# Supplementary Information

**Interpretable formation-energy structure from composition, density, and band-gap descriptors**

Helena Cristina Vasconcelos; Maria Meirelles

## S1. Reproducibility details

- **Data source:** Materials Project (MP) via the MP API accessed through the *pymatgen* interface.
- **API access date:** 22 May 2026.
- **Dataset size:** 5000 crystalline inorganic compounds after filtering for availability of formation energy per atom $E_f$, mass density $\rho$, and DFT band gap $E_g$.
- **Train/test split:** 4000/1000 IID split with fixed random seed = 42.
- **Rare-element OOD split:** $k = 7$, $S = \{\mathrm{Tc}, \mathrm{Pa}, \mathrm{Os}, \mathrm{Re}, \mathrm{Ac}, \mathrm{Lu}, \mathrm{Np}\}$, $N_{\mathrm{OOD}} = 200$.
- **Model:** Random Forest regressor; forest size $N_{\mathrm{trees}} = 600$; minimum leaf size $n_{\mathrm{leaf}} = 1$.
- **Uncertainty proxy:** ensemble of $N = 5$ Random Forest models trained with different random seeds; prediction spread defined as the standard deviation across ensemble predictions.
- **UMAP visualisation (parameters used):** neighbourhood size $n_{\mathrm{nb}} = 15$; minimum-distance parameter $d_{\mathrm{min}} = 0.1$; Euclidean metric; random seed = 42.

## S2. Supplementary data files

Two supplementary data files accompany this manuscript:

- **Supplementary Data 1:** *Supplementary_Data_1_material_ids.csv* — list of the 5000 MP material identifiers used in this study.
- **Supplementary Data 2:** *Supplementary_Data_2_dataset.csv* — curated dataset used for modelling, containing the MP material identifier, chemical formula, formation energy per atom $E_f$, DFT band gap $E_g$, and mass density $\rho$.

## S3. Notes on interpretation

Formation energy $E_f$ is reported relative to elemental reference states as provided by MP. Negative $E_f$ indicates energetic favourability relative to the elements but does not guarantee convex-hull stability. The low-dimensional embeddings (PCA/UMAP) are statistical visualisations in descriptor space and should not be interpreted as thermodynamic potential surfaces or physical stability landscapes.