UMA-DOS: TRANSFER LEARNING FROM PRETRAINED ATOMISTIC MODELS FOR DENSITY OF STATES PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Density Of States (DOS) is a fundamental spectral property of crystalline materials that governs a wide range of chemical, electronic, and optical properties. Despite its importance, fast but accurate DOS prediction based on machine learning remains challenging because it involves a high-dimensional structured target, the physically meaningful features of which depend on subtle relations between crystal structure and energy-resolved electronic states. Existing machine-learning approaches primarily focus on learning strong structural representations from crystal graphs, but often rely on DOS supervision alone and do not fully exploit the transferable information already learned by large pretrained atomistic models. In this work, we present UMA-DOS, a transfer-learning framework for total DOS prediction built on pretrained UMA backbones. UMA-DOS uses a pretrained atomistic encoder to represent crystal structure, aggregates node-level features into a graph-level representation, and maps this representation to an energy-resolved DOS through a structured spectral decoder with explicit shape–amplitude factorization. Under a unified Materials Project evaluation pipeline, UMA-DOS achieves the strongest controlled performance among the reproduced baselines, with its clearest advantages appearing in shape-sensitive measures of spectral fidelity. Transfer ablations further show that the value of pretraining is realized only through downstream adaptation: frozen transfer underperforms scratch training, whereas end-to-end fine-tuning yields the best results. These findings show that pretrained atomistic representations can transfer effectively to total DOS prediction when paired with a shape-aware spectral decoder and adapted end-to-end.

1 INTRODUCTION

The electronic Density Of States (DOS) is among the most fundamental quantities in condensed matter physics and materials science (Cohen & Chelikowsky, 1988). It describes the distribution of electronic energy levels in a material, encoding, in a single compact quantity, information that governs many key physical properties, especially in the context of electrical and optical behavior. Despite this centrality, obtaining accurate DOS remains a persistent bottleneck in computational materials science.

First-principles electronic structure methods, rooted in density functional theory (DFT) (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), offer a rigorous framework for computing the DOS directly from crystal structure. Although they can be highly accurate, DFT calculations are computationally prohibitive for high-throughput screening of a wide range of systems, demanded by modern materials discovery. This computational cost bottleneck has motivated a growing interest in machine learning (ML) as a means of making large-scale, accurate electronic structure predictions, as shown in Figure 1. Several studies have tackled DOS prediction directly from structural input using ML, with an increasing emphasis on universal models trained across broad chemical spaces, as opposed to tailored models for specific systems. Graph neural network (GNN) architectures have emerged as the dominant approach, and some studies have also explored transformers and LSTM-based architectures to better capture energy correlations in the spectral output (Kong et al., 2022; Kaundinya et al., 2022; Bai et al., 2023; Lee et al., 2023; Cui et al., 2023; Shermukhamedov et al., 2024; How et al., 2026). These efforts represent significant progress: modern models generalize across

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

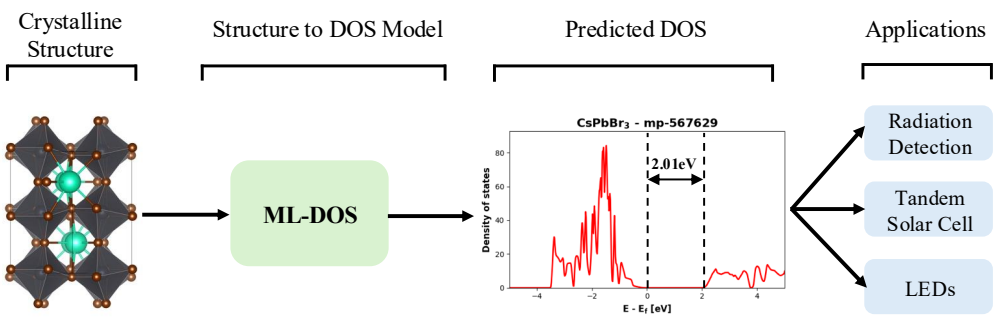


Figure 1: Overview of the structure-to-DOS prediction problem. Given a crystalline structure, ML-DOS predicts the density of states (DOS), an energy-resolved description of the available electronic states relative to the Fermi level (E_F). Features of the predicted DOS, such as the band gap and spectral weight near E_F, are informative for downstream materials screening in applications including radiation detection, solar cells, and light-emitting diodes (LEDs). The application boxes illustrate downstream relevance rather than direct prediction targets in this work.

diverse chemical systems and reproduce the broad shape of the DFT-computed DOS in many cases. However, predictions have not reached the level of accuracy that materials applications demand. Common failure modes include missing sharp peaks, insufficiently accurate peak positions, and distorted spectral features, which are precisely the details that carry critical physical and chemical information (Toriyama et al., 2022). For instance, the presence or absence of states at the Fermi level can determine whether a material is metallic or semiconducting/insulating; a shifted or absent edge feature can make the difference between a promising photocatalyst and an inactive one. Therefore, even impressively low average error metrics can still mask physically consequential inaccuracies, and closing this gap requires not just quantitative improvement, but architectural choices that prioritize spectral fidelity. The difficulty of predicting the DOS is also affected by limited data availability. Although many datasets exist for scalar- or vector-valued quantities such as total energies, band gaps, and forces (Kirklin et al., 2015; Choudhary et al., 2020; Barroso-Luque et al., 2024), datasets for continuous spectral quantities like the DOS are far less common, typically containing at most tens of thousands of materials (Jain et al., 2013; Choudhary et al., 2020). Compared with energies and forces, the DOS is also a much richer target: it is a global, high-dimensional object defined over the energy axis, rather than a scalar or local quantity. This makes learning structure-to-spectrum mappings from DOS supervision alone intrinsically data-limited. To address this challenge, we leverage machine-learning interatomic potentials (MLIPs) pretrained on large-scale datasets of energies and forces (Batatia et al., 2025; Wood et al., 2026), creating an appealing but nontrivial opportunity for transfer to DOS prediction. This raises a central transfer-learning question: *can broad supervised pretraining for lower-dimensional quantities provide a useful prior for predicting high-dimensional electronic spectra, and whether frozen pretrained representations are sufficient, and how much downstream adaptation is needed?*

In this work, we address this question by introducing `UMA-DOS`, a transfer-learning framework for total DOS prediction built on pretrained UMA (Wood et al., 2026) backbones. `UMA-DOS` uses a pretrained atomistic encoder to represent crystal structure and couples it to a task-specific spectral prediction head for energy-resolved regression. We systematically study both frozen and fine-tuned transfer across three pretrained backbones, `UMA-s-1p1`, `UMA-s-1p2`, and `UMA-m-1p1`, and compare mean pooling with learned softmax-weighted pooling to understand how readout design mediates the transfer from atomistic representations to electronic spectra. For `UMA-s-1p2`, we additionally study parameter-efficient adaptation with ELoRA (Wang et al., 2025) to test whether pretrained representations can be adapted effectively without updating the original backbone weights. Because reported DOS results are often confounded by differences in smoothing conventions, energy references, and preprocessing pipelines, we also emphasize a controlled evaluation protocol and treat unmatched external baselines as contextual rather than strictly comparable.

Our contributions are fourfold. First, we formulate DOS prediction as a transfer-learning problem with a distinctive supervision mismatch: broad, data-rich pretraining on lower-dimensional atomistic targets, followed by adaptation to a data-poor, information-rich spectral target. Second, we introduce `UMA-DOS`, a transfer framework that repurposes pretrained UMA representations for DOS predic-

tion through a structured spectral decoder with explicit shape–amplitude factorization. Third, we systematically study the transfer regime, checkpoint identity, and pooling design. Full fine-tuning achieves the strongest spectral fidelity, while ELoRA recovers much of its improvement with fewer trainable parameters. Fourth, under a unified evaluation protocol, we show that pretrained atomistic representations can transfer effectively to total DOS prediction and achieve the strongest controlled performance among the reproduced baselines. Taken together, these results indicate that successful transfer to structured electronic spectra depends not only on the pretrained encoder but also on downstream adaptation and task-aware output design.

2 Related Work

Direct DOS prediction from crystal structure. A growing line of work has addressed DOS prediction directly from the crystal structure, obtained by adapting crystal encoders to structured spectral outputs. ALIGNN (Kaundinya et al., 2022) extended line-graph neural networks to DOS prediction and considered both direct discretized spectra and compressed latent targets. Mat2Spec (Kong et al., 2022) introduced probabilistic embeddings and supervised contrastive learning for spectral prediction, based on the Materials Project dataset (Jain et al., 2013). Xtal2DoS (Bai et al., 2023) reframed DOS prediction as a crystal-to-sequence problem, coupling graph-based crystal encoding with an attention-based decoder to capture correlations along the energy axis. DOSTransformer (Lee et al., 2023) pushed this idea further by explicitly conditioning DOS prediction on energy levels through a prompt-guided multimodal transformer that cross-attends between crystal representations and energy embeddings. More recently, PET-MAD-DOS (How et al., 2026) moved toward broader-domain DOS prediction by adapting the PET architecture to DOS, training on the MAD dataset and evaluating on both in-domain and external benchmarks. Taken together, these studies show steady progress in learning structured electronic spectra from atomistic inputs, but they primarily rely on supervision from DOS targets themselves.

Pretrained atomistic foundation models and transfer. Large atomistic foundation models have emerged as general-purpose structural encoders pretrained on energies, forces, and related atomistic signals. MACE (Batatia et al., 2025) demonstrated that a universal model trained on large atomistic trajectory data can be deployed "out of the box" and further adapted with relatively small downstream datasets. PET-MAD (Mazitov et al., 2025a) emphasized universality through a chemically diverse training corpus spanning both organic and inorganic systems, achieving competitive performance across a wide range of benchmarks despite substantially less training data than several reference universal MLIPs. UMA (Wood et al., 2026) scaled this paradigm further by training on half a billion unique three-dimensional atomistic structures across molecules, materials, and catalysts, and introduced a mixture-of-linear-experts design to increase capacity without sacrificing efficiency. These models suggest that large-scale atomistic pretraining can produce broadly transferable structural representations. However, they were generally neither designed for nor applied to the prediction of structured electronic spectra. Among these models, we use UMA as the backbone for this work because its large-scale multi-domain pretraining, released checkpoint variants, and end-to-end accessible atom-wise representations provide a convenient setting for systematically isolating transfer regime and checkpoint effects.

`UMA-DOS` lies at the intersection of these two directions. Relative to prior DOS predictors such as ALIGNN (Kaundinya et al., 2022), Mat2Spec (Kong et al., 2022), Xtal2DoS (Bai et al., 2023), and DOSTransformer (Lee et al., 2023), our goal is not to learn DOS solely from DOS supervision, but to test whether pretrained atomistic representations provide an effective structural prior for data-limited spectral prediction. Relative to atomistic foundation models such as MACE-MP-0 (Batatia et al., 2025), PET-MAD (Mazitov et al., 2025a), and UMA (Wood et al., 2026), our contribution is not a new interatomic potential, but a downstream adaptation framework that repurposes pretrained atomistic backbones for structured electronic spectra through a task-specific spectral head. PET-MAD-DOS (How et al., 2026) is the closest prior work bridging these directions, but it remains distinct from `UMA-DOS` because PET-MAD-DOS is trained directly on DOS targets, whereas `UMA-DOS` explicitly studies transfer from atomistic pretraining and the role of frozen versus fine-tuned adaptation in enabling that transfer.

We emphasize that direct numerical comparison with prior ML-DOS work should be interpreted cautiously, because target definitions, energy referencing, and preprocessing differ across studies.

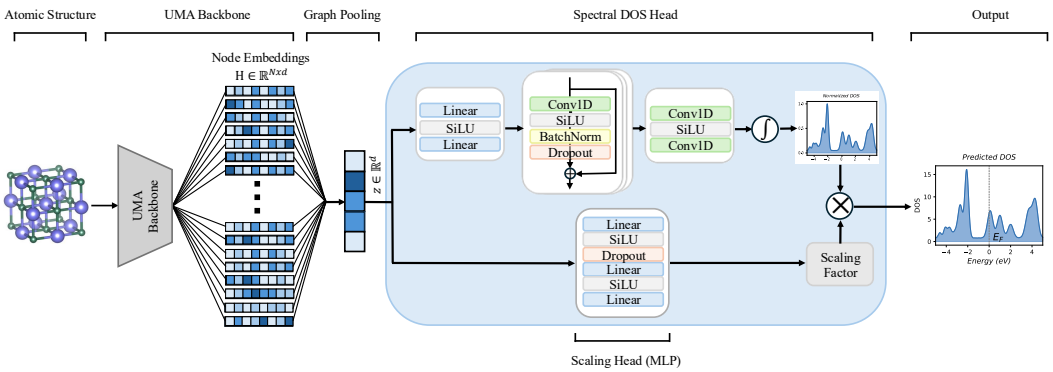


Figure 2: Overview of UMA-DOS. A crystal structure is encoded by a pretrained UMA backbone into node embeddings H. A graph pooling layer aggregates these embeddings into a graph-level descriptor z using either mean pooling or learned softmax-weighted pooling. The spectral DOS head predicts (i) a normalized DOS shape on a fixed energy grid through projection, residual dilated Conv1D blocks, readout, and Softplus, and (ii) a log-scale amplitude through a parallel MLP. The final total DOS is obtained by rescaling the normalized prediction with the predicted amplitude.

For example, Mat2Spec uses a DOS representation aligned so that the Fermi level and band edges are set to 0 eV, with the zero-valued gap region removed for gapped materials; DOSTransformer uses its own Materials Project-derived DOS preprocessing pipeline; and PET-MAD-DOS is trained on the MAD dataset and additionally evaluated on external datasets. We therefore treat unmatched prior results as contextual rather than strictly comparable, and focus our main claims on controlled comparisons under a unified evaluation protocol.

3 METHOD

3.1 PROBLEM FORMULATION AND TARGET PARAMETERIZATION

We begin by formally defining the problem. A crystal structure consists of a parallelepiped unit cell that is repeated periodically throughout space. More precisely, a crystal structure is represented as $\mathcal{C} = \left(L, \{(z_i, \boldsymbol{r}_i)\}_{i=1}^{N}\right)$, where $L = [\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3]$ is the lattice matrix whose columns are the three independent lattice vectors defining the unit cell, and $\{(z_i, \boldsymbol{r}_i)\}_{i=1}^{N}$ denotes the atomic species z_i and positions $\boldsymbol{r}_i$ of the N atoms within the unit cell. Given this representation, our goal is to predict the total DOS, denoted by $D(E)$, from an input crystal structure $\mathcal{C}$. Although $D(E)$ is a non-negative function of energy E, in practice we aim to predict its discretized values on a set of energy grid points $\boldsymbol{e} = [E_1, E_2, \ldots, E_T]$, namely, $\boldsymbol{d} = [D(E_1), D(E_2), \ldots, D(E_T)] \in \mathbb{R}_{\geq 0}$.

This task poses two main challenges. First, labeled DOS data are relatively scarce, making it difficult to train a high-capacity model from scratch. Second, DOS is a structured, high-dimensional target, the overall magnitude of which can vary substantially across materials. To address these challenges, we propose a DOS prediction framework based on two key ideas: transfer learning to leverage a machine-learning interatomic-potential model which trained on large-scale data, and problem decomposition to decouple spectral structure from overall magnitude. We focus on the latter here and describe the transfer-learning component in the next section. Rather than directly regressing $\boldsymbol{d}$, we decompose the target into a normalized spectral shape and an overall scale. Let $m = \max_{1 \leq t \leq T} D(E_t)$. We then define $\boldsymbol{y} = \boldsymbol{d}/m, a = \log(m)$, where $\boldsymbol{y} \in [0, 1]^T$ is the normalized DOS shape, and a is a scalar log-scale target representing the overall DOS magnitude. Under this parameterization, the original DOS can be recovered as $\boldsymbol{d} = \exp(a)\, \boldsymbol{y}$. This decomposition avoids direct regression of $\boldsymbol{d}$, which may cause large-amplitude spectra to dominate the optimization.

3.2 UMA-DOS ARCHITECTURE

Because labeled DOS data are limited, we approach DOS prediction via transfer learning from UMA, a machine-learned interatomic potential pretrained on large-scale energy and force data. For a crystal structure $\mathcal{C}$, the UMA encoder, after removing its final energy and force prediction heads, produces

atomic embeddings $\mathcal{H} = \{\boldsymbol{h}_i\}_{i=1}^{N}$, with $\boldsymbol{h}_i \in \mathbb{R}^{d_\mathrm{h}}$ for the N atoms in the unit cell. From these embeddings, we predict the normalized target $\boldsymbol{y}$ and the scale parameter a.

Transferring a pretrained atomistic model to DOS prediction is non-trivial for three reasons. First, the number of atoms N varies across crystal structures, requiring permutation-invariant aggregation of atomic embeddings into a fixed-dimensional crystal representation $\boldsymbol{z} \in \mathbb{R}^{d_\mathrm{h}}$. Second, the prediction target is a discretized representation of a continuous spectrum, so the output is structured rather than a collection of independent dimensions. Third, to benefit from UMA pretraining, the encoder must be adapted with an effective transfer-learning strategy. Our method, `UMA-DOS`, is designed to address these three challenges jointly. Specifically, `UMA-DOS` maps a crystal structure to a DOS prediction through

$$\mathcal{C} \xrightarrow{f_\theta} \mathcal{H} \xrightarrow{p_\psi} \boldsymbol{z} \xrightarrow{g_\phi} (\boldsymbol{y}, a), \tag{1}$$

where f_θ denotes the pretrained UMA encoder, p_ψ is a pooling operator, and g_ϕ is a DOS-specific spectral decoder. Here, p_ψ converts the variable-size set of atomic embeddings into a crystal-level representation, g_ϕ models the structured spectral output, and fine-tuning θ, ψ, and ϕ transfers pretrained knowledge to the downstream DOS task. We next describe the designs of p_ψ and g_ϕ, followed by the training objective and transfer settings for θ, ψ, and ϕ.

3.2.1 Pooling of UMA output

We study two pooling strategies p_ψ for converting the variable-length set of atomic embeddings produced by UMA into a fixed-length crystal representation.

Mean pooling. The first approach is simple mean pooling over all atomic embeddings. This method treats all atoms equally and summarizes the crystal structure by the average of their embeddings.

Softmax-weighted pooling. The second approach is motivated by the possibility that not all atoms contribute equally to DOS prediction. To allow the model to focus on more informative atoms, we introduce a learnable weighted pooling mechanism. Specifically, each atomic embedding $\boldsymbol{h}_i \in \mathbb{R}^{d_\mathrm{h}}$ is projected to a scalar score using a trainable vector $\boldsymbol{w}_\mathrm{pool} \in \mathbb{R}^{d_\mathrm{h}}$, and the resulting scores are normalized by a softmax function to obtain attention-like weights:

$$q_i = \boldsymbol{w}_\mathrm{pool}^\top \boldsymbol{h}_i, \qquad \alpha_i = \frac{\exp(q_i)}{\sum_{j=1}^{N} \exp(q_j)}, \qquad \boldsymbol{z} = \sum_{i=1}^{N} \alpha_i \boldsymbol{h}_i. \tag{2}$$

Here, α_i represents the relative importance of atom i, and the pooled representation $\boldsymbol{z}$ is given by the weighted sum of the atomic embeddings.

3.2.2 Spectral decoder for DOS prediction

The spectral decoder g_ϕ takes as input the pooled UMA representation $\boldsymbol{z}$ and outputs both the normalized DOS $\boldsymbol{y}$ and a scale factor a through two independent prediction heads. The scale factor a is predicted using a simple multi-layer perceptron (MLP). In contrast, predicting the normalized DOS requires a more structured decoder. A standard MLP head is poorly suited to DOS prediction because it treats the output bins as largely independent. However, neighboring energy bins in a DOS spectrum are strongly correlated, and physically important features such as peaks and band edges are inherently local along the energy axis. We therefore design a spectral decoder that explicitly models local structure across nearby energy bins, rather than predicting all bins independently. Given a pooled crystal-level representation $\boldsymbol{z} \in \mathbb{R}^{d_\mathrm{h}}$, we first project it into a latent spectral tensor with d_s-dimensional embeddings over T energy bins:

$$\boldsymbol{s} = \mathrm{MLP}_\mathrm{proj}(\boldsymbol{z}) \in \mathbb{R}^{d_\mathrm{s} T}, \qquad S^{(0)} = [\boldsymbol{s}_1^{(0)}, \boldsymbol{s}_2^{(0)}, \ldots, \boldsymbol{s}_T^{(0)}] = \mathrm{Reshape}(\boldsymbol{s}) \in \mathbb{R}^{d_\mathrm{s} \times T}, \tag{3}$$

where $\mathrm{Reshape}(\cdot)$ maps the length-($d_\mathrm{s} \times T$) vector into a tensor consisting of T energy bins, each with a d_s-dimensional embedding. Each $\boldsymbol{s}_i^{(0)} \in \mathbb{R}^{d_\mathrm{s}}$ represents the latent spectral feature at the i^th energy bin. This tensor can be viewed as an initial coarse representation of the DOS along the discretized energy axis. We then refine this representation using residual one-dimensional convolutional blocks:

$$S^{(\ell+1)} = S^{(\ell)} + \mathcal{B}_\ell\big(S^{(\ell)}\big), \qquad \ell = 0, \ldots, L-1, \tag{4}$$

where each block $\mathcal{B}_\ell$ consists of a dilated convolution layer followed by a SiLU activation, and batch normalization. Intuitively, the convolution updates each energy bin using features from neighboring

bins. By incorporating this local spectral context at each step, the decoder produces DOS profiles that are smoother and more locally consistent. We use three residual convolutional layers with dilation factors 1, 2, and 4, respectively, so that the decoder captures both short-range and medium-range correlations along the energy axis. After these residual updates, the refined tensor $S^{(L)} \in \mathbb{R}^{d_s \times T}$ is converted into the DOS by a readout module composed of multiple convolutional layers:

$$\boldsymbol{y} = \mathrm{Softplus}\big(\mathrm{Readout}\big(S^{(L)}\big)\big) \in \mathbb{R}_{\geq 0}, \tag{5}$$

where the Softplus activation ensures a non-negative output. Architecture details are provided in Appendix A.1.

3.2.3 Training objective

We train with $\mathcal{L} = \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} + \lambda_{\text{scale}}\,\mathcal{L}_{\text{scale}} + \lambda_{\text{cdf}}\,\mathcal{L}_{\text{cdf}}$, where $\mathcal{L}_{\text{shape}}$ is a critical-point-weighted ℓ_1 loss on the normalized DOS that upweights zero bins, band edges, and turning points; $\mathcal{L}_{\text{scale}}$ supervises log-amplitude; $\mathcal{L}_{\text{cdf}}$ encourages cumulative spectral agreement; and λ_{shape}, λ_{scale}, and λ_{cdf} are weighting hyperparameters; full details are given in Appendix A.2.

3.2.4 Transfer Regimes

The transfer setting is not only a training choice but also a way to probe what pretraining contributes to DOS prediction.

Frozen transfer. We keep the pretrained UMA parameters θ fixed and optimize only the pooling and spectral-head parameters. This setting asks whether DOS-relevant information is already accessible from the pretrained representation with only a task-specific readout.

Parameter-efficient transfer (ELoRA). We keep the pretrained UMA parameters θ fixed and optimize ELoRA adapters (Wang et al., 2025) jointly with the pooling parameters ψ and spectral-decoder parameters ϕ. This setting asks whether adapting pretrained representations through low-rank updates is sufficient for effective transfer to DOS prediction.

Fine-tuned transfer. We optimize the full model end-to-end, updating both the pretrained encoder and the newly introduced pooling and decoder parameters. This setting asks whether pretraining is most useful as an initialization that must be adapted to the downstream spectral task.

Comparing frozen transfer, ELoRA, and full fine-tuning lets us examine whether DOS-relevant information is accessible through fixed pretrained representations, and how much encoder adaptation is needed to improve spectral reconstruction. We additionally compare against scratch training as a non-transfer baseline, testing whether each transfer strategy improves over learning the DOS mapping without pretraining.

4 Experiments

In this section, we design experiments to address the questions: the utility of large-scale atomistic pretraining as a prior for predicting high-dimensional electronic spectra such as the DOS, and the need for end-to-end adaptation as opposed to using frozen representations directly. We begin by evaluating overall performance (Section 4.2) and data efficiency (Section 4.3). We then analyze transfer-learning and pooling strategy ablations to clarify how pretraining contributes to performance (Section 4.4), followed by an examination of architectural design choices (Section 4.5). Finally, we provide a qualitative analysis (Section 4.6).

4.1 Experimental Setup

We evaluate `UMA-DOS` on a publicly available total-DOS dataset derived from the Materials Project repository (Jain et al., 2013). We use a fixed split which contains 23,502 training structures, 2,572 validation structures, and 2,661 test structures. All DOS spectra are aligned to the Fermi level and represented on a fixed grid of 181 bins over the range $[-5, 5]$ eV. Unless otherwise stated, all reported metrics are computed against ground-truth DOS curves, smoothed with a one-dimensional Gaussian kernel with $\sigma = 1.0$ in grid units.

The main `UMA-DOS` configuration uses a pretrained UMA-s-1p2 backbone, softmax-weighted pooling, and the proposed residual dilated spectral DOS head. This configuration was selected using

Table 1: Comparison of DOS predictors under a unified evaluation pipeline on the Materials Project test set. Results are shown as mean ± standard deviation over 3 random seeds. Lower is better for all metrics.

Model	MAE ↓	nMAE ↓	SAM ↓
Mat2Spec-MAE	2.682±0.010	0.166±0.001	0.439±0.001
DOSTransformer	2.579 ±0.012	0.135±0.003	0.403±0.002
UMA-DOS	**2.561**±0.039	**0.112** ±0.003	**0.380** ±0.001

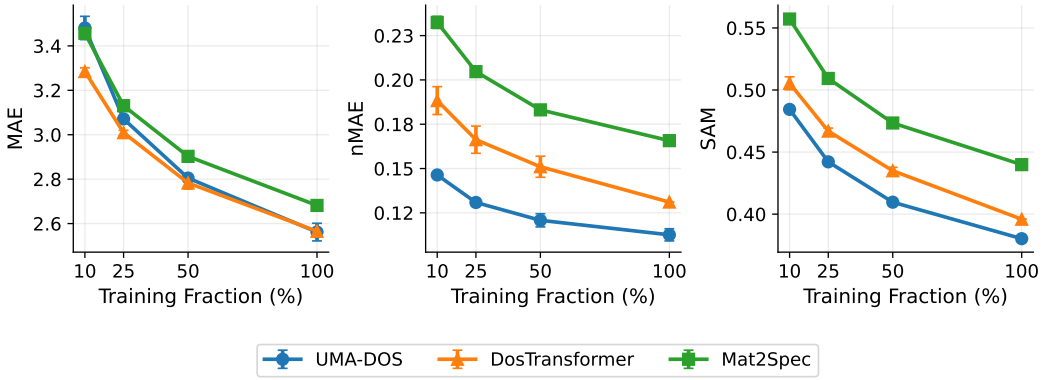


Figure 3: Low-data comparison under the unified Materials Project protocol. Models are trained on 10%, 25%, 50%, and 100% of the training set; error bars show mean ± standard deviation over three seeds. Lower is better. UMA-DOS achieves the best nMAE and SAM at every training fraction, while MAE is competitive at low data and best at full data.

validation performance among the variants described below, and we refer to it as UMA-DOS in the following comparisons. To understand the role of pretraining and downstream adaptation, we compare four training settings: *scratch* training, *frozen* transfer, *ELoRA adaptation*, and *full fine-tuning*. Unless otherwise noted, UMA-DOS denotes the fully fine-tuned UMA-s-1p2 configuration with softmax-weighted pooling. We compare against reproduced structure-to-DOS baselines, including Mat2Spec (Kong et al., 2022) and DOSTransformer (Lee et al., 2023), under the same split and evaluation pipeline; PET-MAD-DOS (How et al., 2026) is reported only as a contextual comparison because it is trained on a different dataset and does not cover the full test set.

We report mean absolute error (MAE) for compatibility with prior DOS-prediction work and because it reflects absolute reconstruction error. Because DOS quality also depends strongly on relative spectral shape, we additionally emphasize normalized MAE (nMAE) and spectral angle mapper (SAM), which better capture scale-removed spectral fidelity and amplitude-insensitive shape agreement, respectively. For the main comparison models, we report the mean ± standard deviation over three random seeds; secondary ablations are single-run unless otherwise noted. Complete dataset preprocessing and filtering are provided in Appendix B.1, evaluation metrics in Appendix B.2, optimization details in Appendix B.3, and baseline-specific settings in Appendix B.4.

4.2 Comparison under a Unified Evaluation Pipeline

Table 1 compares UMA-DOS with reproduced structure-to-DOS baselines under a unified Materials Project evaluation pipeline. Under this controlled comparison, UMA-DOS achieves the best performance on all three reported metrics. Relative to the strongest reproduced absolute-scale baseline, DOSTransformer, UMA-DOS reduces nMAE from 0.135 to 0.112 (17.0%) and SAM from 0.403 to 0.380 (5.7%), while also achieving the lowest MAE. The clearest gains therefore appear in the shape-sensitive metrics, indicating that the main advantage of UMA-DOS lies in more faithful recovery of DOS structure rather than average pointwise error alone. Figure 5 in Appendix C.1 shows that this nMAE advantage is broad across the test set rather than driven by a small number of favorable structures, and Table 6 in Appendix C.1 shows that the same pattern persists when the comparison is expanded to the full set of reproduced baseline and UMA-DOS variants. We also performed contextual comparison with PET-MAD-DOS in Appendix C.2.

4.3 Low-data Comparison Across Methods

To assess data efficiency under limited DOS supervision, we train UMA−DOS, Mat2Spec, and DOS-Transformer using 10%, 25%, 50%, and 100% of the training set. For each training fraction, all methods use the same training subset and the same fixed validation and test sets. Appendix C.3 gives the full split-construction and fairness protocol. Figure 3 reports test MAE, nMAE, and SAM as the amount of supervision increases.

All three models improve with additional training data. UMA−DOS achieves the strongest nMAE and SAM at every training fraction, with the clearest gains in the lowest-data regime, while MAE remains comparable to the strongest competing baseline at low supervision and becomes best at full data. These results suggest that the MLIP-pretrained backbone provides physically meaningful structural representations before DOS supervision, enabling UMA−DOS to adapt effectively even when only a small number of DOS-labeled samples are available.

4.4 Transfer and Adaptation Ablations

Table 2 compares four adaptation regimes with UMA-s-1p2, weighted pooling, and the spectral decoder fixed. Full fine-tuning performs best, whereas frozen transfer underperforms scratch in MAE and SAM. ELoRA improves all three metrics over frozen transfer, recovering 91.9% of the frozen-to-full MAE gain using 7.37% of full fine-tuning's trainable parameters. Thus, adapting pretrained representations improves spectral fidelity without requiring full parameter updates. ELoRA details appear in Appendix C.1.1; pooling and checkpoint comparisons are provided in Appendices C.1 and C.4.

Table 2: Adaptation with UMA-s-1p2 and weighted pooling. Original regimes: three-seed mean ± standard deviation.

Regime	MAE ↓	nMAE ↓	SAM ↓
Scratch	$3.016_{\pm 0.030}$	$0.118_{\pm 0.001}$	$0.412_{\pm 0.003}$
Frozen transfer	$3.763_{\pm 0.019}$	$0.118_{\pm 0.001}$	$0.447_{\pm 0.001}$
ELoRA ($r = 16$)	$2.658_{\pm 0.001}$	$0.114_{\pm 0.001}$	$0.406_{\pm 0.001}$
Full fine-tuning	$\mathbf{2.561}_{\pm 0.039}$	$\mathbf{0.112}_{\pm 0.003}$	$\mathbf{0.380}_{\pm 0.001}$

4.5 Decoder and Shape–amplitude Factorization Ablations

We next study the contribution of the DOS decoder design under a fixed backbone, pooling operator, and training protocol. Table 3 compares linear, MLP, Conv1D, and spectral decoders using the same fine-tuned UMA-s-1p2 backbone with weighted pooling. The proposed spectral decoder outperforms all simpler alternatives on MAE, nMAE, and SAM, indicating that explicitly structured decoding along the energy axis improves both absolute DOS accuracy and spectral fidelity. Table 4 studies whether explicitly separating normalized spectral shape from global amplitude is beneficial in the same fine-tuned setting. Adding the scale branch improves all reported metrics, indicating that shape–amplitude factorization helps reconstruct the total DOS more faithfully. Appendix C.6 extends both ablations across additional training regimes.

Table 3: Decoder ablation under the same backbone, pooling, and training protocol. Lower is better for all metrics.

Decoder	MAE ↓	nMAE ↓	SAM ↓
Linear readout	$2.725_{\pm 0.007}$	$0.141_{\pm 0.002}$	$0.404_{\pm 0.001}$
MLP decoder	$2.702_{\pm 0.024}$	$0.135_{\pm 0.001}$	$0.400_{\pm 0.001}$
Conv1D decoder	$2.715_{\pm 0.023}$	$0.138_{\pm 0.003}$	$0.403_{\pm 0.002}$
Spectral decoder (ours)	$\mathbf{2.561}_{\pm 0.039}$	$\mathbf{0.112}_{\pm 0.003}$	$\mathbf{0.380}_{\pm 0.001}$

Table 4: Shape–amplitude factorization ablation for the spectral decoder under the fine-tuned setting. Lower is better for all metrics.

Variant	MAE ↓	nMAE ↓	SAM ↓
No scale branch	$2.606_{\pm 0.011}$	$0.136_{\pm 0.001}$	$0.417_{\pm 0.001}$
+ Scale branch (ours)	$\mathbf{2.561}_{\pm 0.039}$	$\mathbf{0.112}_{\pm 0.003}$	$\mathbf{0.380}_{\pm 0.001}$

4.6 Qualitative Spectral Fidelity

Scalar metrics do not fully capture DOS quality because physically important errors are often concentrated at narrow peaks and sharp edges. We therefore complement the quantitative comparison with direct inspection of predicted spectra. Figure 4 compares UMA−DOS with the same DOSTransformer

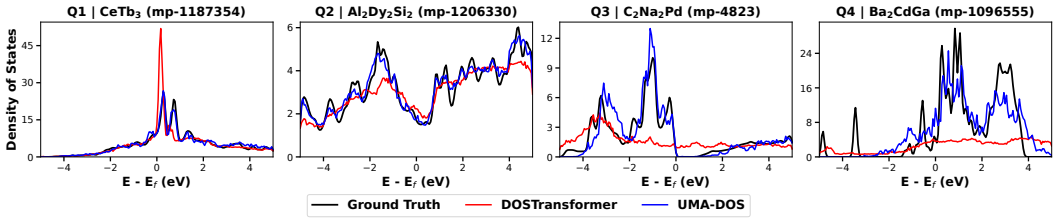


Figure 4: Qualitative comparison of UMA-DOS and DOSTransformer, the same baseline used in Table 1. Panels show one Materials Project test structure from each quartile of pairwise difficulty, defined as the mean per-structure nMAE of the two models. Panels are ordered from Q1 (easiest) to Q4 (hardest); Appendix C.5 gives the selection protocol.

configuration used in Table 1. The four panels span the test-set difficulty range from Q1 (easiest) to Q4 (hardest); the quartile construction and deterministic selection rule are given in Appendix C.5. Figure 12 in Appendix C.7 extends this comparison to four representative examples per quartile. We also provide qualitative comparison with Mat2Spec in Figures 13, 14 and 15 in Appendix C.7.

Across the quartiles, both models usually recover the coarse DOS envelope, but UMA-DOS more faithfully preserves local spectral structure. In Q1, the difference is mainly in local peak heights, valley depths, and fine oscillatory detail. In Q2 and Q3, the gap becomes more visible around sharper features, where DOSTransformer tends to smooth peaks or shift their positions, whereas UMA-DOS better matches the relative spectral shape. Q4 remains challenging for both methods due to narrow peaks and closely spaced extrema, but UMA-DOS still tracks the overall spectral geometry more consistently. Thus, the progression from Q1 to Q4 should be read less as a strictly monotonic visual worsening and more as a shift in the dominant failure mode, from local distortions in easier cases to unresolved high-curvature structure in the hardest cases. This qualitative pattern is consistent with Table 1, where UMA-DOS shows its advantage in the shape-sensitive metrics nMAE and SAM.

5 Conclusions

We introduced UMA-DOS, a transfer-learning framework for total DOS prediction that combines a pretrained UMA backbone with graph pooling, a structured spectral decoder, and explicit shape–amplitude factorization of the DOS target. Under a unified Materials Project evaluation pipeline, the selected UMA-DOS configuration achieves the strongest controlled performance among the reproduced baselines, with the clearest gains in the shape-sensitive metrics nMAE and SAM.

The empirical results support two main conclusions. First, transfer from pretrained atomistic representations is beneficial for total DOS prediction, especially when DOS supervision is limited: UMA-DOS achieves the strongest shape-sensitive performance across the controlled comparison and retains this advantage in the low-data regime. Second, downstream adaptation improves the utility of pretrained representations for DOS prediction. Full fine-tuning achieves the best results, while parameter-efficient transfer methods like ELoRA offer a competitive alternative. Thus, useful adaptation does not require updating every original backbone parameter.

Our ablations further show that architecture matters at the output level. The structured spectral decoder improves DOS fidelity relative to simpler decoder families, and explicit shape–amplitude factorization further improves reconstruction quality in the fine-tuned setting used by the main UMA-DOS model. Taken together, these results suggest that successful transfer to structured spectral targets depends not only on the pretrained encoder but also on a downstream prediction head designed for spectral structure.

These conclusions should be interpreted within the scope of the current study, focusing on minimally smoothed total DOS under a fixed Materials Project split. Within this setting, the evidence supports a clear takeaway: pretrained atomistic encoders can transfer effectively beyond scalar property prediction to total DOS, provided that they are adapted to the spectral task and paired with a shape-aware spectral decoder. A natural next step is to extend this framework from total DOS to partial DOS, where the same transfer-learning questions arise under richer spectral supervision.

AI USE STATEMENT

In this work, we used generative AI tools to draft the manuscript. All other tasks for which disclosure is required were either performed without generative AI assistance or were not applicable to this work. We also used these tools for literature review, language editing, and refinement of the manuscript's structure. We reviewed the entire manuscript to ensure that it accurately reflects our intended scientific claims, and we independently verified all references. We take full responsibility for the final content of this work, including all text and scientific claims.

ETHICS STATEMENT

This work studies electronic density-of-states prediction using publicly available atomistic datasets and pretrained models. It does not involve human participants, personal data, or animal experiments. The intended contribution is methodological and supports computational materials research. Predictions are subject to dataset coverage and approximation errors, and application-specific conclusions require appropriate validation. Limitations and potential downstream impacts are discussed in Appendices D.1 and D.2.

REPRODUCIBILITY STATEMENT

Appendix A describes the architecture and training objective. Appendix B documents dataset preprocessing, evaluation metrics, optimization, and baseline comparison protocols. Appendix C provides additional experiments and configuration details, including ELoRA adaptation, MAD-DOS dataset evaluation, low-data split construction, and qualitative example selection. Compute resources and runtime are reported in Appendix D.3, and code availability is described in Appendix D.4.

REFERENCES

Junwen Bai, Yuanqi Du, Yingheng Wang, Shufeng Kong, John Gregoire, and Carla Gomes. Xtal2dos: Attention-based crystal to sequence learning for density of states prediction, 2023. URL https://arxiv.org/abs/2302.01486.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, and Zachary W. Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models, 2024. URL https://arxiv.org/abs/2410.12771.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, John L. A. Gardner, Mikolaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Cheuk Hin Ho, Stephan Hofmann, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2025. URL https://arxiv.org/abs/2401.00096.

Kamal Choudhary, Kevin F. Garrity, Andrew C. E. Reid, Brian DeCost, Adam J. Biacchi, Angela R. Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A. Gilad Kusne, Andrea Centrone, Albert Davydov, Jie Jiang, Ruth Pachter, Gowoon Cheon, Evan Reed, Ankit Agrawal, Xiaofeng Qian, Vinit Sharma, Houlong Zhuang, Sergei V. Kalinin, Bobby G. Sumpter, Ghanshyam Pilania, Pinar Acar, Subhasish Mandal, Kristjan Haule, David Vanderbilt, Karin Rabe, and Francesca Tavazza. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6(1):173, 2020. 10.1038/s41524-020-00440-1.

Marvin L Cohen and James R Chelikowsky. *Electronic structure and optical properties of semiconductors*. Springer Science & Business Media, 1988.

Yaning Cui, Kang Chen, Lingyao Zhang, Haotian Wang, Lei Bai, David Elliston, and Wei Ren. Atomic positional embedding-based transformer model for predicting the density of states of crystalline materials. *The Journal of Physical Chemistry Letters*, 14(35):7924–7930, 2023.

P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864–B871, November 1964. doi: 10.1103/PhysRev.136.B864.

Wei Bin How, Pol Febrer, Sanggyu Chong, Arslan Mazitov, Filippo Bigi, Matthias Kellner, Sergey Pozdnyakov, and Michele Ceriotti. A universal machine learning model for the electronic density of states, 2026. URL https://arxiv.org/abs/2508.17418.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi: 10.1063/1.4812323. URL http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi.

Prathik R Kaundinya, Kamal Choudhary, and Surya R Kalidindi. Prediction of the electron density of states for crystalline compounds with atomistic line graph neural networks (alignn). *Jom*, 74(4): 1395–1405, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *The Third International Conference on Learning Representations (ICLR 2015)*, 2015.

Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1):15010, 2015. 10.1038/npjcompumats.2015.10.

W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, November 1965. doi: 10.1103/PhysRev.140.A1133.

Shufeng Kong, Francesco Ricci, Dan Guevarra, Jeffrey B Neaton, Carla P Gomes, and John M Gregoire. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nature communications*, 13(1):949, 2022.

Namkyeong Lee, Heewoong Noh, Sungwon Kim, Dongmin Hyun, Gyoung S. Na, and Chanyoung Park. Density of states prediction of crystalline materials via prompt-guided multi-modal transformer. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 61678–61698. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c23fdcb9f8e28af705a87de1375a705c-Paper-Conference.pdf.

Arslan Mazitov, Filippo Bigi, Matthias Kellner, Paolo Pegolo, Davide Tisi, Guillaume Fraux, Sergey Pozdnyakov, Philip Loche, and Michele Ceriotti. Pet-mad, a lightweight universal interatomic potential for advanced materials modeling, 2025a. URL https://arxiv.org/abs/2503.14118.

Arslan Mazitov, Sofiia Chorna, Guillaume Fraux, Marnik Bercx, Giovanni Pizzi, Sandip De, and Michele Ceriotti. Massive atomic diversity: a compact universal dataset for atomistic machine learning. *Scientific data*, 12(1):1857, 2025b.

Shokirbek Shermukhamedov, Dilorom Mamurjonova, Thana Maihom, and Michael Probst. Structure to property: Chemical element embeddings for predicting electronic properties of crystals. *Journal of Chemical Information and Modeling*, 64(15):5762–5770, 2024.

Michael Y. Toriyama, Alex M. Ganose, Maxwell Dylla, Shashwat Anand, Junsoo Park, Madison K. Brod, Jason M. Munro, Kristin A. Persson, Anubhav Jain, and G. Jeffrey Snyder. How to analyse a density of states. *Materials Today Electronics*, 1:100002, 2022. doi: https://doi.org/10.1016/j.mtelec.2022.100002. URL https://www.sciencedirect.com/science/article/pii/S277294942200002X.

Chen Wang, Siyu Hu, Guangming Tan, and Weile Jia. ELoRA: Low-rank adaptation for equivariant GNNs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hcoxm3Vwgy.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. Uma: A family of universal models for atoms, 2026. URL https://arxiv.org/abs/2506.23971.

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Method **4**

3.1 Problem Formulation and Target Parameterization 4

3.2 UMA-DOS Architecture 4

3.2.1 Pooling of UMA output 5

3.2.2 Spectral decoder for DOS prediction 5

3.2.3 Training objective 6

3.2.4 Transfer Regimes 6

4 Experiments **6**

4.1 Experimental Setup 6

4.2 Comparison under a Unified Evaluation Pipeline 7

4.3 Low-data Comparison Across Methods 8

4.4 Transfer and Adaptation Ablations 8

4.5 Decoder and Shape–amplitude Factorization Ablations 8

4.6 Qualitative Spectral Fidelity 8

5 Conclusions **9**

A Method Details **15**

A.1 Architecture details 15

A.2 Training Objective 16

B Experimental Setup and Comparison Protocol **17**

B.1 Dataset Filtering and Preprocessing 17

B.2 Evaluation Metrics 18

B.3 Optimization 19

B.4 Baseline-Specific Settings and Comparison Protocol 19

B.5 Why Literature Results are not Directly Comparable 20

C Additional Results **21**

C.1 Extended Controlled Comparison Across Baseline and UMA-DOS Variants 21

C.1.1 ELoRA configuration and parameter accounting 22

C.2 Comparison with PET-MAD-DOS 23

C.2.1 Contextual evaluation on Materials Project dataset 23

C.2.2 Supervised evaluation on MAD-DOS dataset 24

C.3 Low-data Split Construction and Fairness Protocol 25
C.4 Exploratory Backbone Ablations Across UMA Checkpoints 26
C.5 Qualitative Example Selection Protocol 26
C.6 Extended Decoder and Shape–amplitude Factorization Ablations 27
C.6.1 Decoder Ablations 28
C.6.2 Shape–amplitude Factorization 30
C.7 Extended Qualitative Experiments 31

D Scope **36**
D.1 Limitations and Scope of Claims 36
D.2 Broader Impact 36
D.3 Compute Resources and Runtime 36
D.4 Code Availability 37

A METHOD DETAILS

A.1 ARCHITECTURE DETAILS

Table 5 provides the full architecture of UMA-DOS. Given a crystal structure $\mathcal{C}$ with N atoms, the pretrained UMA encoder f_θ first produces atom-wise embeddings in $\mathbb{R}^{128\times N}$. These variable-length atomic features are then aggregated by a permutation-invariant pooling operator p_ψ to obtain a fixed-dimensional crystal representation $\boldsymbol{z} \in \mathbb{R}^{128}$, where we consider either mean pooling or learnable softmax-weighted pooling. The pooled representation is projected by $\mathrm{MLPproj}$ and reshaped into a latent spectral tensor in $\mathbb{R}^{64\times 181}$, corresponding to 64 feature channels over the fixed grid of 181 energy bins. This latent tensor is refined by three residual 1D convolution blocks with dilation factors 1, 2, and 4, which model local and medium-range correlations along the energy axis. A convolutional readout head then maps the refined spectral features to the normalized DOS prediction $\boldsymbol{y} \in \mathbb{R}^{181} \geq 0$ through a final Softplus activation, while a separate MLP scale head predicts the scalar amplitude parameter a from the pooled crystal embedding $\boldsymbol{z}$.

We use UMA as a publicly available atomistic foundation model family pretrained on large-scale energy and force data. For the Materials Project benchmark, we use the omat task embedding, corresponding to UMA's OMat24 materials setting (Barroso-Luque et al., 2024). This choice matches the crystalline inorganic domain of our benchmark and is held fixed across the UMA transfer, pooling, and checkpoint ablations. UMA therefore provides a common platform for studying these design choices, with UMA-s-1p2 selected using validation performance. Please note that we do not treat the UMA task embedding as a tunable hyperparameter; our goal is to analyze DOS transfer with a fixed materials-oriented pretrained encoder, not to optimize task-tag selection.

Table 5: Detailed architecture of UMA-DOS. The model maps a crystal structure $\mathcal{C}$ to a DOS prediction in four stages: UMA encoding, pooling, latent spectral projection, and residual spectral decoding. We denote by N the number of atoms.

Block	**Output dims**	**Details**
UMA encoder f_θ	$(128, N)$	Pretrained UMA encoder. The encoder outputs atom-wise embeddings.
Pooling p_ψ	128	Aggregates the variable-length set of atomic embeddings into a crystal-level representation $\boldsymbol{z}$. Two variants are considered: (i) mean pooling (ii) softmax-weighted pooling
Projection to spectral dimension $\mathrm{MLP}_{\mathrm{proj}}$ + Reshape	(64, 181)	Maps the pooled crystal embedding to a latent spectral representation over 181 energy bins. $\mathrm{MLP}_{\mathrm{proj}}$: Linear → SiLU → Linear (dims: 128→128→64×181)
Residual convolution block ×3	(64, 181)	Residual 1D convolutional refinement along the energy axis. Each block $\mathcal{B}_\ell$ consists of: dilated Conv1D → SiLU → BatchNorm Dilation factors: $1, 2, 4$ Kernel size: 3
Readout + Softplus activation	181	Maps the refined features to the normalized DOS output $\boldsymbol{y}$. Consists of two 1D convolutional layers: Conv1D → SiLU → Conv1D (dims: 64→32→1) Kernel size: 1, 3
Scale head	1	MLP for predicting the scale factor a from $\boldsymbol{z}$. Linear → SiLU → Dropout → Linear → SiLU → Linear (dims: $128 \to 64 \to 32 \to 1$)

A.2 Training Objective

Our training objective reflects three complementary desiderata: local agreement in spectral shape, accurate prediction of the overall DOS scale, and consistency in cumulative spectral weight.

Critical-point-weighted shape loss To supervise the normalized DOS shape, we use a target-dependent critical-point-weighted ℓ_1 loss. For a DOS spectra with T energy bins, the loss is defined as

$$\mathcal{L}_{\text{shape}} = \frac{1}{T}\sum_{t=1}^{T} w_t \, |\hat{y}_t - y_t|, \tag{6}$$

where $\hat{y}_t$ and y_t denote the predicted and target normalized DOS values, respectively. The weights w_t are computed from the target DOS only, and are therefore independent of the prediction.

The purpose of the weighting is to assign larger importance to spectral regions that are structurally important for the DOS shape, such as sharp peaks, valleys, band edges, and zero-DOS regions. Let

$$c_t = \text{Crit}(\boldsymbol{y})_t \in \{0, 1\} \tag{7}$$

be a binary mask indicating whether bin t is regarded as a critical location. We first assign unnormalized weights

$$u_t = 1 + 3c_t. \tag{8}$$

Thus, non-critical bins receive weight 1, while critical bins receive weight 4 before normalization.

To keep the overall scale of the loss comparable to the ordinary mean absolute error, we normalize these weights over the full target tensor passed to the loss:

$$w_t = u_t \cdot \frac{T}{\sum_{t'=1}^{T} u_{t'}}. \tag{9}$$

This normalization enforces unit mean weight,

$$\frac{1}{T}\sum_{t=1}^{T} w_t = 1, \tag{10}$$

so the weighting changes the relative contribution of different energy bins without changing the overall loss scale.

We next describe how the critical-point mask $\boldsymbol{c} = \text{Crit}(\boldsymbol{y})$ is constructed for a target spectrum $\boldsymbol{y} = [y_1, \ldots, y_T]$. The detector is fixed and is applied only to the target spectrum. Its output is the union of three binary indicators.

First, we mark zero-DOS bins, i.e., band gaps:

$$g_t = \mathbf{1}[y_t = 0]. \tag{11}$$

Second, we mark band-edge bins. These are nonzero bins that are immediately adjacent to a zero-valued bin:

$$\beta_t = \mathbf{1}\big[y_t \neq 0 \ \wedge \ \big((t > 1 \wedge y_{t-1} = 0) \ \vee \ (t < T \wedge y_{t+1} = 0)\big)\big]. \tag{12}$$

Third, we mark turning points based on sign changes in the discrete first derivative. Let

$$\Delta y_t = y_{t+1} - y_t, \qquad t = 1, \ldots, T-1. \tag{13}$$

For interior bins, we define

$$e_t = \mathbf{1}[\text{sign}(\Delta y_{t-1}) \neq \text{sign}(\Delta y_t)], \qquad t = 2, \ldots, T-1, \tag{14}$$

with $e_1 = e_T = 0$. We use $\text{sign}(0) = 0$, so this detector marks not only strict peak and valley transitions, but also transitions between flat and non-flat local slopes.

The final critical-point mask is the union of these three indicators:

$$c_t = \mathbf{1}[g_t \vee \beta_t \vee e_t]. \tag{15}$$

Consequently, the loss places greater emphasis on bins corresponding to zero-DOS regions, band edges, and local extrema or slope-transition points, while preserving the global scale of a standard ℓ_1 shape loss.

Scale loss. To supervise the per-sample DOS magnitude, we regress the log-amplitude with an ℓ_1 loss:

$$\mathcal{L}_{\text{scale}} = |\hat{s} - s|, \tag{16}$$

where $\hat{s}$ and s denote the predicted and target DOS log-amplitudes, respectively.

Cumulative-spectrum loss. Pointwise agreement alone does not constrain whether spectral weight accumulates correctly over energy. We therefore add a cumulative-spectrum term,

$$\mathcal{L}_{\text{cdf}} = \frac{1}{T} \sum_{t=1}^{T} \left| \sum_{\tau=1}^{t} \hat{y}_\tau - \sum_{\tau=1}^{t} y_\tau \right|, \tag{17}$$

which encourages agreement in the integrated spectral weight.

Total objective. The full training loss is

$$\mathcal{L} = \lambda_{\text{shape}} \mathcal{L}_{\text{shape}} + \lambda_{\text{scale}}\, \mathcal{L}_{\text{scale}} + \lambda_{\text{cdf}}\, \mathcal{L}_{\text{cdf}}, \tag{18}$$

where λ_{shape}, λ_{scale}, and λ_{cdf} are weighting hyperparameters, which we set to $(1.0, 0.1, 0.01)$, respectively. The shape term is the main supervision signal, the scale term restores absolute DOS magnitude, and the cumulative term regularizes the global spectral profile. In practice, the final objective is averaged over the mini-batch during training.

B Experimental Setup and Comparison Protocol

B.1 Dataset Filtering and Preprocessing

For each material, we shift the energy axis by the Fermi energy,

$$E' = E - E_F, \tag{19}$$

so that all spectra share a common reference with $E = 0$ corresponding to E_F. The total DOS is then interpolated onto a fixed uniform grid using piecewise cubic Hermite interpolation (PCHIP) over the interval $[-5, 5]$ eV with 181 bins. This produces a fixed-length target vector for every retained material. Any negative interpolation artifacts are clipped to zero.

Dataset filtering. If the original DOS energy range does not fully cover the interpolation interval, the sample is flagged and may be excluded so that all retained materials have valid DOS targets over the full energy window. We further apply several material-level quality filters. First, we retain only materials for which the integrated DOS agrees with the expected valence electron count within 10%, using electron-count agreement as a quality-control heuristic. Second, we remove any material containing a noble-gas element. Third, we exclude SCAN and r^{2}SCAN calculations to obtain a more homogeneous dataset.

Gaussian smoothing. To mitigate the impact of noise in the DFT calculations used when generating the training data, Gaussian smoothing was applied to the targets. Let $\boldsymbol{d} = [d_1, \ldots, d_T] \in \mathbb{R}^T_{\geq 0}$ denote the reference DOS on the fixed grid. We smooth each spectrum by convolving it with a normalized discrete Gaussian kernel. In the implementation, the kernel size is

$$K = \lfloor 6\sigma + 1 \rfloor. \tag{20}$$

Let

$$\mathcal{J} = \{\, 1 - \lceil K/2 \rceil, \ldots, \lfloor K/2 \rfloor \,\}. \tag{21}$$

The kernel coefficients are

$$k_j = \frac{\exp\!\left(-\frac{1}{2}(j/\sigma)^2\right)}{\sum_{\ell \in \mathcal{J}} \exp\!\left(-\frac{1}{2}(\ell/\sigma)^2\right)}, \qquad j \in \mathcal{J}, \tag{22}$$

and the smoothed DOS is

$$\tilde{d}_t = \sum_{j \in \mathcal{J}} k_j\, d_{t-j}, \tag{23}$$

implemented as a one-dimensional convolution. In all experiments reported in the main paper, we evaluated predictions against this smoothed reference DOS.

B.2 Evaluation Metrics

All reported metrics are computed on DOS spectra defined on the common energy grid described in Section 4.1. Unless otherwise stated, the reference DOS used for evaluation is first smoothed with a one-dimensional Gaussian kernel with width $\sigma = 1.0$ in grid units. We predict the normalized spectrum and the log magnitude. Let $\hat{\boldsymbol{y}}$ and $\hat{s}$ denote the predicted normalized spectrum and log magnitude, respectively, and let y and s denote their corresponding targets. The predicted DOS is then reconstructed as $\hat{\boldsymbol{d}} = \exp(\hat{s})\hat{\boldsymbol{y}}$ while the target DOS is written as $\boldsymbol{d} = \exp(s)\boldsymbol{y}$. For these quantities, we consider the following five metrics: three primary metrics and two additional metrics. Each metric is designed to evaluate a different aspect of the prediction error, and we describe them in detail below. We report the dataset mean of these metrics.

Mean Absolute Error (MAE). To evaluate the accuracy of the final spectral prediction, we compute the MAE between target $\boldsymbol{d}$ and predicted DOS $\hat{\boldsymbol{d}}$.

$$\mathrm{MAE}(\hat{\boldsymbol{d}}, \boldsymbol{d}) = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{d}_t - d_t\right|. \tag{24}$$

Normalized MAE (nMAE). Because DOS amplitudes vary substantially across materials, we also report a normalized MAE that emphasizes relative spectral shape. The normalized MAE is defined as:

$$\mathrm{nMAE}(\hat{\boldsymbol{d}}, \boldsymbol{d}) = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{y}_t - y_t\right|. \tag{25}$$

This metric is insensitive to overall scale and is therefore more aligned with relative spectral-shape fidelity.

Spectral Angle Mapper (SAM). We further evaluate the global agreement of spectral shape using the spectral angle mapper (SAM). Unlike the MAE, which measures pointwise absolute deviations in the reconstructed DOS, SAM measures the angular discrepancy between two spectra regarded as vectors in $\mathbb{R}^T$. Given target $\boldsymbol{d}$ and predicted DOS $\hat{\boldsymbol{d}}$, let

$$\cos\alpha = \frac{\langle \hat{\boldsymbol{d}}, \boldsymbol{d}\rangle}{\|\hat{\boldsymbol{d}}\|_2\,\|\boldsymbol{d}\|_2 + 10^{-8}}, \tag{26}$$

with the cosine value clipped to the interval $[-1, 1]$ for numerical stability. The spectral angle mapper is then simply $\alpha \in [0, \pi]$. Because the SAM depends on the angle between the predicted and reference spectra rather than their overall magnitude, lower values indicate better agreement in spectral shape.

NS3. NS3 is a hybrid spectral dissimilarity metric that combines pointwise intensity errors with global shape mismatch. The RMSE term measures the average pointwise deviation between two spectra, while the angular term, $1 - \cos\alpha$, measures the discrepancy in their overall spectral direction. Specifically, let

$$\mathrm{RMSE}(\hat{\boldsymbol{d}}, \boldsymbol{d}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{d}_t - d_t)^2}, \tag{27}$$

and let $\alpha = \mathrm{SAM}(\hat{\boldsymbol{d}}, \boldsymbol{d})$. Then the NS3 score is defined as

$$\mathrm{NS3}(\hat{\boldsymbol{d}}, \boldsymbol{d}) = \sqrt{\mathrm{RMSE}(\hat{\boldsymbol{d}}, \boldsymbol{d})^2 + (1 - \cos\alpha)^2}. \tag{28}$$

Thus, NS3 combines a pointwise Euclidean discrepancy term with an angular mismatch term. Lower values indicate better agreement.

Overlap coefficient (OVL). As an additional diagnostic, we measure the overlap between the predicted and target DOS after nonnegative area normalization. First, the predicted and reference DOS are clipped to be nonnegative and then normalized by their respective areas, yielding area-normalized

spectra $\hat{\boldsymbol{p}}$ and $\boldsymbol{p}$. This normalization removes the effect of the overall DOS magnitude and allows us to compare how the spectral weight is distributed along the energy axis.

Conceptually, for continuous area-normalized spectra, the overlap coefficient is defined as

$$\mathrm{OVL}(\hat{p}, p) = \int \min\{\hat{p}(E), p(E)\}\, dE. \tag{29}$$

At each energy E, the quantity $\min\{\hat{p}(E), p(E)\}$ represents the portion of spectral weight that is common to both the predicted and target spectra. Therefore, integrating this minimum over the energy axis gives the total shared spectral weight between the two area-normalized DOS profiles. In practice, all spectra are represented on the fixed energy grid $\{E_t\}_{t=1}^{T}$, so we evaluate this quantity numerically using the trapezoidal rule:

$$\mathrm{OVL}(\hat{p}, p) \approx \sum_{t=1}^{T-1} \frac{E_{t+1} - E_t}{2} \left[\min(\hat{p}_t, p_t) + \min(\hat{p}_{t+1}, p_{t+1})\right]. \tag{30}$$

A larger OVL indicates that the predicted DOS places spectral weight in similar energy regions as the target DOS, whereas a smaller value indicates poorer agreement in the distribution of spectral weight. Thus, OVL complements pointwise error metrics by focusing on the overlap of the area-normalized spectral profiles rather than on absolute amplitude errors.

Metric interpretation. No single scalar metric fully captures DOS quality. The MAE reflects the absolute reconstruction error but can be dominated by large-amplitude spectra. The nMAE emphasizes the relative spectral shape by removing the scale. SAM is amplitude-insensitive and quantifies angular mismatch between spectra. The overlap coefficient (OVL) measures the fraction of shared spectral weight between the area-normalized prediction and the reference, with values in $[0, 1]$ and higher values indicating better agreement. NS3 combines pointwise and angular discrepancies into a single score. In the main paper, we therefore report MAE, nMAE, and SAM, and complement the quantitative evaluation with a qualitative inspection of predicted spectra; OVL and NS3 are provided as additional diagnostics.

B.3 Optimization

All `UMA-DOS` models are trained with Adam optimizer (Kingma & Ba, 2015) using a batch size of 32 and a maximum of 100 epochs. The learning rate is 1.0×10^{-4} for the spectral head and, when fine-tuning is enabled, 1.0×10^{-4} for the backbone as a separate parameter group. We use a ReduceLROnPlateau scheduler with a factor of 0.8 and patience of 3, monitored on the validation loss. Validation is performed every 1000 optimizer steps, and training is stopped early if validation loss does not improve for 10 consecutive evaluations. Unless otherwise stated, reported results use the checkpoint with the best validation loss.

B.4 Baseline-Specific Settings and Comparison Protocol

Our main claims are based on controlled comparisons under a unified Materials Project evaluation pipeline. Unless explicitly noted otherwise, all reproduced baselines are trained and evaluated on the same Materials Project split, with the same DOS smoothing used for `UMA-DOS`. The rationale for these evaluation settings is provided in Appendix B.5. We use the native validation objective of each method for model selection to avoid post hoc adjustment of checkpoints to a single reported test metric; then, all evaluation metrics are calculated on the selected checkpoint in the same test pipeline.

Representative baseline selection. To keep the main comparison compact, Table 1 reports one representative configuration per baseline family. For Mat2Spec, this representative row is **Mat2Spec-MAE**; for DOSTransformer, it is the **absolute-scale** variant. Within each reported variant, checkpoints are selected by the lowest validation loss. Alternative baseline variants are reported in Table 6.

Mat2Spec variants. We reproduce three Mat2Spec (Kong et al., 2022) training objectives under the unified evaluation pipeline: **Mat2Spec-MAE**, trained with an MAE reconstruction objective; **Mat2Spec-KL**, trained with a KL-divergence objective; and **Mat2Spec-WD**, trained with a Wasserstein-distance objective. All three variants are trained on the same Materials Project split

and evaluated with the same smoothing and metrics as UMA-DOS. Table 6 reports all three variants. The official repository from which the model is taken is https://github.com/gomes-lab/Mat2Spec.

DOSTransformer variants. We reproduce DOSTransformer (Lee et al., 2023) under the unified evaluation pipeline in two forms. The first is an **absolute** scale variant, trained to predict smoothed total DOS directly on its original scale; this is the variant reported in Table 1 and used as the controlled baseline for the prediction of full total DOS. The second is a **peak-normalized** variant which follows the original DOSTransformer target parameterization and is trained on per-spectrum peak-normalized DOS targets. Because this original formulation predicts only a peak-normalized DOS shape, not the absolute DOS scale, it cannot be evaluated as a full total-DOS predictor without using ground-truth test information. We therefore do not evaluate this variant as an absolute-scale DOS predictor and instead report it only as an auxiliary *shape-only* comparison using scale-insensitive metrics, for fairness and completeness. The official repository from which the model is taken is https://github.com/HeewoongNoh/DOSTransformer.

PET-MAD-DOS contextual comparison. PET-MAD-DOS (How et al., 2026) is not a directly controlled baseline, since it is trained on the MAD dataset (Mazitov et al., 2025b) rather than on the Materials Project split used for UMA-DOS, Mat2Spec, and DOSTransformer. We therefore report PET-MAD-DOS only as a contextual comparison. Importantly, we do *not* retrain, fine-tune, or otherwise adapt PET-MAD-DOS on our Materials Project training split. Instead, we use the publicly released model as provided by the authors and run inference directly on the Materials Project test set. Because PET-MAD-DOS does not return predictions for every Materials Project test structure, we evaluate it only on the subset of structures for which it produces valid predictions, and evaluate UMA-DOS on that same subset for reference. As a complementary experiment that removes this training-dataset mismatch, we also train UMA-DOS on MAD-DOS and report the resulting comparison in Appendix C.2.2. The official repository from which the model is taken is https://github.com/lab-cosmo/upet.

B.5 Why Literature Results are not Directly Comparable

The original published settings of prior DOS prediction models differ substantially in target definition, preprocessing, and dataset construction, leading to materially different learning tasks.

Mat2Spec (Kong et al., 2022) operates on a DOS representation aligned to band edges, using a 128-point energy grid over $[-4, 4]$ eV. In this formulation, both the Fermi level and the band edges are shifted to 0 eV, and the zero-valued band-gap region is removed for gapped materials. Mat2Spec also reports multiple scaling and loss settings, including Standard-MAE and SumNorm-KL/WD, which further change the target representation and optimization objective.

DOSTransformer (Lee et al., 2023) uses its own Materials Project-derived preprocessing workflow and, in its original formulation, predicts a peak-normalized DOS target rather than absolute-scale total DOS. Because this representation retains relative spectral shape but discards absolute DOS magnitude, recovering the full total DOS would require external scale information at inference time. For this reason, the literature-native DOSTransformer task is not identical to the unified absolute-scale preprocessing and evaluation protocol used in our controlled experiments.

PET-MAD-DOS (How et al., 2026) is trained on the MAD dataset rather than the Materials Project split used in our experiments, and is additionally evaluated on other public datasets. We therefore treat PET-MAD-DOS as a contextual comparison rather than a directly controlled baseline.

Because these differences materially affect target definition, task difficulty, and evaluation, we do not interpret literature-native results as strictly comparable to UMA-DOS. Instead, our main claims are based on controlled comparisons performed under a unified preprocessing and evaluation pipeline.

C Additional Results

C.1 Extended Controlled Comparison Across Baseline and UMA−DOS Variants

Table 6 expands the main-paper comparison by reporting the full set of reproduced baseline variants together with the internal UMA−DOS variants under the unified Materials Project evaluation pipeline. For Mat2Spec, we report the MAE-, KL-, and Wasserstein-trained variants. For DOSTransformer, the Table reports the absolute-scale target parameterization. For UMA−DOS, we report scratch training, frozen transfer, and fine-tuned transfer, each with mean or weighted pooling, all using the UMA-s-1p2 backbone. All entries use the same split, preprocessing, smoothing, and evaluation metrics as in the main paper, and results are reported as mean ± standard deviation over three random seeds.

Table 6: Extended comparison of DOS predictors under a common evaluation pipeline on the Materials Project test set. Results are reported as mean ± standard deviation over 3 random seeds.

Configuration	MAE ↓	nMAE ↓	SAM ↓	NS3 ↓	Overlap ↑
Literature baselines					
Mat2Spec-MAE	$2.682_{\pm 0.010}$	$0.166_{\pm 0.001}$	$0.439_{\pm 0.001}$	$4.347_{\pm 0.019}$	$0.803_{\pm 0.001}$
Mat2Spec-KL	$2.761_{\pm 0.017}$	$0.152_{\pm 0.002}$	$0.426_{\pm 0.002}$	$4.312_{\pm 0.026}$	$0.806_{\pm 0.001}$
Mat2Spec-WD	$3.663_{\pm 0.019}$	$0.173_{\pm 0.001}$	$0.543_{\pm 0.002}$	$5.279_{\pm 0.030}$	$0.740_{\pm 0.001}$
DOSTransformer (absolute-scale)	$2.579_{\pm 0.012}$	$0.135_{\pm 0.003}$	$0.403_{\pm 0.002}$	$4.082_{\pm 0.014}$	$0.817_{\pm 0.001}$
UMA−DOS variants (UMA-s-1p2 backbone)					
Scratch + mean pooling	$3.016_{\pm 0.032}$	$0.119_{\pm 0.001}$	$0.413_{\pm 0.001}$	$4.706_{\pm 0.036}$	$0.814_{\pm 0.001}$
Scratch + weighted pooling	$3.016_{\pm 0.030}$	$0.118_{\pm 0.001}$	$0.412_{\pm 0.003}$	$4.690_{\pm 0.045}$	$0.814_{\pm 0.001}$
Frozen transfer + mean pooling	$3.753_{\pm 0.037}$	$0.123_{\pm 0.000}$	$0.441_{\pm 0.001}$	$5.685_{\pm 0.043}$	$0.799_{\pm 0.001}$
Frozen transfer + weighted pooling	$3.763_{\pm 0.019}$	$0.118_{\pm 0.001}$	$0.447_{\pm 0.001}$	$5.726_{\pm 0.027}$	$0.801_{\pm 0.001}$
ELoRA ($r = 8$) + mean pooling	$2.715_{\pm 0.011}$	$0.116_{\pm 0.002}$	$0.413_{\pm 0.001}$	$4.318_{\pm 0.007}$	$0.816_{\pm 0.001}$
ELoRA ($r = 8$) + weighted pooling	$2.701_{\pm 0.008}$	$0.114_{\pm 0.001}$	$0.410_{\pm 0.001}$	$4.304_{\pm 0.019}$	$0.816_{\pm 0.001}$
ELoRA ($r = 16$) + mean pooling	$2.681_{\pm 0.012}$	$0.116_{\pm 0.001}$	$0.409_{\pm 0.001}$	$4.266_{\pm 0.022}$	$0.818_{\pm 0.001}$
ELoRA ($r = 16$) + weighted pooling	$2.658_{\pm 0.001}$	$0.114_{\pm 0.001}$	$0.406_{\pm 0.001}$	$4.240_{\pm 0.007}$	$0.818_{\pm 0.001}$
Fine-tuned + mean pooling	$2.584_{\pm 0.016}$	$0.113_{\pm 0.003}$	$0.382_{\pm 0.002}$	$4.138_{\pm 0.020}$	$0.831_{\pm 0.001}$
Fine-tuned + weighted pooling[†]	$\mathbf{2.561}_{\pm 0.039}$	$\mathbf{0.112}_{\pm 0.003}$	$\mathbf{0.380}_{\pm 0.001}$	$\mathbf{4.107}_{\pm 0.044}$	$\mathbf{0.832}_{\pm 0.001}$

[†] Configuration reported as UMA−DOS in Table 1.

Several patterns are consistent with the conclusions of the main paper. Within baseline families, the training objective or target parameterization materially affects the metric profile. Within the UMA−DOS family, the dominant effect is the transfer regime rather than the pooling operator: fine-tuned transfer clearly outperforms scratch training and frozen transfer across the reported metrics, while weighted pooling provides a smaller improvement over mean pooling. The fine-tuned weighted-pooling configuration is the variant reported as UMA−DOS in Table 1.

Table 7 reports the original peak-normalized DOSTransformer formulation as an auxiliary shape-only comparison. Its lower nMAE is expected because both the training target and nMAE remove per-spectrum scale. However, UMA−DOS remains stronger on SAM and OVL while also predicting the full absolute-scale total DOS. We therefore use the absolute-scale DOSTransformer as the main controlled baseline and report the peak-normalized variant only for fairness and completeness.

Per-structure improvement over DOSTransformer (absolute-scale). To test whether the nMAE gain in Table 1 is broad or driven by a small number of favorable structures, Figure 5 plots the

Table 7: Auxiliary shape-only evaluation of the original peak-normalized DOSTransformer formulation. Because this variant predicts only peak-normalized DOS and not absolute DOS magnitude, only scale-insensitive metrics are reported. UMA−DOS is included as a reference under the same shape-only evaluation. Lower is better for nMAE and SAM; higher is better for OVL.

Model	nMAE ↓	SAM ↓	OVL ↑
DOSTransformer (peak-normalized; original)	$0.098_{\pm 0.001}$	$0.394_{\pm 0.004}$	$0.821_{\pm 0.002}$
UMA−DOS (reference)	$0.112_{\pm 0.003}$	$0.380_{\pm 0.001}$	$0.832_{\pm 0.001}$

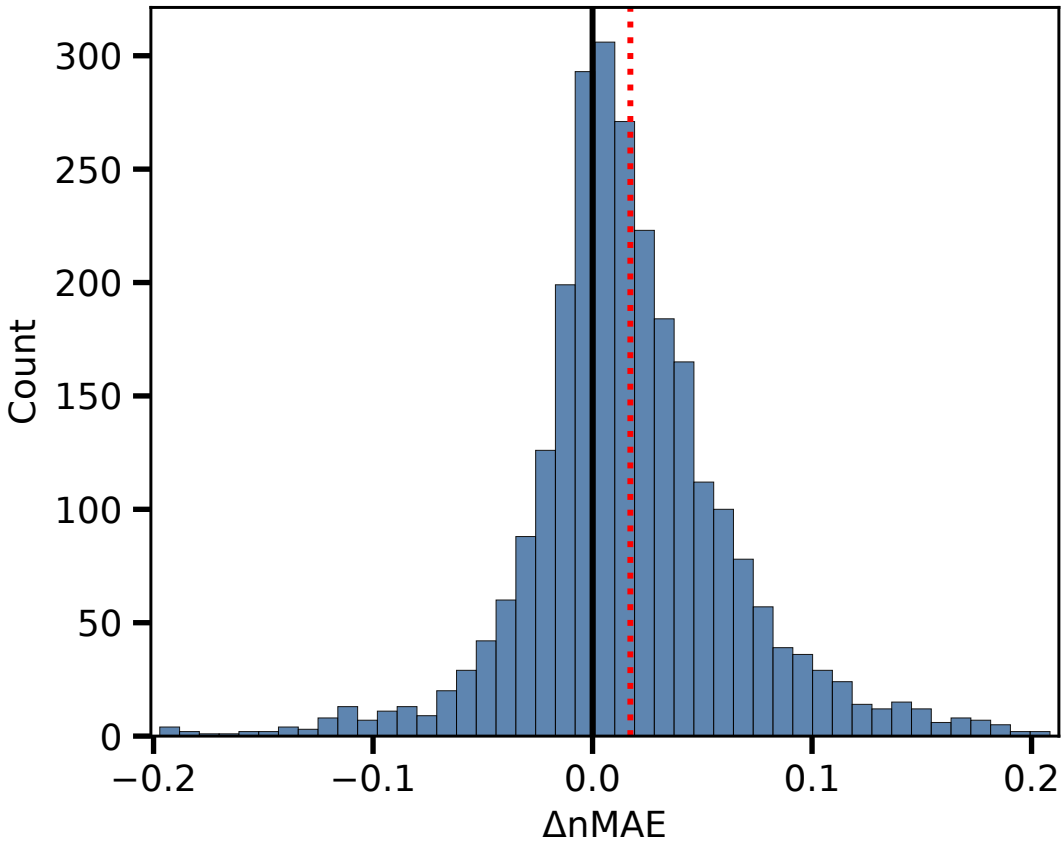


Figure 5: Distribution of per-structure normalized MAE differences between DOSTransformer (absolute-scale) and UMA−DOS on the Materials Project test set under the unified evaluation pipeline. We plot $\Delta\mathrm{nMAE} = \mathrm{nMAE}_{\mathrm{DOSTransformer(abs)}} - \mathrm{nMAE}_{\mathrm{UMA-DOS}}$, so positive values favor UMA−DOS. The solid black line marks parity ($\Delta\mathrm{nMAE} = 0$), and the dashed red line marks the mean difference. UMA−DOS achieves lower per-structure nMAE on 66.3% of test structures, with mean $\Delta\mathrm{nMAE} = 0.017$ and median $\Delta\mathrm{nMAE} = 0.013$, indicating that the shape-sensitive advantage over DOSTransformer (absolute-scale) is broad across the test set rather than driven by a small number of outliers.

per-structure difference $\Delta\mathrm{nMAE} = \mathrm{nMAE}_{\mathrm{DOSTransformer(abs)}} - \mathrm{nMAE}_{\mathrm{UMA-DOS}}$. The distribution is shifted to positive values, indicating that the stronger shape-sensitive performance of UMA−DOS reflects a broad per-structure advantage over the controlled DOSTransformer baseline rather than a few outliers.

C.1.1 ELoRA CONFIGURATION AND PARAMETER ACCOUNTING

We apply rank-8 and rank-16 ELoRA adapters to the 24 MoLE and 8 SO(3) layers of UMA-s-1p2. The original backbone parameters remain fixed, while the adapters, softmax-weighted pooling, and both DOS prediction heads are optimized. The pooling and decoder architectures are identical to those used in the corresponding frozen-transfer and full-fine-tuning configurations. For rank-8 ELoRA, the adapters contain 10.01M trainable parameters. Including weighted pooling and the prediction heads gives 11.57M trainable parameters, corresponding to 3.95% of the parameters updated during full fine-tuning. While for rank-16 ELoRA, the adapters contain 20.02M trainable parameters. Including weighted pooling and the prediction heads gives 21.58M trainable parameters, corresponding to 7.37% of the parameters updated during full fine-tuning. Table 6 reports the complete five-metric comparison.

C.2 Comparison with PET-MAD-DOS

C.2.1 Contextual evaluation on Materials Project dataset

PET-MAD-DOS (How et al., 2026) is not a directly controlled baseline for our main study, since it is trained on MAD-DOS, a newly curated dataset constructed by performing DOS calculations on structures from the MAD dataset (Mazitov et al., 2025b), rather than on the Materials Project split used for `UMA-DOS`, Mat2Spec, and DOSTransformer. Moreover, it does not return predictions for every structure in the Materials Project test set. In our test split, PET-MAD-DOS produces valid predictions for 2476 of 2661 structures (93.0% coverage), leaving 185 structures without predictions. We therefore evaluate PET-MAD-DOS only on this covered subset and evaluate the selected `UMA-DOS` model on the same subset as a reference.

Table 8: Contextual comparison between PET-MAD-DOS and the `UMA-DOS` model on the PET-MAD-DOS-covered subset of the Materials Project test set (2476/2661 structures; 93.0% coverage). PET-MAD-DOS is trained on the MAD dataset rather than on the Materials Project split used in the main controlled comparisons, so this result should be interpreted as contextual rather than directly comparable. Both models are evaluated on the same covered subset. Metrics are computed per structure with smoothing ($\sigma = 1$). Lower is better for MAE, nMAE, SAM, and NS3; higher is better for Overlap.

Model	MAE ↓	nMAE ↓	SAM ↓	NS3 ↓	Overlap ↑
PET-MAD-DOS	4.723	0.230	0.683	7.565	0.678
`UMA-DOS`	**2.685**	**0.105**	**0.396**	**4.283**	**0.823**

Table 8 reports the resulting comparison on the PET-MAD-covered subset of the Materials Project test set (2476 of 2661 structures). Figure 6 complements the table by showing the distribution of per-structure nMAE on the same subset; the `UMA-DOS` distribution is substantially shifted toward lower error, indicating that the improvement is broad rather than driven by a small number of outliers.

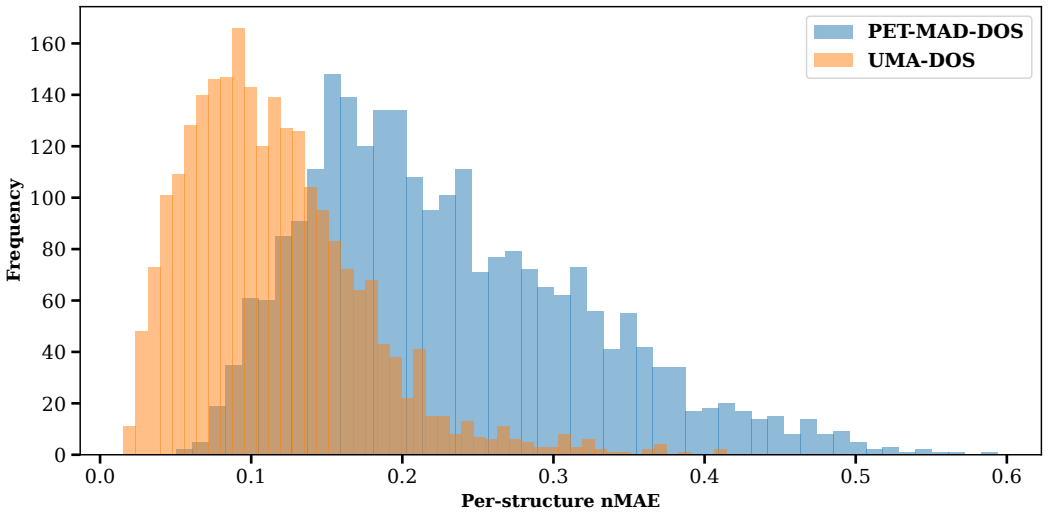


Figure 6: Distribution of per-structure normalized MAE (nMAE) for PET-MAD-DOS and the selected `UMA-DOS` model on the subset of the Materials Project test set for which PET-MAD-DOS returns valid predictions (2476/2661 structures; 93.0% coverage). The remaining 185 structures are excluded from both models in this comparison. Lower is better. The `UMA-DOS` distribution is shifted substantially toward lower error and has a lower mean nMAE, indicating that the improvement over PET-MAD-DOS is broad across the covered subset rather than driven by a small number of outliers. Because PET-MAD-DOS is trained on the MAD dataset rather than on the Materials Project split used in the main paper, this comparison is contextual rather than directly controlled.

Figure 7, included as a qualitative illustration, shows randomly selected examples from the same covered subset and exhibits the same trend at the spectral level: UMA-DOS more closely follows the ground-truth DOS, especially around sharp peaks, local oscillatory structure, and near-Fermi features, whereas PET-MAD-DOS often oversmooths the spectrum or introduces broad spurious peaks. Because PET-MAD-DOS is trained on a different dataset and does not cover the full Materials Project test set, we interpret these results as contextual evidence. For Figure 7, examples are drawn only from the PET-MAD-covered subset of the Materials Project test set and are sampled randomly using a fixed random seed.

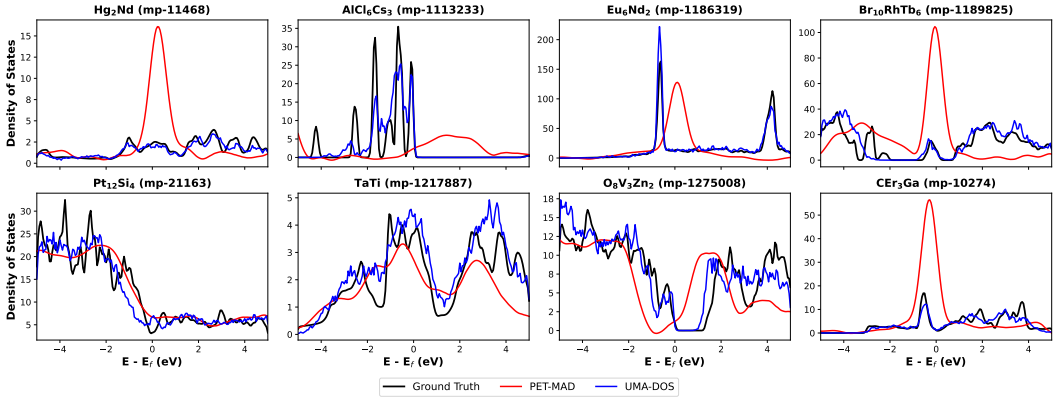


Figure 7: Qualitative comparison of predicted DOS spectra for PET-MAD-DOS and UMA-DOS on random examples from the PET-MAD-covered subset of the Materials Project test set. Black curves denote the ground-truth DOS, red curves denote PET-MAD-DOS, and blue curves denote UMA-DOS; panel titles give the chemical formula and Materials Project identifier. Across these illustrative cases, UMA-DOS more closely follows the ground-truth spectral profile, especially around sharp peaks, local oscillatory structure, and features near the Fermi level, whereas PET-MAD-DOS often oversmooths the spectrum or introduces broad spurious peaks.

C.2.2 Supervised evaluation on MAD-DOS dataset

To complement the contextual comparison on Materials Project, we additionally evaluate UMA-DOS on MAD-DOS. We retain the existing MAD-DOS splits, comprising 72,805 training, 9,097 validation, and 9,110 test structures. For UMA-DOS training, the native DOS targets containing 4,606 energy points are linearly interpolated onto a uniform grid of 181 points over $[-5, 5]$ eV. Because the supplied targets are already smoothed, no additional Gaussian smoothing is applied.

We retain the reference validity masks supplied with MAD-DOS throughout training and evaluation. During UMA-DOS training, excluded bins do not contribute to either DOS-shape or cumulative-DOS loss, and the peak used for target normalization is computed from valid bins only. Although the model predicts all 181 bins, excluded regions do not receive direct DOS supervision for the corresponding structure. At evaluation, UMA-DOS and both PET-MAD-DOS variants are scored using the same reference mask for each structure. The mask excludes bins for 6,908 of the 9,110 test structures; consequently, the reported metrics describe reconstruction over each structure's valid reference region within the selected energy window.

We compare UMA-DOS with raw and denoised predictions obtained from pretrained PET-MAD-DOS weights (How et al., 2026), without retraining PET-MAD-DOS. Table 9 summarizes the comparison under this common masked evaluation protocol.

Under this evaluation protocol, UMA-DOS outperforms both PET-MAD-DOS variants across all five metrics. Relative to raw PET-MAD-DOS, UMA-DOS reduces MAE by approximately 36.4%, nMAE by 37.7%, and SAM by 36.7%, while increasing spectral overlap from 0.842 to 0.902. These improvements indicate better absolute DOS reconstruction and spectral-shape agreement over the valid reference regions. The denoised PET-MAD-DOS predictions perform worse than the raw predictions under the reported metrics. These results extend our evaluation beyond Materials Project

Table 9: Comparison on the MAD-DOS test split using the same reference validity mask for all predictors. UMA−DOS is trained on the adapted MAD-DOS targets, while PET-MAD-DOS uses pretrained weights. Lower is better for MAE, nMAE, SAM, and NS3; higher is better for overlap (OVL).

Model	MAE ↓	nMAE ↓	SAM ↓	NS3 ↓	OVL ↑
PET-MAD-DOS (raw)	2.460	0.120	0.314	3.294	0.842
PET-MAD-DOS (denoised)	2.679	0.120	0.324	3.641	0.835
UMA−DOS	**1.565**	**0.075**	**0.199**	**2.148**	**0.902**

and support the applicability of UMA−DOS to a second DOS dataset. The comparison evaluates the predictors under a common masked scoring protocol.

Figure 8 complements the quantitative comparison with eight selected MAD-DOS test structures. Within the valid reference regions, UMA−DOS closely reproduces the reference DOS in most displayed cases, capturing the dominant peak positions, amplitudes, and overall spectral profiles. Predictions outside the reference validity mask are displayed for completeness, and do not contribute to the reported metrics.

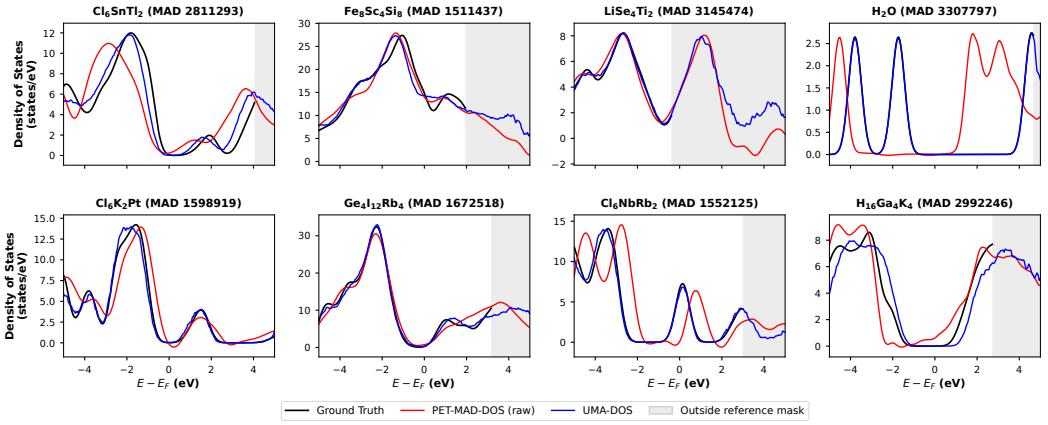


Figure 8: Qualitative DOS comparison for eight randomly selected structures from the MAD-DOS test split. Black curves denote the reference DOS, red curves denote raw PET-MAD-DOS predictions, and blue curves denote UMA−DOS predictions. Energies are displayed relative to the Fermi level, $E - E_F$, and DOS is reported in states/eV. Grey shading marks energies outside each structure's reference-validity mask.

The target construction is also relevant to interpreting these results. MAD-DOS uses Gaussian broadening with $\sigma = 0.3$ eV (How et al., 2026), which attenuates narrow spectral features and can merge closely spaced peaks. Several of the reference spectra shown here consequently exhibit broad, smooth profiles, whereas the Materials Project examples in the main paper display more pronounced local spectral structure. This distinction motivates our use of Materials Project as the primary benchmark for evaluating the reconstruction of sharp peaks and fine spectral detail. The MAD-DOS evaluation provides complementary evidence of the applicability of UMA−DOS under a different target construction and dataset distribution.

C.3 Low-data Split Construction and Fairness Protocol

The low-data comparison in Figure 3 of the main text uses the same fixed Materials Project train/validation/test split as the rest of the paper. For each training fraction $f \in \{10, 25, 50, 100\}$, we construct the reduced training set from the full training split using a single fixed random permutation of the training examples. The subset for fraction f is then obtained by retaining the first $f\%$ of that permutation. The same reduced training subsets are reused for UMA−DOS, Mat2Spec, and DOSTransformer. The validation and test sets remain unchanged across all fractions and all methods.

To control optimization variance, each method at each training fraction is trained with the same three random seeds. These seeds define three independent training runs per method and fraction and are shared across methods, so the reported means reflect matched variability under a common split and training protocol rather than differences in subset construction. Checkpoints are selected using the same validation-based model selection rule as in the main paper.

This protocol ensures that the low-data comparison isolates differences between model families rather than differences in which training examples were observed during training. In particular, performance differences in Figure 3 cannot be attributed to one method receiving an easier reduced training subset than another.

Subset construction detail. The 10%, 25%, and 50% training subsets are nested: each smaller subset is contained within every larger subset. Thus, increasing the training fraction adds examples without replacing previously seen ones, making the low-data scaling curves directly comparable across fractions.

C.4 Exploratory Backbone Ablations Across UMA Checkpoints

Table 10 extends the main-paper transfer and pooling study to additional UMA checkpoints under the same Materials Project split and unified evaluation pipeline. Here, `s` and `m` denote smaller and medium UMA backbone families, while `1p1` and `1p2` denote different pretrained checkpoint versions. Two observations are most relevant.

First, checkpoint identity matters materially for DOS transfer, and larger backbone size does not automatically translate into better downstream DOS prediction. Among the explored settings, UMA-s-1p2 remains the strongest overall checkpoint: fine-tuned transfer with weighted pooling gives the best MAE and NS3 in this table, ties for the best nMAE, and matches the best Overlap. By contrast, UMA-s-1p1 shows weaker transfer, with only limited MAE improvement under fine-tuning, although its fine-tuned variants improve shape-sensitive metrics relative to scratch and frozen transfer. The medium backbone UMA-m-1p1 does not overturn this conclusion. Under weighted pooling, scratch and frozen transfer remain clearly weaker than fine-tuning, so the qualitative transfer-regime conclusion remains unchanged. However, despite its larger capacity, fine-tuned UMA-m-1p1 does not outperform UMA-s-1p2 overall: it matches the best SAM in this single-run table, but remains worse on MAE, nMAE, and NS3, with nearly identical Overlap. Taken together, these exploratory results reinforce that DOS transfer quality depends more on checkpoint identity and end-to-end adaptation than on nominal model size alone.

Second, pooling remains a secondary and non-uniform factor. Where both pooling variants are available, weighted pooling sometimes improves nMAE, SAM, or Overlap, but the gains are modest and not universal across backbones or transfer regimes. This is consistent with the main paper's seeded ablations, where the dominant effect comes from the transfer regime rather than the readout alone.

Overall, these exploratory results support selecting UMA-s-1p2 with weighted pooling as the default `UMA-DOS` configuration, while also indicating that checkpoint version is more consequential than scaling from UMA-s to UMA-m within this limited exploratory study.

C.5 Qualitative Example Selection Protocol

To make the qualitative figures reproducible, we use explicit selection rules for each family of qualitative comparisons. Unless otherwise stated, all examples are drawn from the fixed Materials Project test set and panels are ordered from Q1 (easiest) to Q4 (hardest).

Pairwise baseline comparisons. For any qualitative comparison between `UMA-DOS` and a baseline variant b, including the main-paper DOSTransformer comparison and the appendix galleries for DOSTransformer and Mat2Spec variants, we define the per-structure difficulty of a test structure x as

$$\delta_b(x) = \frac{1}{2}\Big(\mathrm{nMAE}_{\text{UMA-DOS}}(x) + \mathrm{nMAE}_b(x)\Big). \tag{31}$$

We sort the test structures by $\delta_b(x)$, and partition the sorted list into four equal-sized quartiles. We display the four structures within each quartile whose difficulty is closest to the median of that

Table 10: Exploratory backbone ablations for UMA-DOS on the Materials Project test set. Results are single-run evaluations under the same split and unified evaluation pipeline used in the main paper, and should therefore be interpreted as directional trends rather than seeded estimates. For UMA-m-1p1, we report only weighted pooling, since the main seeded ablations and the smaller-checkpoint exploratory results indicate that pooling effects are secondary relative to transfer regime and checkpoint choice. Lower is better for MAE, nMAE, SAM, and NS3; higher is better for Overlap.

Backbone	Configuration	MAE ↓	nMAE ↓	SAM ↓	NS3 ↓	Overlap ↑
UMA-s-1p1	Scratch + mean pooling	3.007	0.118	0.411	4.687	0.815
UMA-s-1p1	Scratch + weighted pooling	3.043	0.117	0.414	4.725	0.814
UMA-s-1p1	Frozen transfer + mean pooling	4.742	0.131	0.469	6.941	0.789
UMA-s-1p1	Frozen transfer + weighted pooling	4.743	0.126	0.402	4.951	0.791
UMA-s-1p1	Fine-tuned + mean pooling	3.078	0.114	0.389	4.818	0.827
UMA-s-1p1	Fine-tuned + weighted pooling	3.086	0.114	0.389	4.830	0.827
UMA-s-1p2	Scratch + mean pooling	3.048	0.119	0.414	4.742	0.814
UMA-s-1p2	Scratch + weighted pooling	3.033	0.117	0.411	4.714	0.814
UMA-s-1p2	Frozen transfer + mean pooling	3.769	0.123	0.451	5.732	0.798
UMA-s-1p2	Frozen transfer + weighted pooling	3.789	0.119	0.437	5.724	0.801
UMA-s-1p2	Fine-tuned + mean pooling	2.571	**0.113**	0.395	4.296	**0.832**
UMA-s-1p2	Fine-tuned + weighted pooling[†]	**2.519**	**0.113**	**0.381**	**4.057**	**0.832**
UMA-m-1p1	Scratch + weighted pooling	3.036	0.121	0.413	4.722	0.814
UMA-m-1p1	Frozen transfer + weighted pooling	4.597	0.131	0.460	6.725	0.792
UMA-m-1p1	Fine-tuned + weighted pooling	3.002	0.120	**0.381**	4.681	0.831

[†] Single-run realization of the configuration reported as UMA-DOS in the main paper; values therefore differ from the 3-seed mean reported in Table 1.

quartile. Because δ_b depends on the comparator, the quartile boundaries are recomputed separately for each baseline variant.

Decoder comparison across transfer regimes. The frozen-transfer and fine-tuned decoder figures are intended as a matched comparison across transfer regimes and therefore use the same four test structures. To define these shared examples, we construct a single reference difficulty stratification in the fine-tuned setting:

$$\delta_{\mathrm{dec}}(x) = \frac{1}{4}\Big(\mathrm{nMAE}^{\mathrm{FT}}_{\mathrm{Linear}}(x) + \mathrm{nMAE}^{\mathrm{FT}}_{\mathrm{MLP}}(x) + \mathrm{nMAE}^{\mathrm{FT}}_{\mathrm{Conv1D}}(x) + \mathrm{nMAE}^{\mathrm{FT}}_{\mathrm{Spectral}}(x)\Big). \quad (32)$$

We sort test structures by $\delta_{\mathrm{dec}}(x)$ and partition them into four equal-sized quartiles. We then sample one structure from each quartile and reuse those same four structures in both the frozen-transfer and fine-tuned figures. Accordingly, the Q1–Q4 labels in the frozen-transfer figure refer to this shared fine-tuned stratification rather than quartiles recomputed under frozen transfer.

Shape–amplitude factorization across transfer regimes. For the qualitative shape–amplitude factorization figure, we again use a shared stratification so that the same structures can be compared across scratch, frozen, and fine-tuned training. We define

$$\delta_{\mathrm{scale}}(x) = \frac{1}{2}\Big(\mathrm{nMAE}^{\mathrm{FT}}_{\mathrm{no\ scale}}(x) + \mathrm{nMAE}^{\mathrm{FT}}_{+\mathrm{scale}}(x)\Big), \quad (33)$$

partition the test structures into quartiles by $\delta_{\mathrm{scale}}(x)$, select one structure for each quartile, and reuse the same four structures across all rows.

C.6 Extended Decoder and Shape–amplitude Factorization Ablations

The main paper studies decoder choice and shape–amplitude factorization in the fine-tuned UMA-s-1p2 setting with weighted pooling. Tables 11 and 12 extend those ablations across additional training regimes under the same Materials Project split and unified evaluation pipeline. Because these are secondary ablations, all results in this section are single-run evaluations and should be interpreted as directional rather than seeded estimates.

Table 11: Extended decoder ablation for UMA-DOS under a fixed UMA-s-1p2 backbone and weighted pooling reported for one training run with a single random seed. We compare four decoder families under frozen transfer and fine-tuned transfer. Lower is better for MAE, nMAE, SAM, and NS3; higher is better for Overlap.

Decoder	MAE ↓	nMAE ↓	SAM ↓	NS3 ↓	Overlap ↑
Frozen transfer					
Linear readout	4.684	0.255	0.534	6.734	0.752
MLP decoder	3.854	0.196	0.475	5.799	0.783
Conv1D decoder	4.298	0.234	0.551	6.285	0.746
Spectral decoder	3.789	0.119	0.437	5.724	0.801
Fine-tuned					
Linear readout	2.720	0.138	0.404	4.302	0.819
MLP decoder	2.705	0.138	0.402	4.285	0.821
Conv1D decoder	2.741	0.143	0.407	4.345	0.818
Spectral decoder (ours)	**2.519**	**0.113**	**0.381**	**4.057**	**0.832**

C.6.1 Decoder Ablations

Decoder Ablation Across Transfer Regimes. Table 11 extends the ablation of the main decoder on paper by comparing the same four decoder families in both frozen transfer and fine-tuned transfer, while keeping the UMA-s-1p2 backbone and the weighted pool fixed. Unlike the main paper comparison, which reports only the fine-tuned setting, this appendix table examines whether the benefit of structured spectral decoding persists when the pretrained backbone is not end-to-end adapted. The answer is yes. The proposed residual dilated spectral decoder is the strongest decoder

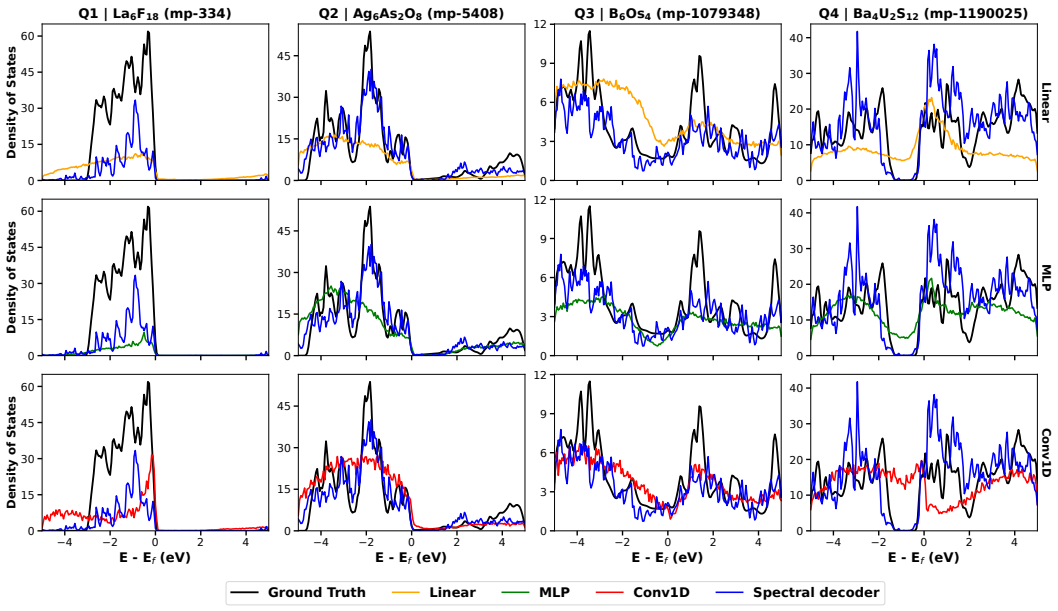


Figure 9: Qualitative decoder comparison under frozen transfer for a fixed UMA-s-1p2 backbone with weighted pooling. Panels show four representative Materials Project test structures selected once from the shared fine-tuned decoder-difficulty quartiles and reused for matched comparison across transfer regimes. Black curves denote the smoothed ground-truth DOS, blue curves denote the proposed spectral decoder, and orange, green, and red curves denote Linear, MLP, and Conv1D decoders, respectively. The frozen transfer spectral decoder more consistently preserves local peak structure, valley depth, and low-DOS regions than the simpler decoder heads.

in both transfer regimes, achieving the best value on every reported metric. The improvement is especially pronounced under frozen transfer, where simpler linear, MLP, and Conv1D decoders are

substantially worse on the shape-sensitive metrics nMAE and SAM. Under fine-tuned transfer, the MAE gaps are smaller, but the spectral decoder still gives the best overall performance and remains clearly strongest on nMAE and SAM.

This pattern is consistent with the broader empirical picture in the paper: the main gains from the proposed decoder arise in preserving relative spectral structure rather than only reducing absolute pointwise error. Because these experiments are single-run evaluations under the fixed Materials Project split, we interpret them as directional support rather than seeded estimates. Nevertheless, together with the main-paper decoder ablation, they strengthen the conclusion that explicit structured decoding along the energy axis is an important part of the `UMA-DOS` design rather than an effect that appears only in the fine-tuned setting.

Qualitative Decoder Comparisons. Figures 9 and 10 complement Table 11 by comparing the proposed spectral decoder against linear, MLP, and Conv1D alternatives under a fixed UMA-s-1p2 backbone and weighted pooling. To facilitate a matched comparison across transfer regimes, both figures use the same four Materials Project test structures. These structures are selected once from a shared reference difficulty stratification defined in the fine-tuned setting, with one example chosen from each quartile, and are then reused in the frozen-transfer figure. Consequently, the Q1–Q4 labels in Figure 9 refer to this shared fine-tuned stratification rather than to quartiles recomputed under frozen transfer. This design ensures that visual differences can be attributed to decoder family and transfer regime rather than to changes in the example set.

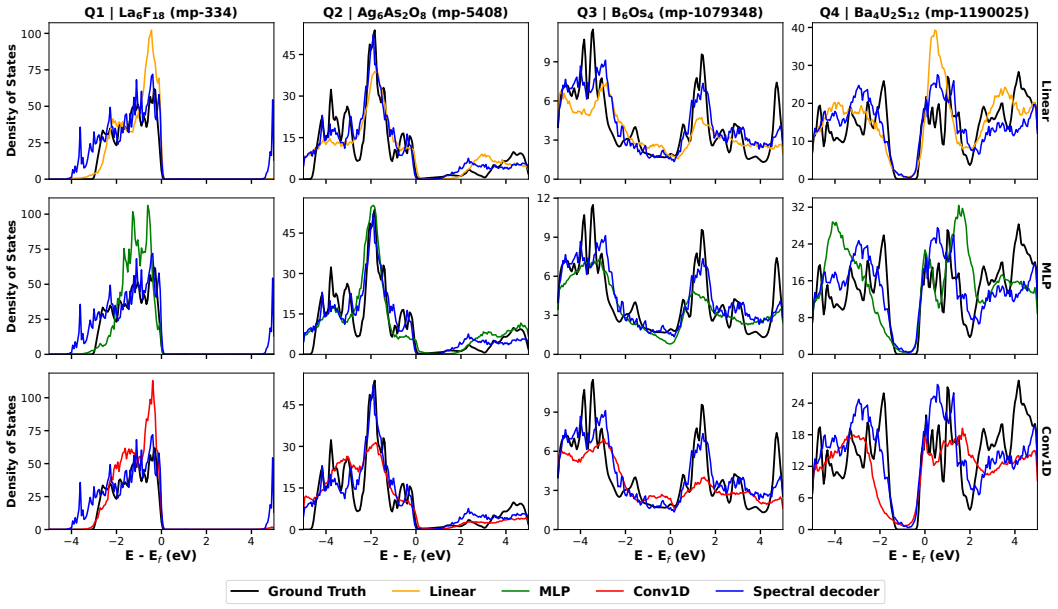


Figure 10: Qualitative decoder comparison under finetuned transfer for a fixed UMA-s-1p2 backbone with weighted pooling. Panels show four representative Materials Project test structures, one from each per-structure difficulty quartile. Black curves denote the smoothed ground-truth DOS, blue curves denote the proposed fine-tuned spectral decoder, and orange, green red curves denote Linear, MLP and Conv1D decoders respectively. The fine-tuned spectral decoder more consistently preserves local peak structure, valley depth, and low-DOS regions than the simpler decoder heads.

Frozen Transfer. In frozen transfer, the qualitative gap between decoder families is large. The linear and Conv1D heads often recover only a coarse spectral envelope, smoothing away narrow peaks, underestimating valley depth, and distorting low-DOS or gap-adjacent regions. The MLP decoder is stronger, but still misses several local extrema and short-range oscillations that the spectral decoder tracks more faithfully. Across the four examples, the proposed spectral decoder remains the most consistent in preserving local peak placement and rapidly varying structure along the energy axis. This visual pattern matches Table 11, where the spectral decoder is the strongest decoder under

frozen transfer on every reported metric, with especially large gains on the shape-sensitive metrics nMAE and SAM.

Fine-tuned. After end-to-end fine-tuning of the UMA backbone, all decoder families improve, but the spectral decoder still gives the most faithful reconstruction of local spectral structure. Relative to the linear, MLP, and Conv1D alternatives, it more consistently preserves sharp peaks, valley depth, and oscillatory behavior while better respecting low-density regions around the Fermi level. The remaining advantage is more pronounced in spectral shape than in coarse envelope matching, which is consistent with Table 11: the MAE gaps narrow under fine-tuning, but the spectral decoder still achieves the best overall performance and remains clearly strongest on nMAE and SAM. Taken together, the frozen and fine-tuned figures support the conclusion that explicit structured decoding contributes beyond the effect of backbone adaptation alone.

Table 12: Extended shape–amplitude factorization ablation for `UMA-DOS` across scratch training, frozen transfer, and fine-tuned transfer evaluated on one training run. We compare the spectral decoder with and without the explicit scale branch under a fixed UMA-s-1p2 backbone and weighted pooling. Lower is better for MAE, nMAE, SAM, and NS3; higher is better for Overlap.

Variant	MAE ↓	nMAE ↓	SAM ↓	NS3 ↓	Overlap ↑
Scratch					
No scale branch	3.230	0.141	0.439	4.918	0.807
+ Scale branch	3.033	0.117	0.411	4.714	0.813
Frozen					
No scale branch	3.799	0.140	0.503	5.763	0.782
+ Scale branch	3.789	0.119	0.438	5.724	0.801
Fine-tuned					
No scale branch	2.599	0.138	0.419	4.411	0.820
+ Scale branch (ours)	**2.519**	**0.113**	**0.381**	**4.057**	**0.832**

C.6.2 Shape–amplitude Factorization

Shape–amplitude Factorization Across Training Regimes. Table 12 compares the spectral decoder with and without the explicit scale branch under scratch training, frozen transfer, and fine-tuned transfer. Under fine-tuned transfer, the setting used by the selected `UMA-DOS` model in the main paper, adding the scale branch, improves all reported metrics, consistent with the main text conclusion that explicit shape–amplitude factorization improves total DOS reconstruction. In scratch training, the scale branch also shows a clear improvement.

Together, these results in the Appendix strengthen the design rationale of `UMA-DOS`. The spectral decoder is a robust improvement across transfer regimes, whereas the gains from explicit shape–amplitude factorization are strongest in the end-to-end fine-tuned setting used in the main paper.

Qualitative Effect of Shape–amplitude Factorization. Figure 11 complements Table 12 by comparing the spectral decoder with and without the explicit scale branch under scratch training, frozen transfer, and fine-tuned transfer on the same four representative Materials Project test structures. One structure is shown from each difficulty quartile, and the same structures are reused across all three training regimes so that visual differences can be attributed to the scale branch and transfer regime rather than to changes in the example set. Across these shared examples, the main qualitative effect of the scale branch is improved calibration of the overall DOS magnitude, along with more faithful preservation of relative peak prominence and valley depth. Without the scale branch, the decoder more often under- or over-scales the spectrum and correspondingly distorts local spectral structure; with the scale branch, the model can treat global magnitude and normalized shape separately, leading to a closer reconstruction of the total DOS. This benefit is most visible in the scratch and fine-tuned settings, where the +scale branch model more consistently matches sharp peaks, low-DOS regions, and near-Fermi behavior, while the frozen-transfer differences are smaller and more case-dependent. Overall, the figure supports the same conclusion as Table 12: explicit shape–amplitude factorization improves total-DOS reconstruction, with the clearest gains in the fine-tuned regime used by the main `UMA-DOS` model.

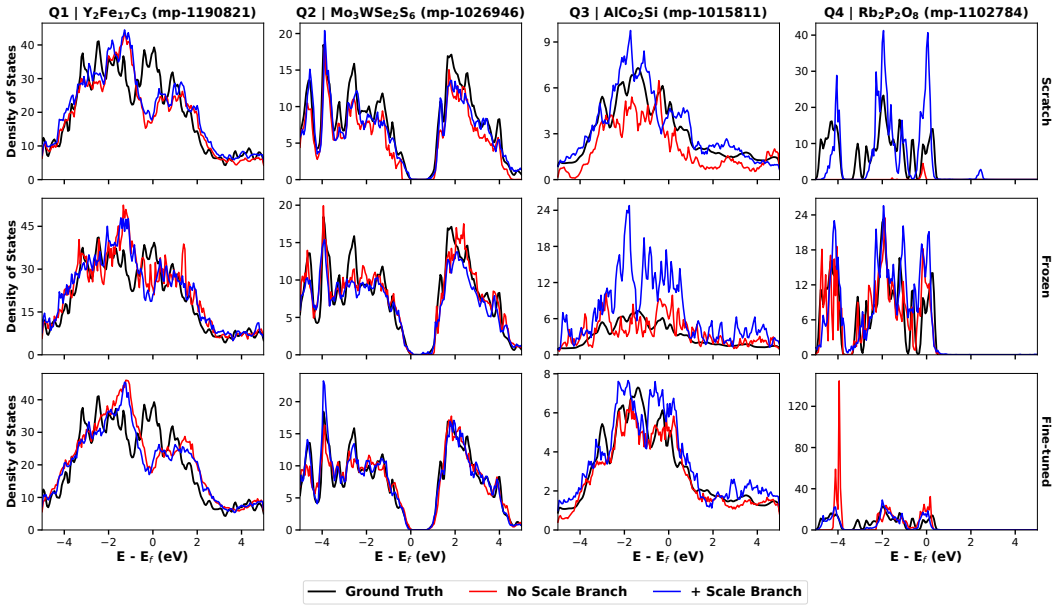


Figure 11: Qualitative effect of shape–amplitude factorization across training regimes. Rows correspond to scratch training, frozen transfer, and fine-tuned transfer; columns show one representative test structure from each difficulty quartile. Black curves denote the ground-truth DOS, red curves denote the variant without the explicit scale branch, and blue curves denote the model with the scale branch. Quartiles are defined once using the fine-tuned pair, and the same four structures are reused across all rows. The explicit scale branch most clearly improves total DOS reconstruction in the fine-tuned setting, while also reducing amplitude miscalibration and local spectral distortion in scratch and frozen transfer.

C.7 Extended Qualitative Experiments

This section expands the qualitative analysis from Section 4.6. Because the main paper can only show a small number of examples, we provide larger galleries that make the selection protocol explicit and illustrate how qualitative behavior varies across the test set.

Figure 12 extends the main-paper Figure 4 to a larger gallery using DOSTransformer (absolute-scale). Across these examples, UMA-DOS more consistently tracks the ground-truth DOS, especially around narrow peaks, local oscillations, and near-Fermi structure. The difference is clearest in the intermediate quartiles, where DOSTransformer often captures the coarse spectral envelope but smooths or distorts finer local features. In the hardest quartile, both models remain challenged, but UMA-DOS still better preserves the overall spectral geometry and relative peak structure. This qualitative pattern is consistent with the stronger shape-sensitive performance of UMA-DOS in quantitative comparisons.

Figure 13 extends the qualitative comparison to Mat2Spec-MAE using the same quartile-based selection protocol. In the easier quartiles, both models often recover the broad DOS profile, but UMA-DOS more faithfully tracks local peak heights, valley depth, and short-range spectral structure. The gap becomes more visible in the intermediate and hard quartiles, where Mat2Spec-MAE more frequently smooths narrow features, broadens sharp peaks, or distorts near-Fermi behavior, whereas UMA-DOS remains more consistent in preserving the underlying spectral geometry. Both models remain challenged on the hardest spectra, but the examples overall are qualitatively consistent with the stronger quantitative performance of UMA-DOS under the unified evaluation pipeline.

Figure 14 extends the qualitative comparison to Mat2Spec-KL using four examples from each difficulty quartile. In the easier quartiles, both models often recover the broad DOS profile, but UMA-DOS more consistently matches local peak heights, valley depth, and short-range spectral structure. The difference becomes more visible in the intermediate and hard quartiles, where Mat2Spec-KL more frequently smooths narrow features, broadens sharp peaks, or distorts near-Fermi

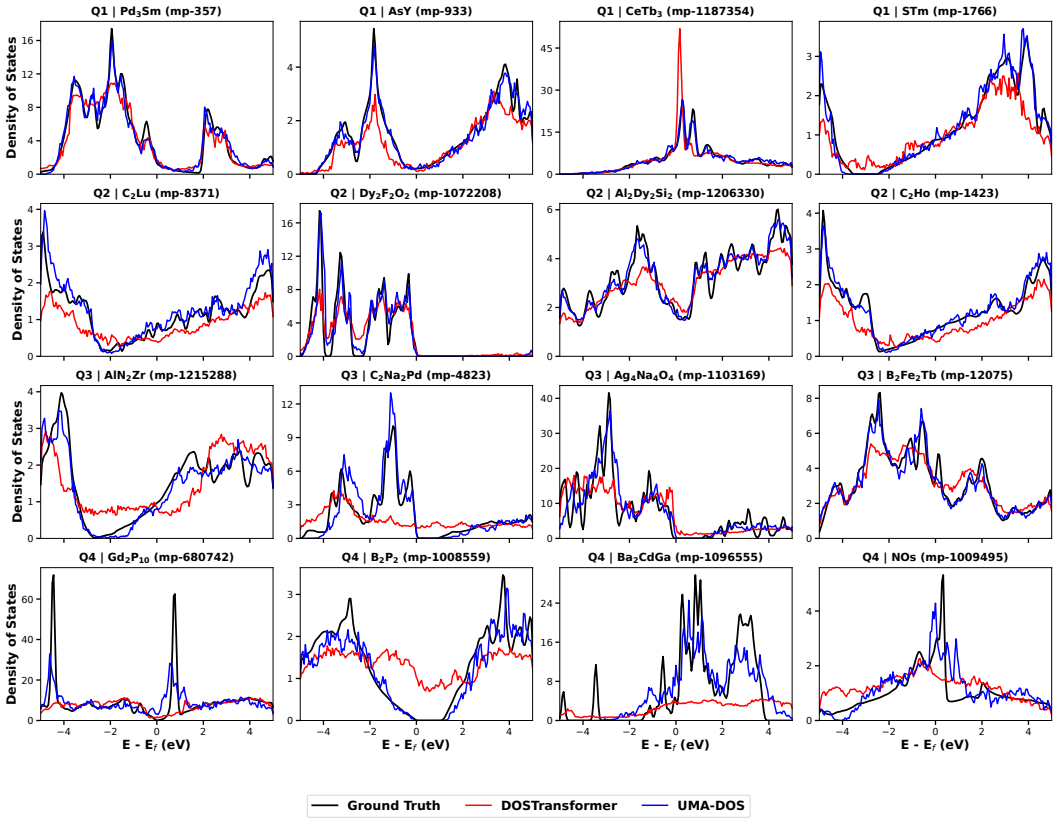


Figure 12: Extended qualitative comparison of UMA-DOS and DOSTransformer (absolute-scale), the same DOSTransformer variant used in Table 1 and Figure 4. We show four representative test structures from each per-structure difficulty quartile. In each quartile, the example closest to the quartile median is the one shown in Figure 4. Panels are ordered from Q1 (easiest) to Q4 (hardest) with quartile intervals Q1:[0.016, 0.080], Q2:[0.080, 0.114], Q3:[0.114, 0.157], and Q4:[0.157, 0.360]. Black curves denote the smoothed ground-truth DOS, while blue and red curves denote predictions from UMA-DOS and DOSTransformer, respectively. UMA-DOS remains more consistent across the quartiles, particularly in preserving local peak placement, valley depth, and rapidly varying spectral structure.

behavior, while UMA-DOS remains more faithful to the underlying spectral geometry. There are still individual hard cases where neither model is uniformly best on every local feature, but overall the gallery is qualitatively consistent with the stronger shape-sensitive performance of UMA-DOS in the quantitative comparison.

Figure 15 extends the qualitative comparison to Mat2Spec-WD using four examples from each difficulty quartile. Among the reproduced Mat2Spec variants, the Wasserstein-trained model shows the least stable qualitative behavior: in many examples it introduces spurious spikes, noisy local oscillations, or broadened distortions that are not present in the ground-truth DOS. By contrast, UMA-DOS remains more consistent in tracking the underlying spectral geometry, especially around narrow peaks, valley depth, and near-Fermi structure. The contrast is already visible in the easier quartiles and becomes more pronounced in the intermediate and hard quartiles, where Mat2Spec-WD often departs substantially from the target spectrum. Although both models remain challenged on the hardest spectra, the overall gallery is qualitatively consistent with the weaker quantitative performance of Mat2Spec-WD and the stronger spectral fidelity of UMA-DOS.

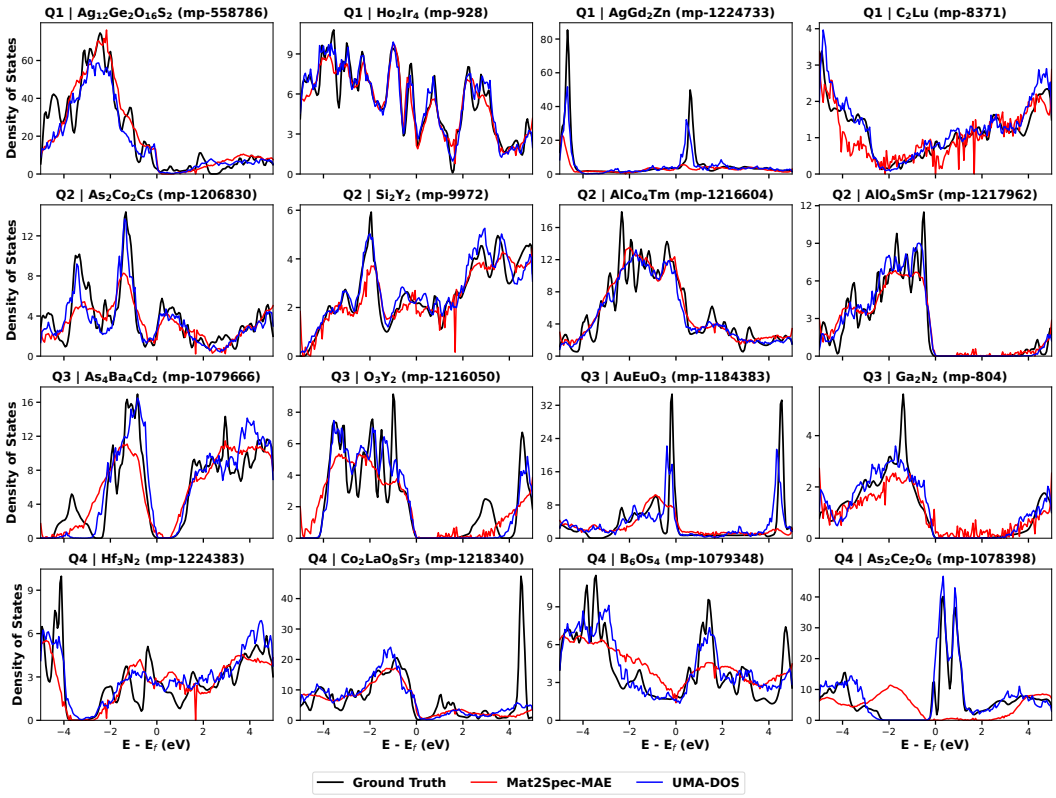


Figure 13: Extended qualitative comparison of UMA-DOS and Mat2Spec-MAE on Materials Project test structures. Four examples are shown from each quartile of per-structure difficulty, with rows ordered from Q1 (easiest) to Q4 (hardest)with quartile intervals Q1:[0.014, 0.093], Q2:[0.093, 0.129], Q3:[0.129, 0.175], and Q4:[0.175, 0.434]. Black curves denote the smoothed ground-truth DOS, while blue and red curves denote predictions from UMA-DOS and Mat2Spec-MAE, respectively. Across the quartiles, UMA-DOS more consistently preserves sharp peaks, local spectral structure, and near-Fermi behavior.

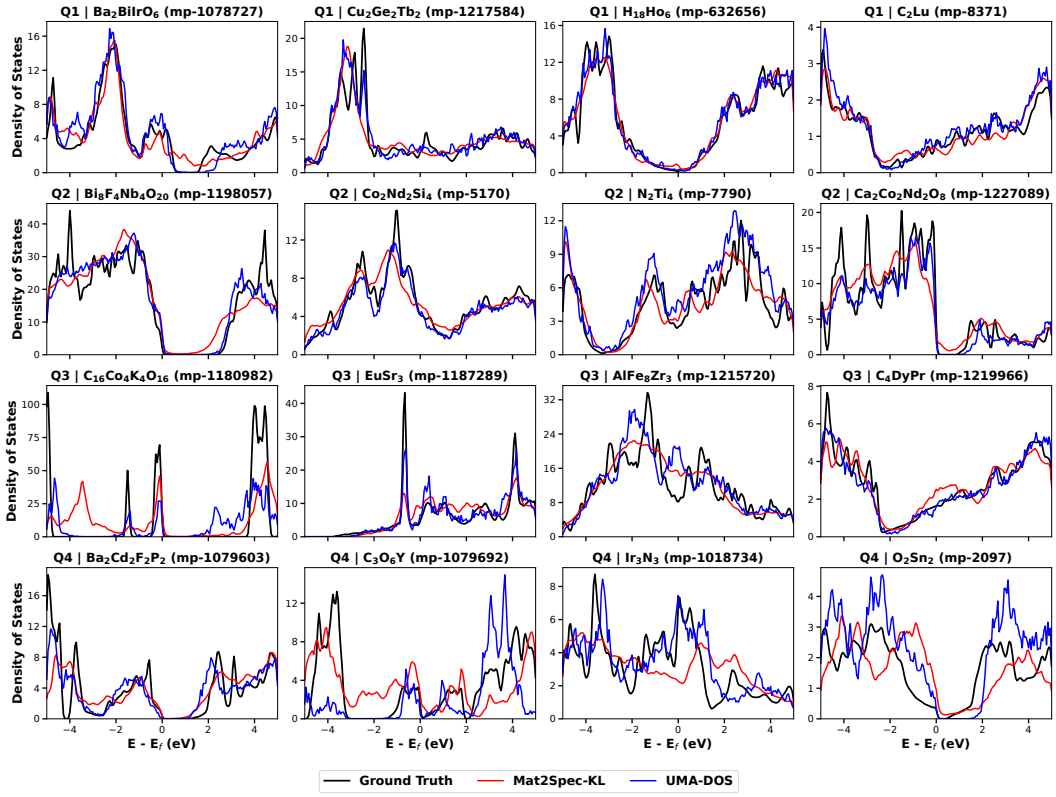


Figure 14: Extended qualitative comparison of UMA−DOS and Mat2Spec-KL on Materials Project test structures. Four examples are shown from each quartile of per-structure difficulty, with rows ordered from Q1 (easiest) to Q4 (hardest). Black curves denote the smoothed ground-truth DOS, while blue and red curves denote predictions from UMA−DOS and Mat2Spec-KL, respectively. Across the quartiles, UMA−DOS more consistently preserves local spectral structure, sharp peaks, and near-Fermi behavior.

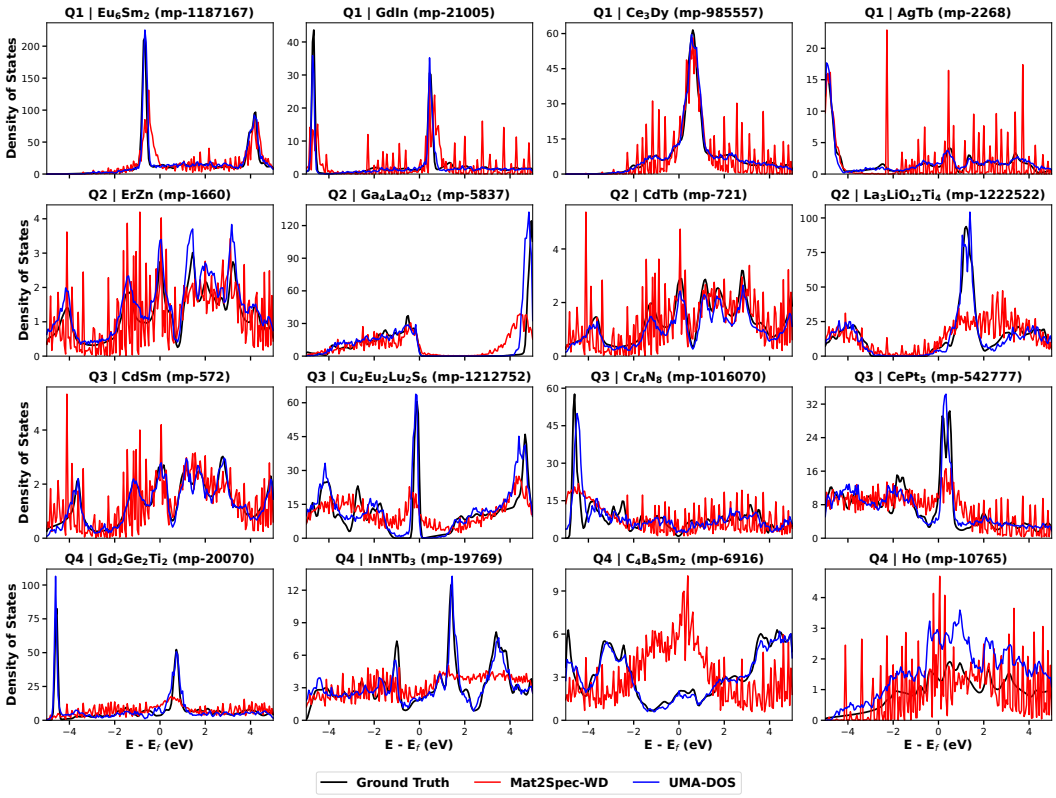


Figure 15: Extended qualitative comparison of UMA-DOS and Mat2Spec-WD on Materials Project test structures. Four examples are shown from each quartile of per-structure difficulty, with rows ordered from Q1 (easiest) to Q4 (hardest). Black curves denote the smoothed ground-truth DOS, while blue and red curves denote predictions from UMA-DOS and Mat2Spec-WD, respectively. Across the quartiles, UMA-DOS more consistently preserves local spectral structure, sharp peaks, and near-Fermi behavior, whereas Mat2Spec-WD often exhibits spurious spikes, noisy local fluctuations, or broader spectral distortions.

D SCOPE

D.1 LIMITATIONS AND SCOPE OF CLAIMS

Our empirical claims concern the prediction of total DOS under the unified Materials Project evaluation protocol used throughout this work. The main quantitative conclusions are drawn from controlled comparisons on the same split, energy reference, smoothing, and evaluation metrics whenever such a reproduction is possible. Results obtained under unmatched preprocessing or different target definitions are therefore treated as contextual rather than strictly head-to-head. Within this scope, the evidence supports improved total-DOS fidelity for UMA-DOS, with the clearest gains appearing in shape-sensitive metrics.

Several limitations remain. Our experiments focus on total DOS, and all main claims are based on a single Materials Project preprocessing pipeline. In addition, some appendix ablations are single-run exploratory studies rather than fully seeded comparisons. We mitigate these limitations by using a unified split-and-evaluate protocol for the main baselines, reporting three-seed results for the central comparisons, conducting matched low-data experiments, and using deterministic example-selection rules for qualitative figures. Therefore, our claims should be interpreted as controlled evidence for improved total-DOS prediction under this protocol, rather than as a universal statement about all spectral materials tasks.

D.2 BROADER IMPACT

UMA-DOS is intended to support early-stage computational materials screening by providing fast estimates of total density of states from crystal structures. Methods of this type may reduce the number of expensive first-principles electronic-structure calculations needed during initial candidate ranking, potentially accelerating the exploration of materials for electronic, optical, and energy-related applications. The direct contribution of this work is methodological: we study whether pretrained atomistic representations can transfer to structured electronic spectra, rather than proposing an autonomous materials-discovery system. The main risks are misuse through overinterpretation and dataset-induced bias.

D.3 COMPUTE RESOURCES AND RUNTIME

Table 13: Representative compute resources and runtime for the main UMA-DOS training regimes. All runs use a single 40 GB GPU. Wall-clock time includes intermediate validation/evaluation.

Training regime	GPU	Memory	Epochs	Wall-clock time
Scratch	1× A100-SXM4	40 GB	100	07:06:07
Frozen transfer	1× A100-SXM4	40 GB	100	03:45:42
ELoRA	1× A100-SXM4	40 GB	100	11:20:40
Full Fine-tuned	1× A100-SXM4	40 GB	100	05:20:27

All experiments were run on a single GPU, using an NVIDIA A100-SXM4-40GB. No distributed or multi-GPU training was used. Table 13 reports representative end-to-end wall-clock runtimes for the main UMA-DOS training regimes. Each run was trained for 100 epochs. The reported times include periodic validation and evaluation during training, and should therefore be interpreted as practical end-to-end runtimes rather than pure forward/backward optimization time. Although ELoRA updates fewer parameters than full fine-tuning, its measured training time is approximately 2.1 times longer in our implementation. Its reduction in trainable parameters, therefore, does not translate into shorter wall-clock training time in this setting.

Since the main seeded UMA-DOS results are reported over three random seeds, these timings correspond to approximately 21.3, 11.3, 33.6, and 16.0 single-GPU hours, respectively, for the scratch, frozen-transfer, ELoRA, and full fine-tuning configurations.

D.4 Code Availability

For reproducibility, we provide an anonymous repository containing the data-preparation pipeline, the `UMA-DOS` model implementation, and the training and evaluation scripts used in this work: `https://anonymous.4open.science/r/UMA_DOS-F520`.