuracy, and structural analysis.

Prompts. Same prompt utilized for all the models in order to keep the benchmarking consistent. The prompts are constructed as below:

```
messages = [
    {
        "role": "system",
        "content": """You are a materials science expert specializing in
            ↪ analyzing TiO2 nanoparticles.
Your task is to generate precise captions describing the structural properties
    ↪ of nanoparticles based on both visual and atomic coordinate data.
You should predict both the exact temperature within the given range and the
    ↪ crystal phase (anatase, brookite, or rutile),
and determine the precise rotation applied to the structure if it is not the
    ↪ original configuration."""
    },
    {
        "role": "user",
        "content": [
```

```
            {
                "type": "text",
                "text": """Analyze this TiO2 nanoparticle structure. The
                    ↪ temperature is between 0K and 1000K. This is {"the
                    ↪ original" if is_original else "a rotated"} configuration.

Here is the XYZ structural data:
{xyz_content}

Based on the structural data and image, perform the following tasks:

1. **Predict the crystal phase**: (options: anatase, brookite, rutile)
2. **Predict the exact temperature** within the given range.
3. **Determine the precise rotation angles** if this is a rotated configuration
    ↪ .

Then, generate a caption in the following exact format (replace the
    ↪ placeholders with your predictions):

"This [predicted_phase] configuration at [predicted_temperature]K consists of [
    ↪ total_atoms] atoms, including [ti_atoms] titanium atoms and [o_atoms]
    ↪ oxygen atoms, resulting in a Ti:O ratio of approximately [ratio]:1. The
    ↪ nanoparticle spans about [x_dimension]    in x, [y_dimension]    in y,
    ↪ and [z_dimension]    in z. [Original/Rotation Information]"

**Notes:**

- For rotated configurations, replace `[Original/Rotation Information]` with:
  "Rotation applied: x=[x_angle]  , y=[x_angle]  , z=[z_angle]  ."
- For original configurations, replace it with:
  "This is the original configuration (no rotation)."

**Example Output:**
"This anatase configuration at 350K consists of 100 atoms, including 30
    ↪ titanium atoms and 70 oxygen atoms, resulting in a Ti:O ratio of
    ↪ approximately 0.43:1. The nanoparticle spans about 5.0    in x, 3.0    in
    ↪  y, and 2.0    in z. This is the original configuration (no rotation)."

**Important:** Only output the caption as specified above without any
    ↪ additional text or explanations."""
            },
            {
                "type": "image_url",
                "image_url": {
                    "url": "{image_b64}",
                    "detail": "low"
                }
            }
        ]
    }
]
```

Text Similarity Performance. Claude-3 models (Sonnet and Opus) outperform GPT models in text similarity metrics, as shown in Table 7, achieving BLEU scores around 0.440.45 and ROUGE-L scores above 0.70. GPT-4 underperforms, with BLEU scores around 0.15 and ROUGE-L scores near 0.24. Performance

remains stable between ID and OOD scenarios for most models, except GPT-4, which experiences slight degradation in OOD cases.

Table 4: Dimension and Phase Prediction Performance

Metric	**Model**	**ID (%)**	**OOD (%)**
Total Atom Count Match (%)	Claude-3-Sonnet	0.00	0.00
	Claude-3-Opus	0.00	0.00
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	0.00	0.00
Phase Prediction (%)	Claude-3-Sonnet	50.00	50.00
	Claude-3-Opus	50.00	50.00
	GPT-3.5-Turbo	45.67	42.08
	GPT-4o	16.33	13.33
Dimension Exact (%)	Claude-3-Sonnet	0.00	0.00
	Claude-3-Opus	0.00	0.00
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	0.00	0.00
Dimension Within 5%	Claude-3-Sonnet	0.00	1.25
	Claude-3-Opus	6.33	11.67
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	1.33	1.67
Dimension Within 10%	Claude-3-Sonnet	2.00	1.25
	Claude-3-Opus	25.67	26.25
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	6.00	4.58
Dimension Within 15%	Claude-3-Sonnet	7.33	8.33
	Claude-3-Opus	41.67	45.00
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	10.67	7.50
Average Dimension Error (%)	Claude-3-Sonnet	39.68	33.27
	Claude-3-Opus	7.16	6.17
	GPT-3.5-Turbo	44.54	44.23
	GPT-4o	8.12	8.49

Table 5: Atom Count Prediction Performance

Metric	Model	ID (%)	OOD (%)
Ti Atoms Count Exact Match (%)	Claude-3-Sonnet	0.33	0.00
	Claude-3-Opus	13.67	15.42
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	0.00	0.00
Ti Atoms Count Within 5%	Claude-3-Sonnet	0.33	0.42
	Claude-3-Opus	14.33	15.83
	GPT-3.5-Turbo	0.67	1.67
	GPT-4o	1.00	1.67
Ti Atoms Count Within 10%	Claude-3-Sonnet	4.00	5.00
	Claude-3-Opus	47.67	47.50
	GPT-3.5-Turbo	0.67	1.67
	GPT-4o	4.33	5.83
Ti Atoms Count Within 15%	Claude-3-Sonnet	4.67	5.83
	Claude-3-Opus	47.67	47.92
	GPT-3.5-Turbo	4.33	4.58
	GPT-4o	9.33	10.00
Average Ti Atoms Count Error (%)	Claude-3-Sonnet	44.96	45.37
	Claude-3-Opus	6.90	6.65
	GPT-3.5-Turbo	34.79	33.53
	GPT-4o	14.75	12.13
O Atoms Exact Count Match (%)	Claude-3-Sonnet	0.00	0.00
	Claude-3-Opus	0.00	0.00
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	0.33	0.00
O Atoms Count Within 5%	Claude-3-Sonnet	0.00	0.00
	Claude-3-Opus	1.33	1.67
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	1.00	1.25
O Atoms Count Within 10%	Claude-3-Sonnet	0.00	0.00
	Claude-3-Opus	1.67	1.67
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	1.67	1.25
O Atoms Count Within 15%	Claude-3-Sonnet	0.00	0.00
	Claude-3-Opus	1.67	1.67
	GPT-3.5-Turbo	0.00	0.00
	GPT-4o	2.00	1.25
Average O Atoms Count Error (%)	Claude-3-Sonnet	64.18	63.88
	Claude-3-Opus	44.28	43.76
	GPT-3.5-Turbo	69.39	68.05
	GPT-4o	34.13	34.23

Table 6: Temperature Prediction Performance

Metric	**Model**	**ID (%)**	**OOD (%)**
Temperature Exact (%)	Claude-3-Sonnet	5.33	0.00
	Claude-3-Opus	2.00	0.00
	GPT-3.5-Turbo	9.33	0.00
	GPT-4o	2.33	0.00
Temperature Within 50K (%)	Claude-3-Sonnet	14.33	2.92
	Claude-3-Opus	7.67	0.42
	GPT-3.5-Turbo	28.00	0.00
	GPT-4o	5.33	0.00
Temperature Within 100K (%)	Claude-3-Sonnet	27.00	4.58
	Claude-3-Opus	15.67	0.42
	GPT-3.5-Turbo	46.00	0.00
	GPT-4o	9.00	0.42
Temperature Within 200K (%)	Claude-3-Sonnet	40.67	15.00
	Claude-3-Opus	37.33	11.25
	GPT-3.5-Turbo	46.00	0.00
	GPT-4o	14.00	0.83
Average Temperature Error (K)	Claude-3-Sonnet	139.67	357.92
	Claude-3-Opus	154.50	294.08
	GPT-3.5-Turbo	61.15	422.22
	GPT-4o	125.51	339.06

Table 7: Text Similarity Metrics

Metric	Model	ID	OOD
BLEU	Claude-3-Sonnet	0.4433	0.4368
	Claude-3-Opus	0.4544	0.4609
	GPT-3.5-Turbo	0.4064	0.3806
	GPT-4o	0.1500	0.1159
ROUGE1	Claude-3-Sonnet	0.7462	0.7452
	Claude-3-Opus	0.7521	0.7670
	GPT-3.5-Turbo	0.7103	0.6915
	GPT-4o	0.2519	0.2030
ROUGE2	Claude-3-Sonnet	0.5018	0.4989
	Claude-3-Opus	0.5237	0.5385
	GPT-3.5-Turbo	0.4635	0.4470
	GPT-4o	0.1808	0.1441
ROUGEL	Claude-3-Sonnet	0.7104	0.7095
	Claude-3-Opus	0.7218	0.7329
	GPT-3.5-Turbo	0.6649	0.6476
	GPT-4o	0.2445	0.1960