TDCM25: A MULTI-MODAL MULTI-TASK BENCHMARK FOR TEMPERATURE-DEPENDENT CRYSTALLINE MATERIALS

Can Polat[1], **Hasan Kurban**[2]*, **Erchin Serpedin**[1], **and Mustafa Kurban**[3]*

[1]Dept. of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA

[2]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[3]Dept. of Prosthetics and Orthotics, Ankara University, Ankara, Turkey

{can.polat, eserpedin}@tamu.edu, hkurban@hbku.edu.qa, kurbanm@ankara.edu.tr

ABSTRACT

Materials exhibit phase and temperature dependent properties that are critical for applications ranging from catalysis to energy storage and environmental remediation and accurate modeling of these dependencies requires high-quality, multi-modal datasets. In this work, TDCM25 (Temperature Dependent Crystalline Materials 2025) is introduced as a comprehensive dataset featuring approximately 100,000 entries spanning three crystalline phases of TiO_2 (anatase, brookite, and rutile) sampled over 21 temperatures from 0K to 1000K. Each entry comprises 3D atomic coordinates, corresponding RGB molecular images, and detailed textual metadata including Ti:O ratios, temperature, spatial dimensions, and transformation parameters. TDCM25 provides a benchmark for developing and evaluating machine learning methods that integrate multi-modal data to capture temperature dependent material behavior. The dataset is publicly available at https://github.com/KurbanIntelligenceLab/TDCM25.

1 INTRODUCTION

Material behaviors vary significantly with changes in phase and temperature, driven by complex interactions among atomic structures, electronic configurations, and external conditions that determine key characteristics such as bandgap, conductivity, and mechanical stability (Yeomans, 1992; Rashad et al., 2012; Roduner, 2014; Sarkar et al., 2018; Cho et al., 2020).

Titanium dioxide (TiO_2) exemplifies the challenges of modeling temperature and phase dependent behaviors (Zhang et al., 2009; Hanaor & Sorrell, 2011; Kurban et al., 2020). Its three crystalline phases, namely anatase, brookite, and rutile, display distinct physical and chemical properties that evolve with temperature, underpinning its applications in photocatalysis, solar energy conversion, and hydrogen capture (Li et al., 2018; Reinhardt et al., 2020; Zhang & Xu, 2020; Zhang et al., 2021; Kurban et al., 2024). Accurately predicting these properties requires capturing comprehensive structural and contextual information across diverse phases and thermal conditions.

Benchmark datasets have driven significant progress in materials science by enabling standardized evaluation of machine learning models. Datasets such as QM9 (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014),

*Corresponding authors. Dataset is publicly available at https://github.com/KurbanIntelligenceLab/TDCM25.

MD17 (Chmiela et al., 2017), and MatBench (Dunn et al., 2020) have supported tasks in molecular and property prediction, yet many focus on static configurations or limited chemical spaces. This leaves a gap in capturing the dynamic, temperature sensitive, and phase dependent behaviors essential for materials like TiO_2. Moreover, limited rotational and structural diversity in existing datasets hampers the development of models that can effectively learn rotational invariance and phase transitions.

This study addresses these challenges by introducing **TDCM25 (Temperature Dependent Crystalline Materials 2025)**, a comprehensive multi-modal benchmark dataset designed to advance machine learning (ML) in materials science. Simulations were performed using density functional tight binding (DFTB) (Hourahine et al., 2020), balancing computational efficiency with physical accuracy to ensure TDCM25 accurately captures essential temperature- and phase-dependent phenomena. TDCM25 serves as a robust benchmark for classification (phase identification), regression (property prediction), and interpretability (explainability of model decisions). The dataset comprises $99{,}414$ data items spanning the anatase, brookite, and rutile phases of TiO_2, sampled at temperatures from 0 K to 1000 K in 50 K increments. Each data item includes 3D atomic coordinates, corresponding molecular images, and detailed textual metadata capturing phase-specific and temperature-sensitive properties. By integrating structural, visual, and textual data, TDCM25 aims to catalyze AI-driven breakthroughs in materials science and foster structured data sharing within the research community. An overview of the dataset and its associated tasks is presented in Figure 1.

2 Related Work

2.1 DFT and DFTB in Materials Modeling

Density functional theory (DFT) (Hohenberg & Kohn, 1964; Calais, 1993) provides a powerful framework for predicting material properties by solving the many-body Schrödinger equation (Schrödinger, 1926; Atkins & Friedman, 2011) through approximations like the Kohn-Sham formalism (Kohn & Sham, 1965). The total energy of a system in DFT is expressed as:

$$E_{\text{total}}[\rho(\mathbf{r})] = T_s[\rho(\mathbf{r})] + \int v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})\, d\mathbf{r} + E_H[\rho(\mathbf{r})] + E_{\text{xc}}[\rho(\mathbf{r})]. \tag{1}$$

This equation consists of four key terms: the kinetic energy of non-interacting electrons (T_s), the external potential energy (v_{ext}), the Hartree energy (E_H) accounting for classical electrostatic interactions, and the exchange-correlation energy (E_{xc}), which captures quantum mechanical many-body effects. While DFT provides accurate predictions, its computational cost scales as $O(n^3T)$, where n represents the number of electrons or basis functions, and T denotes the number of self-consistent field iterations required for convergence. This scaling makes DFT impractical for large-scale datasets and high-throughput materials screening (Becke, 2014; Ratcliff et al., 2017).

DFTB represents a computationally efficient alternative to DFT by approximating it with parameterized Hamiltonians and overlap matrices. The total energy in DFTB is expressed as $E_{\text{total}} \approx E_0 + \Delta E_{\text{rep}}$, where E_0 represents the band structure energy obtained from the eigenvalues of the effective Hamiltonian, while ΔE_{rep} accounts for short-range repulsive interactions. By leveraging precomputed parameters, DFTB significantly reduces the computational costs while maintaining sufficient accuracy for modeling temperature- and phase-dependent properties. This balance makes it particularly suitable for generating datasets like TDCM25, which require extensive structural and electronic data across varying conditions.

2.2 Crystal Phases and Temperature Dependence

Crystalls exhibit diverse physical and chemical behaviors depending on their phases and temperature conditions (Rohrer, 2001). For instance, TiO_2 exists in three main phases: anatase, brookite, and rutile, each with unique electronic, optical, and structural properties (Murray et al., 1993).

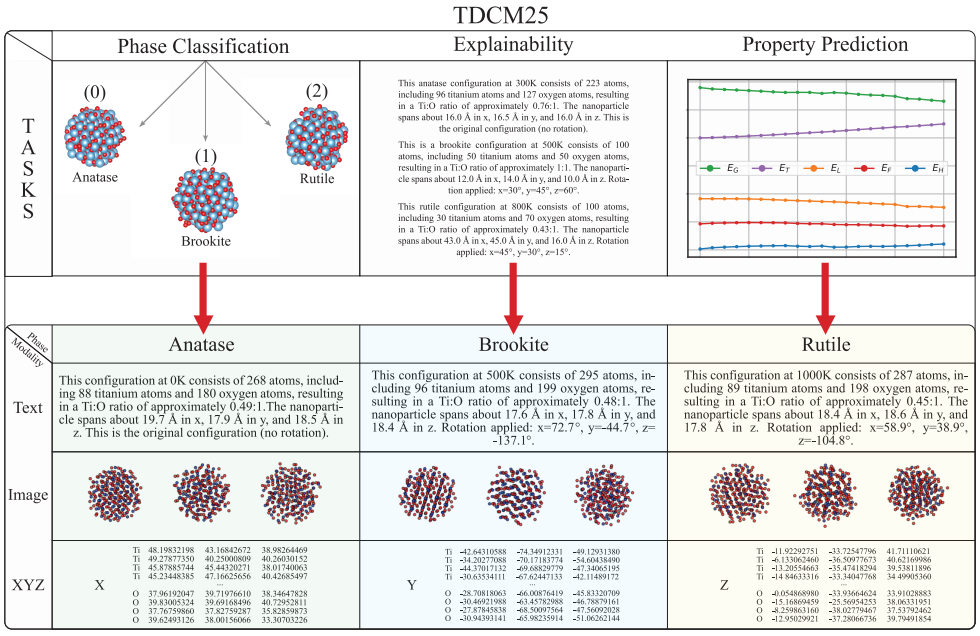

Figure 1: Overview of TDCM25, a multi-modal benchmark for modeling temperature dependent properties in crystalline materials. The figure highlights key tasks: phase classification, property prediction, and explanation generation using large language models, and illustrates the dataset's multi-modal representations (text, images, and 3D coordinates) for TiO_2 in its three phases (anatase, brookite, and rutile) across a range of temperatures.

Anatase is known for its photocatalytic efficiency and high surface reactivity and is widely studied for applications in water splitting and pollutant degradation (Luttrell et al., 2014). *Brookite* displays unique intermediate properties, with potential for niche energy storage applications (Chen & Mao, 2007; Reyes-Coronado et al., 2008). *Rutile* is the thermodynamically stable phase and is commonly used in optical coatings, pigments, and conductive materials due to its density and lower bandgap (Gemming et al., 2010; Buchalska et al., 2015).

Temperature plays a critical role in determining the stability, electronic structure, and phase transitions of these materials (Dubey, 2018). For example, anatase transforms into rutile at elevated temperatures, accompanied by significant changes in its bandgap and charge transport properties (Dette et al., 2014). Capturing these dependencies is essential for modeling real-world applications, making datasets that integrate phase specific and temperature sensitive information invaluable for advancing materials design (Zhang et al., 2013; Hosseini-Sarvari, 2011).

2.3 Datasets in Quantum Chemistry

Benchmark datasets have catalyzed advances in ML for quantum chemistry and materials science by providing structured data for various prediction tasks. For example, QM7 (Blum & Reymond, 2009; Rupp et al., 2012) focuses on small organic molecules with atomization energies as target properties. Datasets such as OC20 (Chanussot et al., 2021) and OC22 (Tran et al., 2023) extend these efforts to catalyst surface interactions, providing geometries, energies, and relaxation trajectories for material-catalyst systems. Similarly, MD22 (Chmiela et al., 2023) includes molecular dynamics trajectories with atomic forces, facilitating the

development of accurate force field models. More recently, PubQChemQC (Kim et al., 2025) offers millions of ground-state molecular structures and electronic properties to support large-scale prediction tasks. Datasets, such as NablaDFT (Khrabrov et al., 2022) and QH9 (Yu et al., 2024) concentrate on Hamiltonian matrix prediction, a core aspect of quantum dynamics, with NablaDFT offering millions of Hamiltonian matrices for various molecular conformers. Despite their significant contributions, many of these datasets lack temperature sensitivity, phase diversity, or multi-modal representations.

In parallel, several benchmarks have been developed to evaluate multi-modal models in scientific contexts. ScienceQA (Lu et al., 2022) and SciBench (Wang et al., 2023) cover topics from elementary to college-level science, while LabBench (Laurent et al., 2024) focuses on figure and table interpretation. More comprehensive benchmarks such as MMMU (Yue et al., 2024) and OlympiadBench (He et al., 2024) extend to research-level content and advanced multi-modal challenges.

Finally, specialized benchmarks like MoleculeNet (Wu et al., 2018) standardize molecular machine learning evaluation by curating datasets, defining metrics, and providing open-source implementations to advance predictive modeling. ChemLit-QA (Wellawatte et al., 2024) provides an expert-validated dataset for evaluating retrieval-augmented generation systems, while HoneyComb (Zhang et al., 2024) enhances materials science reasoning with a curated knowledge base and adaptive tool hub. More recently, MaCBench (Alampara et al., 2024) assesses core competencies in chemistry and materials science, including data extraction and laboratory knowledge.

3 DATASET

The construction of the TDCM25 dataset starts with the electronic structure computations (*e.g.*, using DFTB) and continues with the generation of multi-modal data: 3D coordinates (XYZ files), molecular images, and accompanying textual metadata. Each modality contributes uniquely to model training and broadens applicability.

XYZ files encode precise atomic coordinates, capturing structural configurations, bonding interactions, and phase transitions. These representations enable graph-based neural networks to learn fundamental physical and chemical properties at the atomic level. *Images* provide visual representations of molecular structures, allowing vision-based models to extract structural patterns, morphological variations, and phase-dependent characteristics. This modality is particularly useful for learning spatial and geometric relationships. *Textual descriptions* offer structured metadata, including elemental composition, Ti:O ratio, temperature, and spatial dimensions. These summaries enhance explain ability, support retrieval-augmented generation models, and enable large language models (LLMs) to perform scientific reasoning and explanation tasks.

Through the integration of three distinct data modalities, TDCM25 facilitates comprehensive evaluation across a wide range of model architectures, encompassing graph-based, vision-based, and language-based frameworks. This multimodal approach not only augments predictive performance but also significantly advances AI-driven materials discovery by enabling cross-modal learning in tasks such as classification, regression, and interpretability.

3.1 DFTB SIMULATIONS

Detailed settings and results of in-house DFTB simulations are provided in Appendix A.1 while Figure 2 summarizes the optimized electronic properties for TiO_2 in its three phases: anatase, brookite, and rutile, at 0K, 500K, and 1000K. Specifically, the plots track the evolution of the ground-state energy (E_G), total energy (E_T), LUMO energy (E_L), Fermi energy (E_F), and HOMO energy (E_H), respectively. Additionally, Figure 2(d) presents the maximum atomic displacement, while Figure 2(e) illustrates the volumetric expansion as temperature increases.

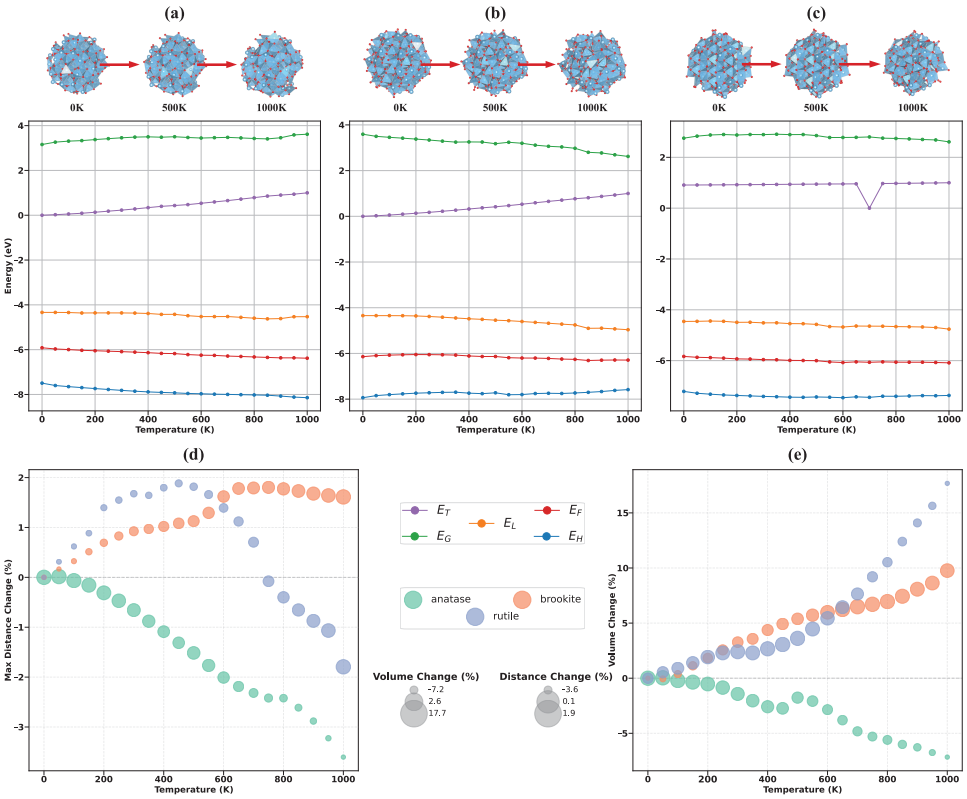

Figure 2: Temperature dependent electronic properties and structural changes of TiO_2 in its anatase, brookite, and rutile phases. The plots track the evolution of E_G, normalized E_T, E_L, E_F, and E_H with temperature. Subplots (d) and (e) show the relative maximum atomic displacement and volumetric expansion, respectively, illustrating the effects of thermal expansion.

3.2 Rotational Diversity and Sampling in 3D Space

To achieve rotational invariance, the dataset incorporates multiple orientations of TiO_2 nanoparticles, systematically sampled from the special orthogonal group $\mathrm{SO}(3)$. A quaternion-based method is employed for uniform sampling, thereby avoiding the clustering issues associated with angle-based parameterizations (Yershova & LaValle, 2004). The total number of orientations, N, is computed using the solid angle $\Omega = 2\pi(1 - \cos(\theta))$, which leads to $N = \frac{4\pi}{\Omega}$. For an angular separation of 5°, this formula yields approximately 526 orientations, providing an optimal balance between rotational diversity and computational efficiency. Smaller angular separations would introduce redundancy and inflate storage requirements, while larger separations could miss important orientations.

3.3 Molecular Images

For each rotated configuration, a high-resolution two-dimensional RGB image is generated using the Matplotlib library (Hunter, 2007). This process is applied across every orientation and temperature for all TiO_2 phases, resulting in a one-to-one correspondence between the 526 orientations and 526 RGB images per configuration.

3.4 TEXTUAL DESCRIPTIONS

In addition to the XYZ files and molecular images, concise human-generated textual descriptions capture key structural properties. Each description details the temperature, total atom count, elemental composition (Ti:O ratio), and approximate nanoparticle dimensions along the Cartesian axes. The quotation below illustrates the format for the original, unrotated configuration at 0 K:

> "This configuration at 0K consists of 268 atoms, including 88 titanium atoms and 180 oxygen atoms, resulting in a Ti:O ratio of approximately 0.49:1. The nanoparticle spans about 19.7 Å in x, 17.9 Å in y, and 18.5 Å in z. This is the original configuration (no rotation)."

When a rotation is applied, the textual description is updated to include the corresponding rotation angles. For example: "Rotation applied: x=170.5, y=13.6, z=66.3." These concise metadata annotations enrich each configuration, thereby enhancing the dataset's utility for multi-modal representation learning in materials science and supporting explainability tasks using LLMs.

3.5 TASKS

To highlight the multi-modal nature of the TDCM25 dataset and evaluate a range of modeling approaches, three core tasks in the dataset are next described: phase classification, property prediction, and explainability.

3.5.1 TASK 1: PHASE CLASSIFICATION

Classification Objective. Classify each TiO_2 nanoparticle into one of three crystalline phases.

Data Splits. Training and Validation: Samples from all three phases within the temperature range of 0K to 800K (excluding 400K to 600K) are used. These samples are randomly partitioned into 80% for training and 20% for validation, ensuring proportional representation of all phases.

In-distribution (ID) Test Set: To assess performance on unseen but in-range data, the temperature range 400K to 600K (inclusive) is reserved as the ID test set. This range is excluded from training and validation to simulate unseen conditions within the overall temperature span.

Out-of-distribution (OOD) Test Set: To evaluate generalization beyond the training range, samples from 800K to 1000K (inclusive) form the OOD test set. These temperatures lie entirely outside the 0K to 800K range used for training and validation.

The splitting strategy enables evaluation of both unseen in-range data and truly out-of-distribution samples, providing insights into the model's robustness and generalization across temperature variations.

Evaluation Metrics. Performance is measured using standard classification metrics, including accuracy, precision, recall, and F1-score.

3.5.2 TASK 2: PROPERTY PREDICTION

Prediction Objective. Predict five key electronic structure properties: E_G, E_T, E_L, E_F, and E_H, using inputs derived from atomic structures, images, textual descriptions, or a combination thereof.

Data Splits. This regression task uses the same dataset splits as defined in Task 1 for both ID and OOD evaluations. The training and validation sets are organized in the same manner, with the difference that the target variables are the corresponding physical property values.

Evaluation metrics. Model performance is evaluated using standard regression metrics, including mean absolute error (MAE) and the standard deviation (STD) of the predictions. These metrics provide insights into both the accuracy and consistency of the predicted electronic structure properties.

3.5.3 TASK 3: EXPLAINABILITY OF MATERIALS

Explainability Objective. Automatically generate human-readable textual descriptions for configurations of TiO_2 nanoparticles across different temperatures and phases using only 2D molecular renders and XYZ data.

Data Splits. For this LLM-based task, the same ID and OOD dataset splits defined in Task 1 are used. Additionally, a fine-tuning stage can be performed on the training and validation sets to further improve LLM performance.

Evaluation Metrics. Model performance is assessed along three dimensions: *Structural Prediction:* Evaluates the ability to capture essential material properties, including atom counts, phase, dimensional accuracy, and atomic ratios. *Temperature Prediction:* Measures accuracy in identifying thermal properties, considering both exact matches and tolerance-based thresholds. *Textual Accuracy:* Quantified using linguistic similarity metrics such as BLEU and ROUGE scores. This comprehensive evaluation framework enables an assessment of both the descriptive quality and the physical accuracy of the generated material explanations.

4 MODEL IMPLEMENTATIONS AND EVALUATION RESULTS

To demonstrate the versatility and challenge of the TDCM25 dataset, a wide range of established models were evaluated across all tasks. This study, exclusively relied on well-known architectures and state-of-the-art (SOTA) methods from the literature during initial exploration, including DTNN (Schütt et al., 2017b), FermiNet (Pfau et al., 2020), SpookyNet (Unke et al., 2021), ForceNet (Hu et al., 2021), PaiNN (Schütt et al., 2021), GNS (Godwin et al., 2021), DeepMoleNet (Liu et al., 2021), PsiFormer (von Glehn et al., 2022), Equiformer-v2 (Liao et al., 2023), Pure2DopeNet (Polat et al., 2024), DeNS (Liao et al., 2024), and QuantumShellNet (Polat et al., 2025).

4.1 CLASSIFICATION RESULTS

For phase classification, three models were evaluated as showcase: ResNet18 (He et al., 2016), SchNet (Schütt et al., 2017a), and DimeNet++ (Gasteiger et al., 2020). ResNet18 was trained on 2D images and adapted for three-class classification, while SchNet and DimeNet++ processed XYZ files using graph neural network architectures, with modifications to their output layers for classification. ResNet18 was implemented using the Transformers library (Wolf, 2019) with pre-trained ImageNet-1k checkpoints. In contrast, SchNet and DimeNet++ were obtained from PyTorch Geometric (Fey & Lenssen, 2019) and trained from scratch. Classification accuracy was reported over three runs, using a subset of 90 out of 526 data points for each configuration.

Table 2 in Appendix A.2 presents the classification results, comparing model performance on both ID and OOD temperature tasks including STD. DimeNet++ achieved the highest accuracy across both ID and OOD datasets. SchNet performed better on OOD data than on ID data, indicating strong generalization capabilities. ResNet18 showed comparable performance to SchNet on ID data but suffered the largest accuracy drop in OOD scenarios. These findings suggest that graph-based models such as SchNet and DimeNet++ generalize more effectively than convolutional neural networks like ResNet-18 for phase classification.

4.2 PREDICTION OF PHYSICAL PROPERTIES

For the regression task, instead of ResNet, a pre-trained ViT (Dosovitskiy, 2020) model (ImageNet-1k via the Transformers library) was used. In addition, SchNet and DimeNet++ were replaced with Equiformer (Liao & Smidt, 2022) and FAENet (Duval et al., 2023), via their official repositories. ViT operated exclusively on

images, while Equiformer and FAENet utilized XYZ files. Unlike the classification task, these models were trained on the full dataset of approximately 100,000 samples.

Table 1 presents the MAE in electron volts (eV) for each architecture across ID and OOD scenarios. An extended analysis with STD values is presented in Appendix A.3. ViT achieved the best performance for predicting E_H, but its performance degraded significantly on OOD data. FAENet produced the lowest errors for E_L predictions and demonstrated strong OOD generalization with minimal accuracy loss. Equiformer yielded mixed results, excelling at some energy levels while struggling with E_G predictions. Notably, all models exhibited the highest MAE in E_T predictions, indicating that E_T remains the most challenging property to predict accurately.

Table 1: Mean absolute error for different models across target properties (E_H, E_L, E_G, E_F, E_T) under ID and OOD settings. Lower values indicate better performance.

Model	E_H		E_L		E_G		E_F		E_T	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
ViT	0.2130	0.2711	0.2161	0.2317	0.3514	0.3791	0.2175	0.2234	0.6620	0.7047
Equiformer	0.3843	0.3794	0.1995	0.2015	0.6288	0.6426	0.5014	0.5110	0.7651	0.7340
FAENet	0.4843	0.4967	0.1670	0.1755	0.4825	0.5087	0.3268	0.3294	0.6584	0.6590

4.3 Explainability Evaluation

SOTA LLMs from multiple providers were evaluated, including OpenAI's GPT-4o and GPT-3.5-Turbo, as well as Anthropic's Claude-3-Sonnet and Claude-3-Opus. The evaluation reveals distinct performance patterns across various metrics. Figure 3 presents a comprehensive performance analysis: Figures 3(a) and (b) illustrate phase and structural prediction performance, respectively; Figure 3(c) shows temperature accuracy for both ID (solid bars) and OOD (hatched bars), with higher percentages indicating better accuracy; and Figure 3(d) compares average temperature errors between ID and OOD cases, where lower values denote improved performance. Detailed numerical values and BLEU/ROUGE scores are provided in Appendix A.4 (Tables 4, 5, 6, and 7) as well as user and system prompts.

Structural Predictions. Results from Figure 3 show that Claude-3 Opus achieves the highest accuracy in structural predictions, particularly in dimensional accuracy (45% within a 15% tolerance) and Ti atom counts (47.92% within a 15% tolerance). However, all models struggle with O atom predictions, with average errors ranging from 34% to 69%. Phase prediction accuracy remains stable at 50% for Claude models across both ID and OOD scenarios, whereas GPT models perform notably worse (13 - 45%). Predictions of the Ti:O ratio remains a challenging task for all models, with average errors between 0.18 and 0.56.

Temperature Prediction. Figure 3 indicates that GPT-3.5-Turbo exhibits strong ID performance in temperature prediction, achieving 46% accuracy within a 200K tolerance and the lowest average error (61.15K). However, its performance deteriorates drastically in OOD scenarios, with accuracy dropping to 0% across all thresholds. In contrast, Claude models maintain more consistent performance across both ID and OOD cases, albeit with higher average errors (139 - 154K for ID and 294 - 357K for OOD).

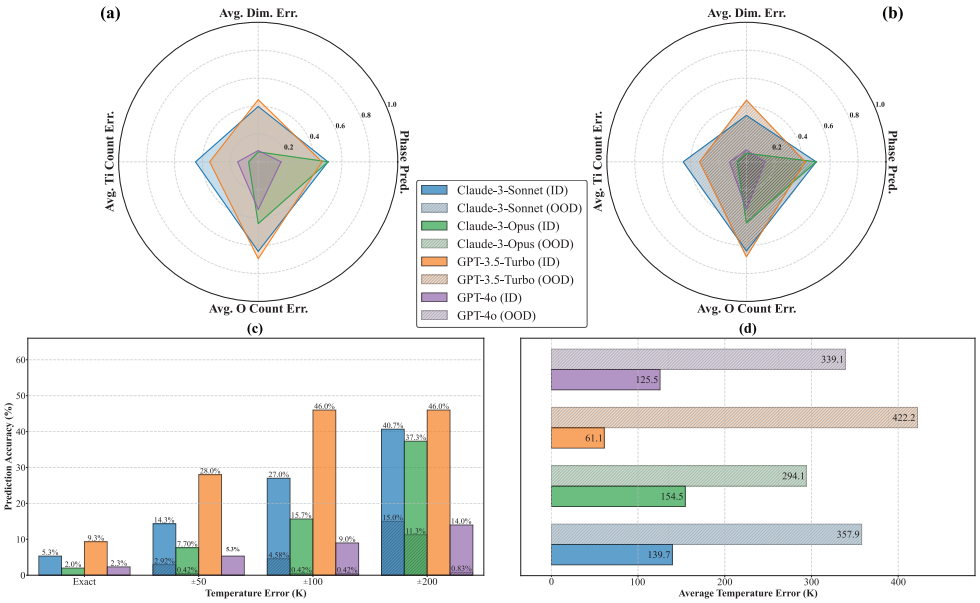

Figure 3: Comprehensive evaluation of LLM performance. (a) and (b) show model performance on ID and OOD data, respectively, reporting phase prediction accuracy, dimensional error, titanium count error, and oxygen count error. (c) Displays temperature prediction accuracy across different temperature ranges for ID (solid bars) and OOD (hatched bars) cases, with higher percentages indicating better accuracy. (d) Compares the average temperature error (in K) between ID and OOD scenarios, where lower values denote improved performance.

5 Discussion and Conclusion

The TDCM25 dataset is a comprehensive benchmark for multi-modal AI in materials science, addressing classification, regression, and explainability tasks. Experiments show that graph-based neural networks outperform convolutional networks in phase classification, while vision transformers yield promising energy predictions, modeling E_T remains particularly challenging, signaling a need for better representations.

LLM evaluations reveal mixed performance in temperature and structural predictions. Some models achieve low temperature prediction errors in ID settings but fail to generalize OOD, whereas others maintain consistent accuracy. In structural predictions, certain LLMs excel in estimating dimensions and titanium counts, yet all models struggle with oxygen counts and Ti:O ratios, indicating a necessity for domain-specific fine-tuning.

TDCM25, while a valuable benchmark for AI-driven materials science, has its limitations. It focuses exclusively on TiO, meaning that incorporating additional materials could enhance its applicability and help validate model findings across diverse systems. Although DFTB simulations offer computational efficiency, they may not match the accuracy of higher-fidelity quantum methods or experimental measurements. Additionally, challenges in integrating multi-modal data and capturing domain-specific nuances might limit the generalizability of models trained on this dataset. Future research should therefore aim to refine model architectures, diversify material inclusion, and utilize advanced simulation techniques to overcome these challenges while further advancing robust, generalizable models for temperature-sensitive materials.

REFERENCES

Dftb+, a software package for efficient approximate density functional theory-based atomistic simulations. *Journal of Computational Chemistry*, 2020. doi: 10.1063/1.5143190. URL https://doi.org/10.1063/1.5143190.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martio Ros-Garca, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research, 2024. URL https://arxiv.org/abs/2411.16955.

Peter W Atkins and Ronald S Friedman. *Molecular quantum mechanics*. Oxford University Press, USA, 2011.

Axel D Becke. Perspective: Fifty years of density-functional theory in chemical physics. *The Journal of Chemical Physics*, 140(18), 2014.

L. C. Blum and J.-L. Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, 131:8732, 2009.

Marta Buchalska, Marcin Kobielusz, Anna Matuszek, Michał Pacia, Szymon Wojtyła, and Wojciech Macyk. On oxygen activation at rutile-and anatase-tio2. *ACS Catalysis*, 5(12):7424–7431, 2015.

Jean-Louis Calais. Density-functional theory of atoms and molecules. rg parr and w. yang, oxford university press, new york, oxford, 1989. ix+ 333 pp. price£ 45.00. *International Journal of Quantum Chemistry*, 47(1):101–101, 1993.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

Xiaobo Chen and Samuel S Mao. Titanium dioxide nanomaterials: synthesis, properties, modifications, and applications. *Chemical Reviews*, 107(7):2891–2959, 2007.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017.

Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T Unke, Adil Kabylda, Huziel E Sauceda, Alexandre Tkatchenko, and Klaus-Robert Müller. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2):eadf0873, 2023.

Y Cho, A Yamaguchi, R Uehara, S Yasuhara, T Hoshina, and M Miyauchi. Temperature dependence on bandgap of semiconductor photocatalysts. *The Journal of Chemical Physics*, 152(23), 2020.

Christian Dette, Miguel A Pérez-Osorio, Christopher S Kley, Paul Punke, Christopher E Patrick, Peter Jacobson, Feliciano Giustino, Soon Jung Jung, and Klaus Kern. Tio2 anatase with a bandgap in the visible region. *Nano Letters*, 14(11):6533–6538, 2014.

Grygoriy Dolgonos, Blint Aradi, Ney H. Moreira, and Thomas Frauenheim. Dftb parameters for ti-o systems. *J. Chem. Theory Comput.*, 6(1):266–278, 2010. doi: 10.1021/ct900422c.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

RS Dubey. Temperature-dependent phase transformation of tio2 nanoparticles synthesized by sol-gel method. *Materials Letters*, 215:312–317, 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.

Alexandre Agm Duval, Victor Schmidt, Alex Hernández-Garcıa, Santiago Miret, Fragkiskos D Malliaros, Yoshua Bengio, and David Rolnick. Faenet: Frame averaging equivariant gnn for materials modeling. In *International Conference on Machine Learning*, pp. 9013–9033. PMLR, 2023.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020.

Sibylle Gemming, Andrey N Enyashin, Johannes Frenzel, and Gotthard Seifert. Adsorption of nucleotides on the rutile (110) surface. *International Journal of Materials Research*, 101(6):758–764, 2010.

Jonathan Godwin, Michael Schaarschmidt, Alexander Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Veličković, James Kirkpatrick, and Peter Battaglia. Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971*, 2021.

Dorian AH Hanaor and Charles C Sorrell. Review of the anatase to rutile phase transformation. *Journal of Materials science*, 46:855–874, 2011.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864, 1964.

Mona Hosseini-Sarvari. Nano-tube tio2 as a new catalyst for eco-friendly synthesis of imines in sunlight. *Chinese Chemical Letters*, 22(5):547–550, 2011.

Ben Hourahine, Bálint Aradi, Volker Blum, Frank Bonafe, Alex Buccheri, Cristopher Camacho, Caterina Cevallos, MY Deshaye, T Dumitrică, A Dominguez, et al. Dftb+, a software package for efficient approximate density functional theory based atomistic simulations. *The Journal of Chemical Physics*, 152(12), 2020.

Weihua Hu, Muhammed Shuaibi, Abhishek Das, Siddharth Goyal, Anuroop Sriram, Jure Leskovec, Devi Parikh, and C Lawrence Zitnick. Forcenet: A graph neural network for large-scale quantum calculations. *arXiv preprint arXiv:2103.01436*, 2021.

John D Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(03):90–95, 2007.

Kuzma Khrabrov, Ilya Shenbin, Alexander Ryabov, Artem Tsypin, Alexander Telepov, Anton Alekseev, Alexander Grishin, Pavel Strashnov, Petr Zhilyaev, Sergey Nikolenko, et al. nabladft: Large-scale conformational energy and hamiltonian prediction benchmark and dataset. *Physical Chemistry Chemical Physics*, 24(42):25853–25863, 2022.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2025 update. *Nucleic Acids Research*, 53(D1): D1516–D1525, 2025.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133, 1965.

Hasan Kurban, Mehmet Dalkilic, Selçuk Temiz, and Mustafa Kurban. Tailoring the structural properties and electronic structure of anatase, brookite and rutile phase tio2 nanoparticles: Dftb calculations. *Computational Materials Science*, 183:109843, 2020.

Mustafa Kurban, Can Polat, Erchin Serpedin, and Hasan Kurban. Enhancing the electronic properties of tio2 nanoparticles through carbon doping: An integrated dftb and computer vision approach. *Computational Materials Science*, 244:113248, 2024.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Zhong Li, ZhengJun Yao, Azhar Ali Haidry, Tomas Plecenik, LiJuan Xie, LinChao Sun, and Qawareer Fatima. Resistive-type hydrogen gas sensor based on tio2: A review. *International Journal of Hydrogen Energy*, 43(45):21114–21132, 2018.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

Yi-Lun Liao, Tess Smidt, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *arXiv preprint arXiv:2403.09549*, 2024.

Ziteng Liu, Liqiang Lin, Qingqing Jia, Zheng Cheng, Yanyan Jiang, Yanwen Guo, and Jing Ma. Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *Journal of Chemical Information and Modeling*, 61(3):1066–1082, 2021.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Tim Luttrell, Sandamali Halpegamage, Junguang Tao, Alan Kramer, Eli Sutter, and Matthias Batzill. Why is anatase a better photocatalyst than rutile?-model studies on epitaxial tio2 films. *Scientific Reports*, 4(1): 4043, 2014.

CBea Murray, David J Norris, and Moungi G Bawendi. Synthesis and characterization of nearly monodisperse cde (e= sulfur, selenium, tellurium) semiconductor nanocrystallites. *Journal of the American Chemical Society*, 115(19):8706–8715, 1993.

David Pfau, James S Spencer, Alexander GDG Matthews, and W Matthew C Foulkes. Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Physical Review Research*, 2(3): 033429, 2020.

Can Polat, Mustafa Kurban, and Hasan Kurban. Multimodal neural network-based predictive modeling of nanoparticle properties from pure compounds. *Machine Learning: Science and Technology*, 5(4):045062, 2024.

Can Polat, Hasan Kurban, and Mustafa Kurban. Quantumshellnet: ground-state eigenvalue prediction of materials using electronic shell structures and fermionic properties via convolutions. *Computational Materials Science*, 246:113366, 2025.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):1–7, 2014.

AM Rashad, Y Bai, PA Muhammed Basheer, NC Collier, and NB Milestone. Chemical and mechanical stability of sodium sulfate activated slag after exposure to elevated temperature. *Cement and Concrete Research*, 42(2):333–343, 2012.

Laura E Ratcliff, Stephan Mohr, Georg Huhs, Thierry Deutsch, Michel Masella, and Luigi Genovese. Challenges in large scale quantum mechanical calculations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(1):e1290, 2017.

Aleks Reinhardt, Chris J Pickard, and Bingqing Cheng. Predicting the phase diagram of titanium dioxide with random search and pattern recognition. *Physical Chemistry Chemical Physics*, 22(22):12697–12705, 2020.

David Reyes-Coronado, G Rodríguez-Gattorno, ME Espinosa-Pesqueira, C Cab, RD De Coss, and G Oskam. Phase-pure tio2 nanoparticles: anatase, brookite and rutile. *Nanotechnology*, 19(14):145605, 2008.

Emil Roduner. Understanding catalysis. *Chemical Society Reviews*, 43(24):8226–8239, 2014.

Gregory S Rohrer. *Structure and bonding in crystalline materials*. Cambridge University Press, 2001.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012.

M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108:058301, 2012.

Abhishek Sarkar, Leonardo Velasco, DI Wang, Qingsong Wang, Gopichand Talasila, Lea de Biasi, Christian Kübel, Torsten Brezesinski, Subramshu S Bhattacharya, Horst Hahn, et al. High entropy oxides for reversible energy storage. *Nature Communications*, 9(1):3400, 2018.

Erwin Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Physical Review*, 28 (6):1049, 1926.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*, 30, 2017a.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pp. 9377–9388. PMLR, 2021.

Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1):13890, 2017b.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

Oliver T Unke, Stefan Chmiela, Michael Gastegger, Kristof T Schütt, Huziel E Sauceda, and Klaus-Robert Müller. Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature Communications*, 12(1):7273, 2021.

Ingrid von Glehn, James S Spencer, and David Pfau. A self-attention ansatz for ab-initio quantum chemistry. *arXiv preprint arXiv:2211.13672*, 2022.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Geemi Wellawatte, Huixuan Guo, Magdalena Lederbauer, Anna Borisova, Matthew Hart, Marta Brucka, and Philippe Schwaller. Chemlit-qa: A human evaluated dataset for chemistry rag tasks. In *AI for Accelerated Materials Design-NeurIPS 2024*, 2024.

T Wolf. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Julia M Yeomans. *Statistical mechanics of phase transitions*. Clarendon Press, 1992.

Anna Yershova and Steven M LaValle. Deterministic sampling methods for spheres and so (3). In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 4, pp. 3974–3980. IEEE, 2004.

Haiyang Yu, Meng Liu, Youzhi Luo, Alex Strasser, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Qh9: A quantum hamiltonian prediction benchmark for qm9 molecules. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Fan Zhang, Rong-Jun Zhang, Dong-Xu Zhang, Zi-Yi Wang, Ji-Ping Xu, Yu-Xiang Zheng, Liang-Yao Chen, Ren-Zhong Huang, Yan Sun, Xin Chen, et al. Temperature-dependent optical properties of titanium oxide thin films studied by spectroscopic ellipsometry. *Applied Physics Express*, 6(12):121101, 2013.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*, 2024.

Huanjun Zhang, Guohua Chen, and Detlef W Bahnemann. Photoelectrocatalytic materials for environmental applications. *Journal of Materials Chemistry*, 19(29):5089–5121, 2009.

Wei Zhang, Haili He, Haoze Li, Linlin Duan, Lianhai Zu, Yunpu Zhai, Wei Li, Lianzhou Wang, Honggang Fu, and Dongyuan Zhao. Visible-light responsive tio2-based materials for efficient solar energy utilization. *Advanced Energy Materials*, 11(15):2003303, 2021.

Yun Zhang and Xiaojie Xu. Machine learning band gaps of doped-tio2 photocatalysts from structural and morphological parameters. *ACS omega*, 5(25):15344–15352, 2020.

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.

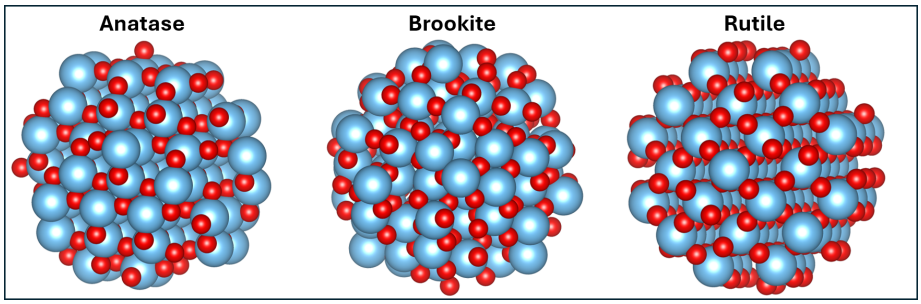

Figure 4: Structural models of anatase, brookite, and rutile phase TiO_2 NPs.

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

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