A APPENDIX

A.1 DATASETS

Table 4: Number of experimental (exp. in the header) and computational (comp. in the header) samples in `small` and `large` datasets.

Dataset	Exp.	Comp.	Total
`small`	0.9k	3.5k	4.4k
`large`	3.5k	21k	24k

A.2 METRICS

To evaluate the model's performance separately on experimental and computational data, we implement a stratified sampling strategy that takes into account the source labels of both data types, which are stratified and split into 80% training and 20% testing datasets. To evaluate the performance of the regression models, we adopt two primary metrics: **mean absolute error (MAE)** and its **standard deviation (STD)** across three fixed random seeds to evaluate the performance of these regression models. These metrics were selected to assess both the accuracy and stability of the model predictions, ensuring robust evaluation of the results. The MAE quantifies the average magnitude of the errors between the predicted bandgap values $\hat{y}_i$ and the true bandgap values y_i, defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where n denotes the total number of samples in the dataset. A lower MAE indicates better predictive accuracy.

The **STD** captures the variability of the prediction errors, reflecting the spread of the differences between the predicted values and the true values. It is computed as:

$$\text{STD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((\hat{y}_i - y_i) - \overline{\delta}\right)^2}$$

where $\delta = \hat{y}_i - y_i$ is the error for the i-th prediction, $\overline{\delta}$ is the mean error across all predictions, and n is the total number of samples. A lower STD suggests that the errors are more tightly clustered around their mean, indicating more consistent predictions. The data shuffling is performed using three fixed random seeds to ensure reproducibility and robustness. The final evaluation results are reported as the average MAE across the three seeds, along with its corresponding STD.

A.3 Embeddings

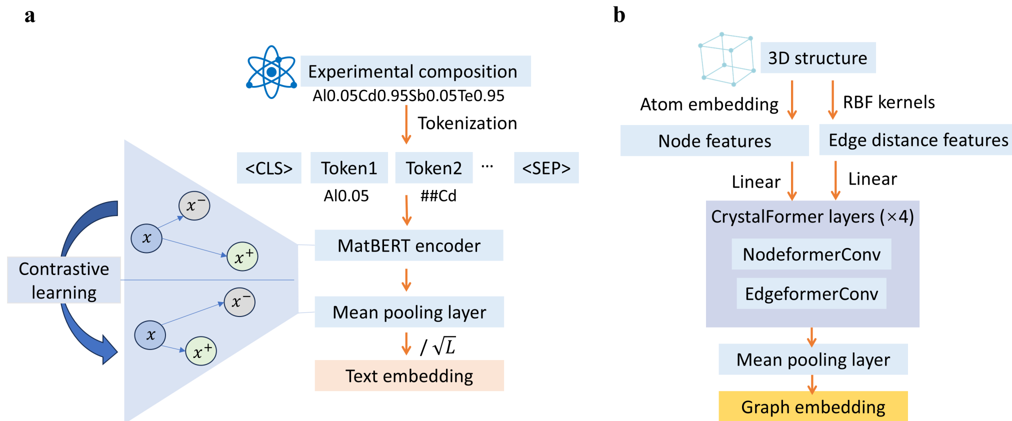

Figure 2: Process of obtaining a) text embeddings from fine-tuned MatBERT and b) structural embedding from CrystalFormer. BERT model use the same framework SBERT (Reimers & Gurevych, 2019) as the fine-tuned MatBERT to obtain text embeddings. LLaMA-based models like DARWIN, LLaMA2, use llm2vec (BehnamGhader et al., 2024) while T5 directly uses Sentence-T5 (Ni et al., 2021)

A.4 Ablation Studies

Our dimensionality reduction experiments evaluate the impact of text embedding dimensions on material bandgap prediction performance. On `small`, reducing contextual and formula embeddings from 768 to 128 dimensions decreases predictive performance, with MAE increasing from 0.8733 to 0.8872(I)/0.8981(D) for contextual embeddings and from 0.8250 to 0.8977(I)/0.9228(D) for formula embeddings. However, when concatenated with graph embeddings, performance remains comparable to original 768-dimensional representations. On `large`, dimensionality reduction affects performance more significantly, particularly for formula embeddings, where MAE increases from 0.5658 to 0.6319(I)/0.6819(D). While reducing dimensions may diminish individual feature representation, combining with complementary information maintains performance, especially for small datasets, providing guidance for dimension selection under computational constraints.

Table 5: Performance comparison between original and reduced dimensions on `small` and `large` datasets.

Data Size	Method	Dimention	MAE_{exp}	MAE_{comp}
Small	T_{ctx}	768	0.8733	0.7657
	T_{ctx}(I)	128	0.8872	0.7844
	T_{ctx}(D)	128	0.8981	0.8200
	$T_{ctx} \| G$	768 + 128	0.6098	0.3219
	$T_{ctx} \| G$(I)	128 + 128	0.6007	0.3220
	$T_{ctx} \| G$(D)	128 + 128	0.5981	0.3243
	T_{fml}	768	0.8250	0.7943
	T_{fml}(I)	128	0.8977	0.8063
	T_{fml}(D)	128	0.9228	0.8453
	$T_{fml} \| G$	768 + 128	0.6039	0.3230
	$T_{fml} \| G$(I)	128 + 128	0.5799	0.3356
	$T_{fml} \| G$(D)	128 + 128	0.6074	0.3257
Large	T_{fml}	768	0.5658	0.6742
	T_{fml}(I)	128	0.6319	0.7182
	T_{fml}(D)	128	0.6819	0.7547
	$T_{fml} \| G$	768 + 128	0.3671	0.2808
	$T_{fml} \| G$(I)	128 + 128	0.3786	0.2763
	$T_{fml} \| G$(D)	128 + 128	0.3804	0.2769

T_{ctx}: Contextual embedding; T_{fml}: Formula embedding; G: Graph embedding; (I): Indirect method (768-dimensional embedding reduced to 128 dimensions); (D): Direct method (128-dimensional embedding extracted directly from text model); $\|$: Feature concatenation; MAE_{exp}: Mean Absolute Error of experimental materials; MAE_{comp}: Mean Absolute Error of computational materials