npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01938-2

Combining feature-based approaches with graph neural networks and symbolic regression for synergistic performance and interpretability

Rogério Almeida Gouvêa[1,2] ✉, Pierre-Paul De Breuck[2,3], Tatiane Pretto[1], Gian-Marco Rignanese[2] ✉ & Marcos José Leite Santos[1]

This study introduces MatterVial, an innovative hybrid framework for feature-based machine learning in materials science. MatterVial expands the feature space by integrating latent representations from a diverse suite of pretrained graph-neural network (GNN) models—including structure-based (MEGNet), composition-based (ROOST), and equivariant (ORB) graph networks—with computationally efficient, GNN-approximated descriptors and novel features from symbolic regression. Our approach combines the chemical transparency of traditional feature-based models with the predictive power of deep learning architectures. When augmenting the feature-based model MODNet on Matbench tasks, this method yields significant error reductions and elevates its performance to be competitive with, and in several cases superior to, state-of-the-art end-to-end GNNs, with accuracy increases exceeding 40% for multiple tasks. An integrated interpretability module, employing surrogate models and symbolic regression, decodes the latent GNN-derived descriptors into explicit, physically meaningful formulas. This unified framework advances materials informatics by providing a high-performance, transparent tool that aligns with the principles of explainable AI, paving the way for more targeted and autonomous materials discovery.

Machine learning has revolutionized materials science, accelerating material discovery and property optimization across various domains[1–3]. The two prominent approaches in this field are feature-based and graph-neural-network (GNN) models, each with distinct advantages and limitations[4,5]. Feature-based models rely on predefined descriptors such as elemental properties, geometric features, and electronic structure information. They are highly interpretable and effective with small datasets, offering insights into structure-property relationships[6,7]. These models adapt well to custom tasks in experimental settings, such as nanocrystal research[8], catalysis[9], and organic photovoltaics[10]. In contrast, GNN models represent materials as graphs, capturing structural information through message passing and learning deep representations with simple atomic descriptors. This often results in more accurate predictions for complex materials, but requires greater computational resources and data for training[11,12]. GNNs are particularly effective in the large-scale screening of materials and for constructing interatomic potentials owing to their efficient computation and local information aggregation[13]; however, they lack interpretability.

Boosting the accuracy of feature-based models to make them competitive on larger datasets usually implies employing neural network models and relying on extensive suites, such as MatMiner[7], to produce meaningful features. This process is particularly time-consuming for sophisticated descriptors like the Orbital Field Matrix (OFM)[14] and the Smooth Overlap of Atomic Positions (SOAP)[15]. A novel strategy to boost these feature-based models involves leveraging the rich latent-space representations learned by GNN models pretrained on vast datasets. Even though neural networks are universal function approximators, easing their burden through well-aligned feature transformations can improve generalization, reduce training time, and stabilize convergence[16,17].

In this work, we address these challenges by proposing a hybrid approach that combines traditional chemically intuitive descriptors with

[1]Laboratory of Applied Materials and Interfaces, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil. [2]Institute of Condensed Matter and Nanosciences, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. [3]Present address: Ruhr-Universität Bochum, Bochum, Germany.
✉e-mail: rogeriog.em@gmail.com; gian-marco.rignanese@uclouvain.be

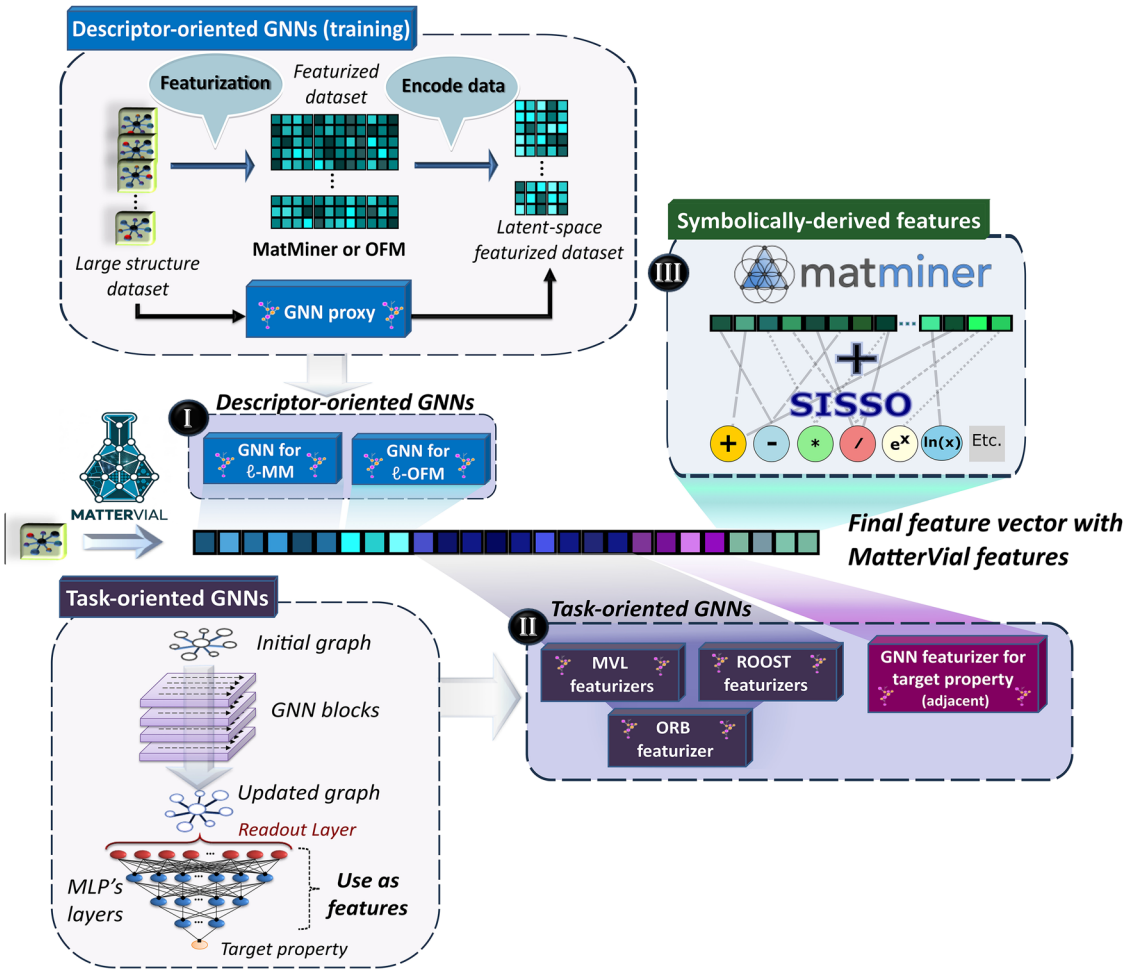

Fig. 1 | Overview of the methodology for leveraging latent-space features from GNN models with MatterVial. On the left (I), the generation and deployment of descriptor-oriented GNN models are illustrated. At the bottom (II), task-oriented GNN models—either pretrained or trained on the fly for adjacent variants—are shown, with feature extraction possible from activation, pooling, or multi-layer perceptron (MLP) layers on the model architecture. On the right (III), formulas from symbolic regression from SISSO are also implemented, leveraging traditional physicochemical descriptors available in MatMiner as a base.

latent features obtained from a diverse set of pretrained models. We incorporate features from both structure-based (MEGNet, coGN)[18,19] and composition-based (ROOST)[20] GNNs, as well as from ORB[21], a powerful equivariant Machine Learning Interatomic Potential (MLIP). To avoid the featurization bottleneck of traditional descriptors, we also leverage GNNs to generate fast, latent-space approximations of MatMiner (ℓ-MM) and Orbital Field Matrix (ℓ-OFM) features. Finally, we augment this feature set with new descriptors derived via symbolic regression. This multifaceted strategy aims to create a more robust, accurate, and versatile featurizer that capitalizes on the distinct strengths of each approach to be useful for a wider range of dataset sizes.

To simplify the generation of all those features, a package was developed named MatterVial, standing for MATerials feaTure ExtRaction Via Interpretable Artificial Learning (MatterVial), which, besides producing all latent-space features from the GNN models, aids in obtaining the interpretable chemical descriptors that correlate to these high-level features. This is achieved through techniques such as SHapley Additive exPlanations (SHAP) analysis in surrogate models and symbolic regression via Sure Independence Screening and Sparsifying Operator (SISSO) to obtain an approximate formula from the most important features. Our results demonstrate an overall improvement in all analyzed datasets compared with the baseline MatMiner featurizer. In addition, it surpassed the performance of the individual GNN models in several cases, indicating that the combination of traditional and latent-space features leads to a more robust generalization.

This work is situated within a recent methodological trend that repurposes GNNs not as end-to-end predictors, but as powerful and data-efficient feature generators for a variety of downstream tasks[22–26]. Our approach bridges feature-based and graph-based methods, leveraging their strengths to develop more versatile and task-agnostic machine learning models in materials science. By enhancing the accuracy, efficiency, and interpretability of property prediction, this framework facilitates the integration of both experimental and simulated data. Moreover, it aligns with the growing demand for explainable AI[27,28], which is essential for the advancement of self-driving laboratories in materials discovery and optimization[29].

Results

We evaluate our approach using the full MatBench v0.1 benchmark[30] with MODNet[4], which is the state-of-the-art feature-based model in materials science[12]. We adopt the same MatMiner featurization as that used in MODNet for MatBench in the original publication[4,31]. These can be complemented by three categories of MatterVial features, as illustrated in Fig. 1:

I. Latent-space features from descriptor-oriented GNNs: Conventional material descriptors are transformed into latent representations using an autoencoder trained on Materials Project[32] (MP) data. These descriptors include the widely used MatMiner features (ℓ-MM) and the features from the Orbital Field Matrix featurizer (ℓ-OFM). A GNN was then trained to replicate these latent features directly from the input

Table 1 | Performance comparison of three MODNet variants against the best multi-purpose MatBench model on each task in the MatBench v0.1 benchmark

MatBench task	n	MODNet@ MM (baseline)	MODNet@ MM+MV (% error reduction*)	MODNet@ MV (% error reduction*)	MatBench record (model)	Best MatterVial groups**
Steels yield strength (MPa)[44]	312	87.76	**85.12 (3.0%)**	120.95 (-37.8%)	MODNet	ROOST
$E_{exfol.}$ (meV/atom)[45]	636	33.19	29.19 (12.1%)	**28.86 (13.0%)**	MODNet	ORB, MVL, ℓ-OFM, ℓ-MM
argmax(PhDOS) (cm^{-1})[46]	1,265	34.27	**30.08 (12.2%)**	30.58 (10.8%)	**28.76** (MegNet)	MVL, ORB, ℓ-OFM, ROOST, SISSO
Exp. band gap (eV)[47]	4,604	0.333	**0.290 (12.9%)**	0.351 (-5.5%)	MODNet	ROOST, SISSO
Refractive index [32,48]	4,764	0.271	0.235 (13.3%)	**0.234 (13.7%)**	MODNet	ORB, ℓ-OFM, MVL, ℓ-MM
Exp. metallicity (eV)[47]	4,921	0.916	**0.976 (71.4%)**	0.898 (-59.3%)	0.921 (AMMExpress)	ROOST
Glass-forming ability [49,50]	5,680	0.936 (**0.960**)†	0.937 (1.6%)	0.904 (-50.0%)	MODNet	ROOST
Logarithmic G^{vrh} (log_{10}GPa)[51]	10,987	0.073	**0.032 (55.5%)**	0.033 (54.8%)	0.067 (coGN)	MVL, ORB, ℓ-MM, ROOST, SISSO
Logarithmic K^{vrh} (log_{10}GPa)[51]	10,987	0.056	**0.027 (49.6%)**	0.028 (50.1%)	0.049 (coNGN)	MVL, ORB, ℓ-MM, ℓ-OFM, ROOST, SISSO
Perovskite ΔH_{form} (eV/unitcell)[52]	18,928	0.0908	0.0386 (57.5%)	0.0389 (57.3%)	**0.0269** (coGN)	ORB, MVL, ℓ-OFM, ℓ-MM, ROOST, SISSO
Band gap (eV)[32]	106,113	0.2199	0.137 (37.6%)	**0.137 (37.8%)**	0.156 (coGN)	MVL, ORB, ROOST, SISSO
Metallicity[32]	106,113	0.904	**0.978 (77.1%)**	0.976 (75.0%)	0.9520 (CGCNN)	ORB, MVL, ℓ-OFM, ROOST
E_f (eV/atom)[32]	132,752	0.0448	0.0147 (67.2%)	**0.0138** (69.2%)	0.0170 (coGN)	MVL, ORB, ℓ-OFM, ℓ-MM

Metrics are reported as mean absolute error (MAE) for regression and area under the receiver-operator curve (AUROC) for classification. MODNet@MM uses only MatMiner features; MODNet@MM + MV augments these features with MatterVial descriptors; and MODNet@MV uses only MatterVial features, essentially substituting MatMiner features by ℓ-MM. For each task, the MatterVial feature group that yields the best result is shown. Scores in bold identify the overall best model per task, and shaded tasks are those in which MODNet was already the best model.

[a]% error reduction $= \frac{MAE_{baseline}-MAE_{model}}{MAE_{baseline}} \times 100\%$ (regression) or $\frac{(1-AUROC_{baseline})-(1-AUROC_{model})}{(1-AUROC_{baseline})} \times 100\%$ (classification).

[b]Subset of descriptors that consistently figured in the most important features in the MODNet@MV models for the task, ordered by importance. MVL, ORB, and ROOST refer to the task-oriented GNN features, respectively, those from MVL MEGNet models for structures, the MLIP Orb-v3, and pretrained ROOST models for compositions. ℓ-MM and ℓ-OFM refer to the descriptor-oriented GNN features; ℓ-MM, when included, substitutes the MatMiner features for faster generation. SISSO refers to the group of features derived from MM features via symbolic regression.

[c]As we were unable to replicate the reported 0.960 AUROC for glass formability using MODNet, we present our best MODNet@MM result as a baseline instead. Despite the lower score, MODNet continues to outperform other models in MatBench for this task.

structures. This method achieves a computational efficiency similar to that of GNNs and still preserves interpretability via decoding.

II. Latent-space features from task-oriented GNNs: These features are extracted directly from the intermediate layers of pretrained GNN models that have been developed for various tasks. Specifically, we incorporate MEGNet models from the Materials Virtual Lab (MVL) that were pretrained for the prediction of elastic constants, band gap, and formation energy, as well as for the metal-insulator classification. We also consider composition-based ROOST models for the band gap and formation energy. In addition, we include the internal layers of ORB-v3, a state-of-the-art equivariant MLIP trained to reproduce energies and forces. This group capitalizes on the strengths of GNN architectures in capturing complex structural representations, aiming to enhance predictive performance on larger datasets.

III. Symbolically derived feature combinations: Here, we use the MatMiner features as a basis to generate new compound features. Through symbolic regression with SISSO, we identify several combinations of pairs of features (rung one) that exhibit enhanced correlations with the target properties of interest in materials science. These derived formulas are then incorporated as new features.

Since the features obtained from task-oriented GNNs are high-level and not directly interpretable as traditional descriptors, we develop a method to decompose them into interpretable descriptors, which is integrated into the Interpreter module in MatterVial. In addition, features from descriptor-oriented GNNs can be decoded into their interpretable counterparts. Equally, the third group of augmented features via symbolic regression can have their formulas retrieved by name. Comprehensive implementation details for each category and for the Interpreter module are available in the "Methods" section and in the Supplementary Information (SI).

MatBench validation of MatterVial features

Table 1 presents the performance of MODNet using MatMiner augmented with MatterVial descriptors (MODNet@MM + MV) and MODNet using

Table 2 | Mean absolute errors (MAEs) for the MatBench task of the heat of formation of perovskites with different models

Reference models	MAE (eV/unit cell)	MatterVial models	MAE (eV/unit cell)[b]
Descriptor-oriented			
AutoMatMiner (MatBench[a])	0.2005 (±0.0085)	MODNet@ℓ-MM	0.1052 (±0.0022)
MODNet@MM (this work)	0.0888 (±0.0025)	MODNet@MM + ℓ-OFM	0.0794 (±0.0016)
MODNet@MM + OFM	0.0751 (±0.0018)	MODNet@ℓ-MM + ℓ-OFM	0.0973 (±0.0016)
Task-oriented (MVL, ROOST) and SISSO			
MODNet@MVL	0.0716 (±0.0020)	MODNet @ℓ-MM + ℓ-OFM + MVL	0.0673 (±0.0015)
MODNet@MVL + ROOST	0.0707 (±0.0017)	MODNet@ℓ-MM + ℓ-OFM+ +MVL + SISSO	0.0653 (±0.0013)
MODNet@MM + ℓ-OFM + + MVL + SISSO + ROOST	0.0637 (±0.001)	MODNet@ℓ-MM + ℓ-OFM + + MVL + SISSO + ROOST	0.0639 (±0.0010)
Task-oriented (ORB featurizer)			
MODNet@MM + ℓ-OFM++MVL + SISSO + ROOST + ORB	0.0386 (±0.0009)		
HackNIP[26] (MODNet@ORB)	0.0397	MODNet@MV[c]	0.0388 (±0.0006)
MV + Adjacent GNN model			
MEGNet (MatBench[a])	0.0352 (±0.0016)		
MEGNet (this work)	0.0685 (±0.0036)	MODNet@ MV+Adj(MEGNet)	0.0343 (±0.0014)
coGN (MatBench[a])	0.0269 (±0.0008)	MODNet@ MV+Adj(coGN)	0.0313 (±0.0012)
coGN (this work)	0.0271 (±0.0008)	MODNet@ MV+Adj(coGN)+hiSISSO	0.0288 (±0.0009)

The values in parentheses represent the standard deviation across the five cross-validation folds.
[a]Data retrieved from *MatBench*[12] in August 2025.
[b]Training logs and fold-wise results for the best models are provided in the Supporting Data repository (https://github.com/rogeriog/MatterVial_SupportData).
[c]For brevity MV = (ℓ-MM + ℓ-OFM + MVL + SISSO + ROOST + ORB), i.e., all pretrained featurizers in MatterVial.

only MatterVial descriptors (MODNet@MV) relative to the baseline model using only MatMiner features (MODNet@MM) in the 13 MatBench tasks. The results show that blending both latent-space representations from task-oriented and descriptor-oriented GNNs with symbolically derived features consistently reduces prediction errors across this diverse array of property prediction tasks.

Our approach significantly improves the performance on smaller datasets, where feature-based models have traditionally outperformed GNNs. Specifically, our models set new performance records for four tasks previously led by MODNet@MM and now achieve a leading performance in metallicity classification from experimental data. Notably, the glass-forming ability task alone did not result in substantial improvements. We highlight that for smaller composition-based datasets, MatMiner featurization is sufficiently fast to make MODNet@MM + MV computationally effective. For larger datasets, in which traditional featurization is very time-consuming, our MODNet@MV models significantly bridge the gap between feature-based and graph-based models, even outperforming state-of-the-art (SOTA) models in predicting properties such as elastic constants, band gap, metallicity, and formation energy. This success demonstrates that our approach effectively addresses the common shortcomings of both feature- and graph-based models.

The validity of these results is underscored by a rigorous analysis of their generalizability. Although the GNNs leveraged by MatterVial are pretrained on large repositories (e.g., MP, OQMD[33], and Alexandria[34]), our analysis (Section S13, Table S15) shows no direct link between dataset overlap and performance gains. This indicates that performance gains arise from the quality and transferability of learned structural representations. Additional benchmarks on 33 external datasets, including JARVIS-DFT[35] and four specialized sets (Section S14, Tables S16–S18), further confirm that the learned representations generalize effectively across materials and properties beyond those seen by incorporated pretrained models, and even varied levels of structural fidelity, such as idealized prototype structures versus optimized geometries.

An analysis of the feature contributions in Table 1 reveals that task-oriented latent features are the primary drivers of performance gains. The inclusion of ROOST aims at enhancing performance in composition-based tasks, and yet the model has reliably improved results in a wide range of tasks that also contain structural information. This performance may be attributed to the attention mechanism that captures unique patterns during activation and material pooling. For structure-based tasks, MVL-derived features have shown a significant positive impact. They boost predictions even when the prediction targets differ from those used in the original models, such as in predicting the perovskite heat of formation and refractive index. The ORB features, derived from an equivariant MLIP, proved particularly impactful, frequently appearing as top contributors. This is chemically intuitive, as the model's training on energies and forces provides a rich, physically meaningful latent space that is useful for transfer learning. This aligns with very recent findings by Kim et al.[26], who also employed ORB features with MODNet for structure-based regression tasks. Our approach achieves enhanced performance by incorporating all Orb-v3 layers and combining these features with diverse descriptor groups within our framework.

The descriptor-oriented and symbolically derived features also provided consistent complementary improvements. The ℓ-OFM features improved performance across most tasks, validating that our GNN-based approximation is an efficient and effective method for incorporating the descriptive power of computationally expensive descriptors, like the Orbital Field Matrix. The ℓ-MM features, designed as a shortcut for MatMiner features via GNN, lead to improved or similar performance on many tasks. Compared to the models that used the full MatMiner features (MODNet@MM + MV), we argue that the reconstruction loss was sufficiently low and that, in some cases, the encoder effectively refined the representation via regularization, improving the metrics. Crucially, these latent-space representations remain decodable, preserving much of the interpretability, which is a hallmark of feature-based models. Finally, the SISSO-derived features, while less universally impactful, still boosted performance in roughly half of the benchmarks. Given that we utilized only first-rung symbolic regression, we conjecture that there is clear potential for further gains with higher-level, more complex formulas. Ultimately, these results show that our approach simultaneously accelerates featurization, improves model performance, and provides valuable chemical insights. This combination of benefits repositions feature-based models as strong and practical alternatives to end-to-end GNNs for property prediction.

Synergy of MatterVial features and the adjacent GNN model

Having demonstrated the performance gains of our method, we now turn to the individual contributions of the MatterVial features. We examine the

synergetic effects of each MatterVial feature group using the perovskite heat of formation task as an example. Table 2 illustrates a step-by-step performance evaluation for this task, revealing how the integration of different MatterVial feature groups leads to cumulative improvements. Starting from our baseline, the MODNet@MM model delivers an MAE of 0.0888 eV/unit cell. This performance serves as a reference point against which the benefits of the additional features can be measured.

The first modification involves introducing descriptor-oriented GNN features, ℓ-OFM, and ℓ-MM, which are designed to be computationally faster approximations of their full counterparts. When MatMiner features are entirely replaced by their latent representation (MODNet@ℓ-MM), the MAE is 0.1052 eV/unit cell. While higher than our MODNet@MM baseline (0.0888 eV/unit cell), this still significantly outperforms AutoMatMiner (0.2005 eV/unit cell), demonstrating ℓ-MM as a viable, faster featurization alternative. Augmenting MatMiner with ℓ-OFM (MODNet@MM + ℓ-OFM) reduces the MAE to 0.0794 eV/unit cell. This is lower than the baseline, although still higher than that obtained using the original computationally intensive OFM features (MODNet@MM + OFM, 0.0751 eV/unit cell). Combining both ℓ-MM and ℓ-OFM (MODNet@ℓ-MM + ℓ-OFM) yields an MAE of 0.0973 eV/unit cell. These results highlight that our proxy GNN featurizers offer a compelling trade-off, capturing essential chemical information with a substantial speedup in featurization.

Building on this foundation, the incorporation of task-oriented GNN features from the MVL pretrained models further boosts performance in the MODNet@ℓ-MM + ℓ-OFM + MVL model, lowering the MAE to 0.0673 eV/unit cell. Clearly, the MVL descriptors capture additional structural and physicochemical details that the MM and OFM features do not, thereby enhancing the ability of the model to predict heat formation (more details on the MVL descriptions and the effect of different layers are provided in SI, Section S6).

Next, the addition of symbolically derived feature combinations via SISSO produces modest refinement, reducing the MAE to 0.0653 eV/unit cell. Although the improvement is small, it underscores the notion that simple algebraic combinations of conventional descriptors can reveal nonlinear relationships, complement the latent-space features, and thereby enhance prediction accuracy.

Further refinement is achieved by incorporating composition-based ROOST features. At first glance, one might not expect an improvement over the MEGNet MVL models since they incorporate structural information alongside composition. However, we believe that the attention-based mechanism present in ROOST is responsible for capturing additional meaningful information to complement other feature groups and achieve an MAE of 0.0639 eV/unit cell. Furthermore, at this point, using the standard MatMiner features instead of their latent representation (ℓ-MM) yields a nearly equivalent performance (MAE of 0.0637 eV/unit cell). These results confirm that the rapidly generated encoded representations can effectively replace the full MatMiner features in tandem with other descriptors. However, eliminating MatMiner features entirely (neither MM nor ℓ-MM) causes a significant decrease in accuracy with 0.0707 eV/unit cell in MODNet@MVL + ROOST and 0.0716 eV/unit cell in MODNet@MVL, indicating that MatMiner features are valuable and not simply redundant to these GNN descriptors. In fact, a synergistic effect among all MatterVial feature groups is observed in this dataset.

ORB features stand apart from other featurizers like MVL and ROOST. While MVL and ROOST were trained on smaller datasets, specifically MP and OQMD[33] (about 1.5 million structures combined), the ORB-v3 featurizer was trained on a significantly larger dataset. This dataset, which combines MP, Alexandria[34], and OMat[36], leverages approximately 120 million calculated structures, which is at least two orders of magnitude larger than either of those datasets. The extraction of features from this model in MatterVial to use in MODNet significantly reduces the mean absolute error in the task, but a slight improvement is still seen with the other MatterVial features that were included. We conjecture that larger reductions might still be achievable by training more task- and descriptor-oriented models in these larger datasets.

Despite the significant reduction, feature-based approaches using pretrained models with MatterVial or HackNIP[26] still fall short of the results obtained purely with GNNs, such as MEGNet and coGN trained in the perovskites dataset. Based on this observation, we incorporate into MatterVial the possibility of training adjacent GNN models on the fly and extracting their features with the AdjacentGNNFeaturizer class. We achieve 0.0343 eV/unit cell using the MEGNet adjacent model features. The MEGNet benchmarked MAE is substantially lower than what we achieved using the default configuration of the model, even with the same elemental embeddings provided by the authors. This discrepancy is possibly due to differences in the hyperparameters, inclusion of additional features, and larger training times employed for the benchmark[37]. Finally, we employed the SOTA coGN model as an adjacent model for feature extraction, and we obtained comparable results to the reported values in MatBench with this model. Incorporating coGN features in our MODNet model reduced the MAE to 0.0313 eV/unit cell, which is much closer to the 0.0269 eV/unit cell record.

Figure 2 graphically depicts the synergistic effects detailed in Table 2 by comparing different models. The feature importance from the mean absolute SHAP values aggregated by feature group in Fig. 2a–c quantifies the contribution of the different groups of features and shows a clear shift in dominance as more powerful features are introduced. However, a closer inspection reveals important nuances regarding how these feature sets interact. In the MODNet@MV(no ORB) model, there is a relatively balanced and significant contribution from all feature groups, led by MVL, ℓ-MM, and SISSO, underscoring their collective utility. This is seen more clearly with the t-SNE projections of the SHAP value vectors for each feature in the model, where we can see these three sets of features covering most regions of the projection, but still some contributions of ROOST and ℓ-OFM features. When ORB features are introduced (Fig. 2b, e), they become the dominant contributor, explaining the dramatic reduction in MAE observed in Table 2. Crucially, the ℓ-MM and SISSO features retain a significant portion of their importance, with SISSO being among the highest contributors. This indicates that they capture complementary chemical information that is not fully encapsulated within the ORB latent space, explaining the slightly better result obtained compared to HackNIP's MODNet@ORB model[26]. This hierarchical and synergistic contribution of features directly explains the visual improvement in the data manifold shown in the t-SNE projections (Fig. 2g–i). The feature space of the MODNet@MV(no ORB) model (Fig. 2g) shows some organization. However, the introduction of ORB features (Fig. 2h) creates a significantly more structured manifold with a smoother gradient along the target property.

This synergistic contribution continues in the final MODNet@MV +Adj(coGN) model. The inclusion of adjacent coGN features (Fig. 2i) results in the most well-defined feature space in the t-SNE projection, with the clearest separation between data points according to the target feature. While the task-specific coGN features predictably take the lead, the pretrained ORB and MVL features remain highly influential, serving as the second- and third-most important groups, respectively (Fig. 2c, f). In contrast, the contributions from ℓ-MM and SISSO are now marginal, as their predictive information has been superseded by more powerful GNN features. This layered view of contributions highlights the interpretability brought by feature-based models. In the following section, we showcase how this interpretability can be deepened using new MatterVial tools.

Interpretability of MatterVial features

We begin by analyzing the most important features of the MODNet@MV(no ORB) model to understand what factors increase its accuracy in predicting the perovskite heat of formation (ΔH_f). Unlike end-to-end GNNs, where features are deeply entangled through message passing, feature-based models have readily decoupled features, and SHAP values can be used to robustly assess the most important features, as shown in the plot in Fig. 3. Utilizing the MatterVial Interpreter module, we can easily obtain SISSO formulas with up to five terms to approximate the GNN features of the included pretrained models. These approximations are based on

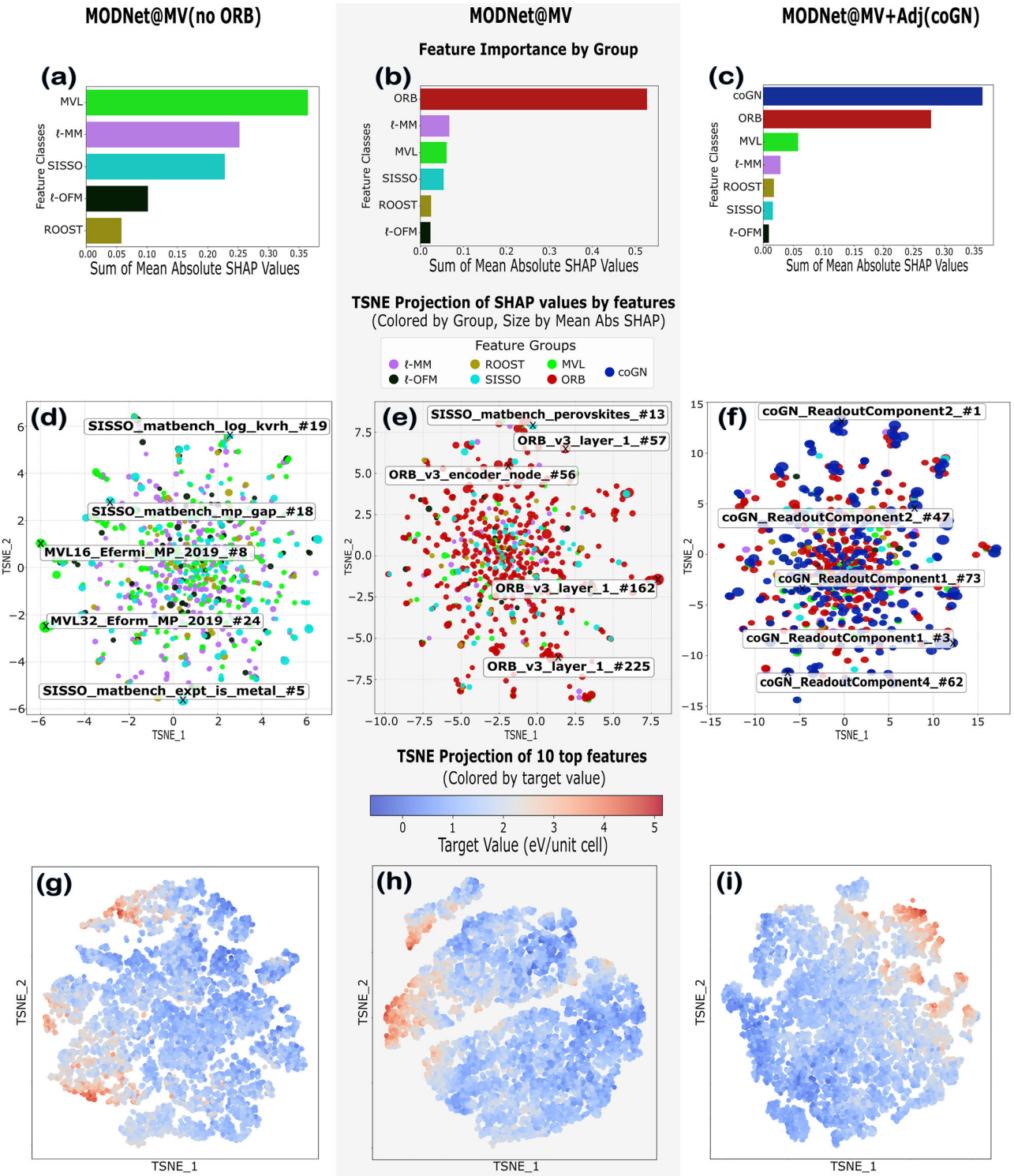

Fig. 2 | The impact of different MatterVial feature sets on the performance of MatBench's perovskite heat of formation task is illustrated using three different types of plots for each of the three models. On the left, MODNet@MV(no ORB), where features from the ORB model are excluded; at the center, with a shaded background, MODNet@MV, which includes all pretrained MatterVial features; and on the right, MODNet@MV+Adj(coGN), which further incorporates features from an adjacent coGN model trained on the task. **a**–**c** Bar plots showing the feature importance aggregated by a group of features through the sum of the mean absolute SHAP values. **d**–**f** t-SNE projections of the SHAP values for each feature in the model, colored by the feature group, and some of the features with the highest contribution are annotated. **g**–**i** t-SNE projections of the top 10 most important features colored by the target value (heat of formation).

interpretable descriptors from MatMiner and OFM. The plot displays the one-term formulas and their corresponding R^2 values, demonstrating that even with relatively simple descriptors, these approximation formulas can achieve high R^2 values for many meaningful GNN features.

This analysis identifies several key feature groups that drive the predictions. Features from the MVL formation energy model, for instance, correlate stability with large electronegativity gaps (promoting ionic character) and low d-electron fractions, which favor early transition metals.

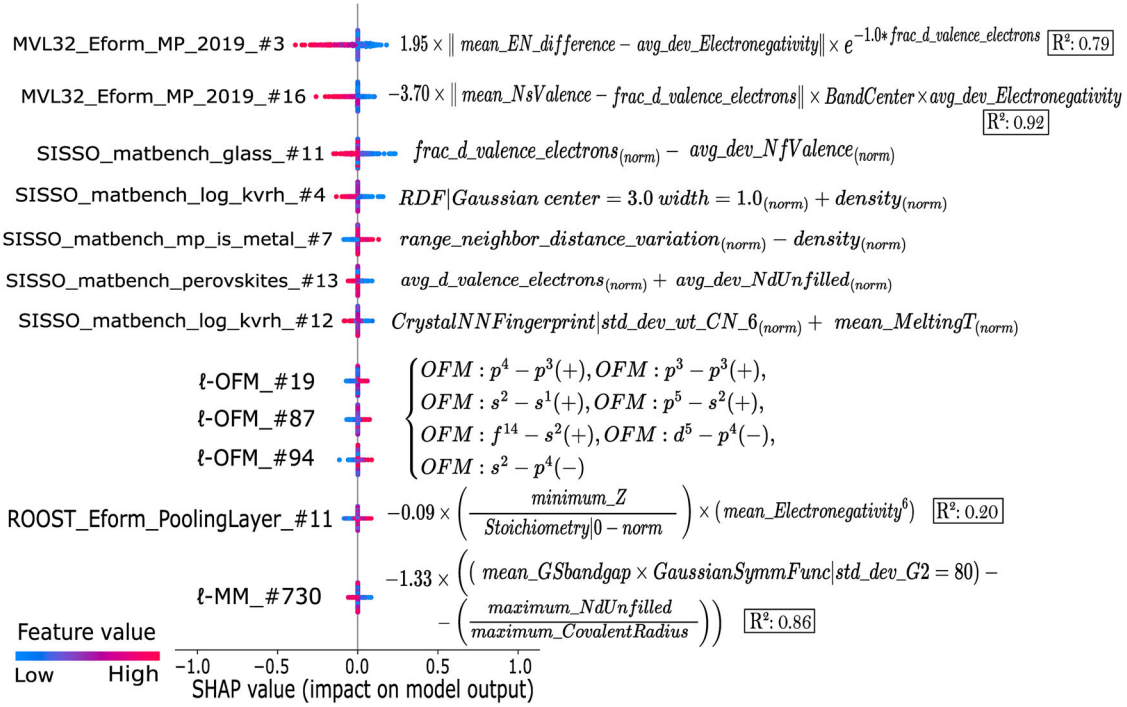

Fig. 3 | SHAP values plot for selected MatterVial features in the MODNet@MV(no ORB) model for the perovskite heat of formation. The plot displays the impact of individual features on the model's output (SHAP value), with the color indicating the feature's value (blue for low, red for high). Alongside each feature, the corresponding 1-term SISSO formula approximation for the MatterVial features and its R^2 value, when appropriate, are shown.

Features generated by SISSO highlight structural drivers, rewarding dense atomic packing, ordered coordination environments, and specific stabilizing factors like 3 Å interatomic contacts, while penalizing destabilizing electronic effects from excess d-electrons. Compositional features from ROOST and encoded MatMiner (ℓ-MM) models capture broader trends, showing that perovskites made of heavier, chemically diverse elements tend to be less stable and illustrating the balance between destabilizing wide-band-gap elements and the stabilizing effect of species with many unfilled d-states. Finally, encoded OFM (ℓ-OFM) features provide a granular view of bonding, distinguishing between the stabilizing interactions characteristic of oxides (e.g., s^2−p^4) and weaker bonds involving pnictogens. Collectively, this demonstrates that the model learns a multifaceted and physically grounded understanding of perovskite stability. A full breakdown of the individual features shown in the figure is provided in the Supplementary Information, Section S10.

A comparative SHAP analysis of the best-performing MODNet@MV and MODNet@MV+adj(coGN) models, which incorporate richer ORB and coGN features (see SI, Section S11, Figs. S4–S7), showed that while the MODNet@MV(no ORB) model primarily relies on fundamental chemical descriptors, the addition of ORB features shifts the emphasis of the model toward geometric information such as packing efficiency. The top-performing MODNet@MV+adj(coGN) model builds on this by capturing the most sophisticated features, representing a complex interplay between chemical and geometric properties. This increase in predictive power is accompanied by a decrease in direct interpretability. As the models become more complex, the ability to approximate their most important features with simple SISSO formulas diminishes (indicated by progressively lower R^2 values), and their correlation with classical descriptors weakens (see discussion in SI, Section S11 and Table S13). This progression highlights the gap between the complex features of high-performing GNNs and the limited descriptive power of interpretable descriptors, emphasizing the need for more flexible descriptors that remain compact for symbolic regression methods and interpretability.

To test the utility of our GNN feature approximations, we conducted a two-stage experiment. In the first stage, we compared two types of SISSO models: a baseline using only MatMiner and OFM descriptors, and an enhanced version that added formulas approximating the GNN's most important features. For both model types, we apply a consistent methodology, utilizing several primary feature preselection algorithms—including mRMR (i-SISSO)[38], random forest importances (rf-SISSO)[39], and our xgb-rfe-SISSO (SI, Section S8). The addition of the GNN-derived features yields a significant and consistent reduction in prediction error, as shown in Fig. 4a. Our approach is analogous to hierarchical SISSO (hiSISSO)[40], but it uniquely feeds back approximations of learned GNN features rather than terms from a prior SISSO model. In the second stage, we extract the terms from this enhanced SISSO model and incorporate them as new "hiSISSO features" to augment the MODNet@MV+adj(coGN) model. This augmented model further reduces the error to 0.0288 eV/unit cell. The t-SNE projection of SHAP value contributions and average feature importance of the classes in Fig. 4b, c confirms their effectiveness, showing the high per-feature predictive power of hiSISSO features complementing the model. This demonstrates that explicit, interpretable formulas can improve generalization and raises the compelling question of whether GNN features could be replaced entirely if more expressive, physically grounded descriptors were available.

Discussion

In this work, we introduced MatterVial, a unified and modular hybrid framework designed to bridge the gap between the predictive power of graph neural networks (GNNs) and the chemical transparency of traditional feature-based models in materials science. By augmenting the state-of-the-art feature-based model MODNet with a diverse and synergistic set of descriptors, this approach elevates its performance to be competitive with, and in several cases superior to, end-to-end GNNs. To summarize our contributions:

(i) MatterVial is a novel open-source Python framework that generates a rich hybrid feature set. It integrates latent-space representations from various pretrained models, including structure-based GNNs such as MEGNet, an equivariant interatomic potential (ORB), and composition-based networks such as ROOST. The framework also

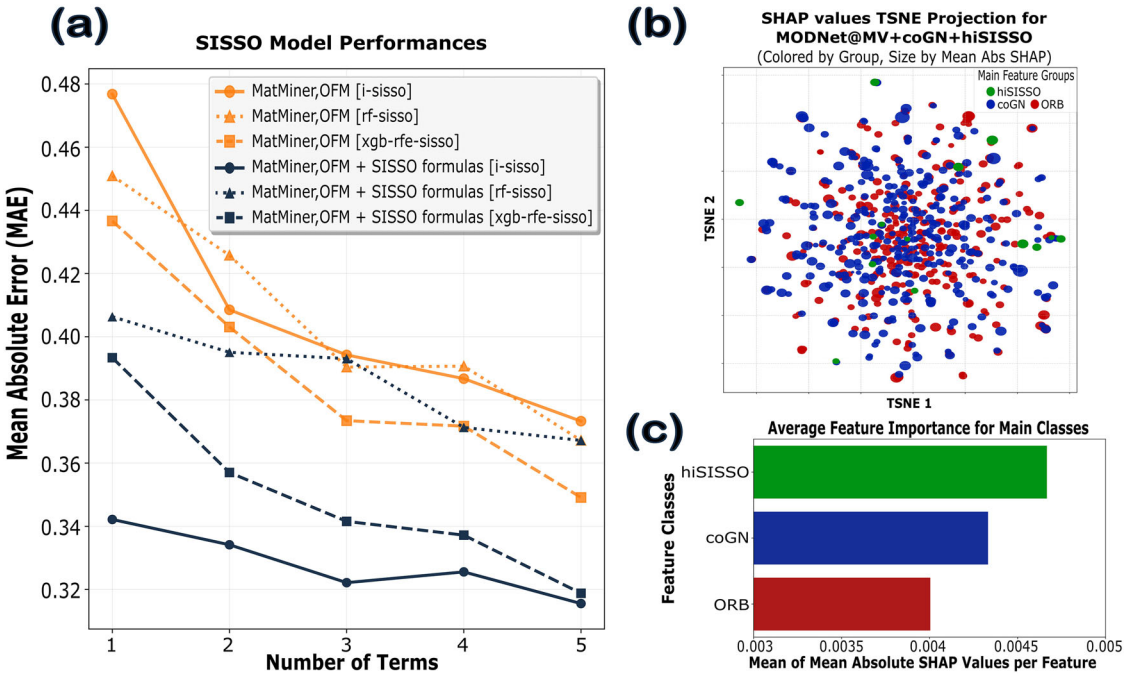

Fig. 4 | SISSO models and hiSISSO-enhanced MODNet model analysis on matbench_perovskites task. a Mean absolute error of models with baseline MatMiner+OFM features (orange) vs. those augmented with SISSO formulas approximating the best GNN features (dark green). Feature selection methods include i-SISSO, rf-SISSO, and xgb-rfe-SISSO; **b** t-SNE projection of SHAP values for top feature groups in the final MODNet@MV+adj(coGN)+hiSISSO model; point size reflects feature impact. **c** Average feature importance across main classes in the final model, calculated from the mean absolute SHAP values.

uses computationally efficient GNN-approximated descriptors (ℓ-MM, ℓ-OFM) and features derived from symbolic regression.

(ii) The hybrid model demonstrates broad applicability and superior performance across the full MatBench v0.1 benchmark and JARVIS-DFT. It consistently reduces prediction errors across nearly all 13 tasks in MatBench and establishes new state-of-the-art records for feature-based models in several categories.

(iii) A key innovation is a method that systematically decodes abstract GNN-derived features into more intuitive formulaic descriptors. This is achieved using surrogate models and symbolic regression to translate latent representations into explicit mathematical expressions based on fundamental physicochemical properties.

(iv) By incorporating features from an adjacent, task-specific GNN model, the framework enables a feature-based model to achieve predictive accuracy that is highly competitive with state-of-the-art GNNs while uniquely maintaining a modular and analyzable feature space.

(v) It was demonstrated that the interpretable formulas extracted from GNNs can be fed back into the model as new "hiSISSO features", leading to a further reduction in the prediction error. This confirms that the interpretability method can capture causally relevant physical information.

In conclusion, this work highlights the continued relevance and power of feature-based modeling in materials informatics. It delivers a practical solution that meets the dual demands of high accuracy and interpretability, a combination that is becoming increasingly critical in the field. While predictive accuracy is essential, interpretability allows researchers to validate that models have learned physically meaningful principles, thereby building trust and moving beyond simple prediction to genuine scientific understanding. This deeper insight accelerates materials discovery by enabling a shift from brute-force screening to more targeted, hypothesis-driven design. Ultimately, this alignment with the principles of explainable AI is a prerequisite for developing the next generation of autonomous discovery platforms, or "self-driving labs", which require models that can not only predict outcomes but also explain the underlying principles to guide subsequent experiments.

Methods

MODNet model training

The MatMiner featurizer used throughout this study is described in detail in the Supplementary Information, Section S1. For all experiments incorporating MatterVial features, since many features are obtained, we perform an initial preselection of features using recursive feature elimination with XGBoost[41] to reduce the pool to 800 features. Subsequently, the built-in MODNet feature selection algorithm is used to select and rank a subset of these features that will be used for training. At this point, we can determine which groups of MatterVial features are relevant for a given task ("Best MatterVial groups" in Table 1). The MODNet models are optimized via a genetic algorithm to select the best hyperparameters, and the optimal models in the validation set form deep ensembles, as described in ref. 31, which are then used for evaluation in the test set and to obtain the final metrics.

The mean absolute error (MAE) serves as the primary evaluation metric in regression tasks, and for classification tasks, the area under the receiver-operator curve (AUROC) is used. We consistently use a fivefold cross-validation method, as described in Matbench[25], in all presented tasks. A Supplementary data repository with detailed results of our work is available at https://github.com/rogeriog/MatterVial_SupportData.

MatterVial implementation

MatterVial is an open-source featurizer tool implemented in Python (available at https://github.com/rogeriog/MatterVial) to enhance material property predictions by integrating pretrained descriptor-oriented and task-oriented GNNs, as well as precomputed symbolic formulas from traditional chemically intuitive descriptors. The package offers significant flexibility and modularity, allowing the extraction of features from different layers of pretrained models and the incorporation of other GNN models as needed. The following outlines each MatterVial featurizer employed:

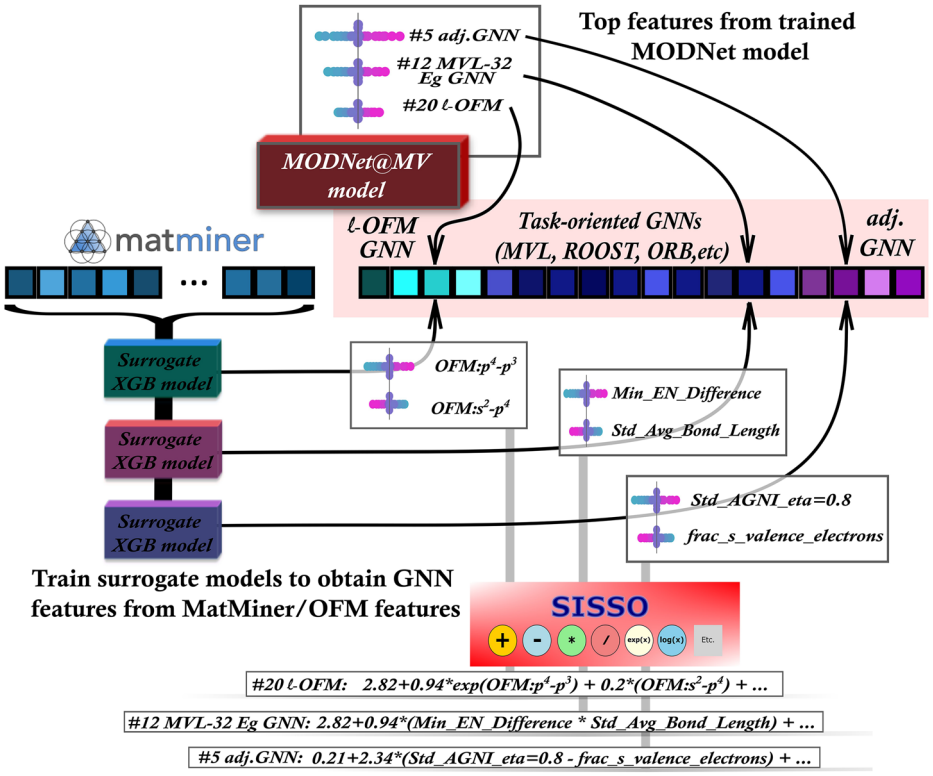

Fig. 5 | Interpretability workflow for MatterVial latent features. Schematic representation of the surrogate model pipeline. XGBoost models are trained on the MP2018-stable dataset to map high-level latent representations to interpretable MatMiner and OFM descriptors. Feature importance is quantified via SHAP analysis (top 30 descriptors), which subsequently informs SISSO++ symbolic regression to derive analytical formulas for each latent feature.

- ℓ-OFM featurizer: the OFM featurizer captures valence electron interactions at each atomic site by employing a weighted vector outer product of one-hot encoded valence orbitals for every atom (details in SI, Section S2, Fig. S1). The structural representation is achieved by averaging all local OFMs. We apply the OFM featurizer to a subset of the MP-crystals-2018.6.1[42] dataset with 106,113 structures whose energy above the convex hull was lower than 150 meV, nicknamed MP2018-stable, followed by training an autoencoder to derive a latent-space representation. The latent OFM features are subsequently used as targets to train a GNN model that generates these features directly from the initial structures.
- ℓ-MM featurizer: following a similar procedure to the OFM featurizer, we encode features obtained from the default MatMiner featurizer of MODNet v.0.1.13 applied to the MP2018-stable dataset, resulting in 1336 MatMiner features. The selected compression level provides latent MatMiner features (ℓ-MM), which are then used as targets to train a GNN model that directly generates these features from the original structures.

  The **DescriptorMEGNetFeaturizer** class in the MatterVial package is implemented to retrieve OFM-encoded and MatMiner-encoded features from the MatterVial package. A thorough investigation of these encoded features, including the determination of optimal compression levels and hyperparameters, was conducted, as detailed in the Supporting Information (Sections S4, S5, S7, and Fig. S3, also Tables S1–S3, S5–S8, S10–S11).
- MVL MatterVial featurizers: Utilizing the MVLFeaturizer class from the MatterVial package, we incorporate five pretrained MEGNet models provided by the Materials Virtual Lab51. Specifically, these are the models trained for the formation energy, Fermi energy, and elastic constants KVRH and GVRH on the 2019.4.1 MP crystals dataset, as well as the band gap regression model trained on the 2018.6.1 MP crystals dataset. The default MEGNet architecture comprises MEGNet blocks followed by an MLP with two dense layers, one with 32 neurons and the other with 16 neurons, before producing the target property (see Section S3, Fig. S2, in SI). The modularity of the MatterVial package allows us to extract features from different layers of these pretrained models. We extract features from the MLP layers preceding the output, specifically from the 32-neuron (layer32) and 16-neuron (layer16) configurations. An investigation was conducted on the effect of using the different layers for prediction, as provided in SI, Section S6, Table S4. In this study, the extracted features of both layers (160 descriptors for layer32 and 80 descriptors for layer16) are concatenated and added to the final feature vector.
- Adjacent GNN featurizer: The AdjacentGNNFeaturizer class from the MatterVial package is employed to train a MEGNet or coGN model on the fly for each fold of the train-test split. This adjacent model captures task-specific data nuances, enhancing prediction accuracy. The default hyperparameters from MEGNet v.1.3.2 and coGN are utilized, as detailed in SI, Section S7.2, Table S9.
- SISSO-based formula featurizer: The SISSO++ framework[43] was used to generate symbolic expressions that approximate target material properties across 15 datasets (see SI, Section S8, Table S12 for details) by transforming MatMiner features. The method begins by recursively applying a predefined set of operators (e.g., addition, subtraction, multiplication, division, sine, cosine, exponential, and logarithm) to expand the feature space, followed by sure-independence screening (SIS) that ranks the resulting candidates by their correlation with the target property and a sparsification step that selects a compact descriptor set. For our configuration, restricted to rung one, this yields 20 paired-feature formulas. By opting for the expressions produced at the SIS step instead of the final SISSO formula, versatility and generalization are assured when integrated with MODNet neural networks. These formulas, derived for each of the 15 tasks, are compiled in the file SISSO_FORMULAS_v1.txt, which is accessed by the get_sisso_features function in MatterVial to process the given MatMiner features (either directly or decoded from ℓ-MM) and outputs a dataframe of evaluated expressions.

In terms of computational cost, MatterVial is substantially more efficient than traditional MatMiner for generating a complete feature set. While

the precise runtime is dataset-dependent, our benchmarks (Table S14) show that MatterVial reduces the feature generation time by at least an order of magnitude. This speedup is most pronounced when leveraging GPU acceleration.

Retrieving interpretability via MatterVial's interpreter module

Before employing MatterVial's interpreter module, we conduct a SHAP value analysis (see SI, Section S9) on our MODNet models to assess feature importance. Using the SHAP Python library, we perform the analysis with 300 samples and 500 perturbations on 24 CPU cores in about 20 min, revealing the features with the greatest impact on the model predictions.

To bridge the gap between high-level latent representations and interpretable chemical descriptors, MatterVial leverages surrogate XGBoost models. These models are trained to predict each latent feature based on the previous assessment using the MP2018-stable dataset featurized with interpretable MatMiner and OFM features. The tree-based additive structure of XGBoost ensures rapid and parallel training as well as efficient SHAP calculations. For each feature, the top 30 most influential interpretable descriptors, as determined by SHAP, are forwarded to SISSO++, which performs symbolic regression to retrieve a symbolic formula that better correlates with the latent feature. This process is illustrated in Fig. 5. The SHAP decompositions and SISSO formula for the latent-space features computed this way can be retrieved by calling the **Interpreter** class and invoking ***get_shap_values*** or ***get_formula*** with the feature name as generated by MatterVial.

Moreover, this interpretability framework extends to adjacent GNN models. Using MatterVial's tools, including the AdjacentGNNFeaturizer, a task-specific GNN model is trained, and its latent features can be interpreted following the previous pipeline for which helper functions are provided. In this way, both the pretrained models imported by MatterVial and the adjacent models trained on the fly benefit from enhanced transparency, enabling users to decode the underlying chemical principles driving the predictions.

Data availability

The supplementary data repository, including detailed results, training logs, and fold-wise results, is available at https://github.com/rogeriog/MatterVial_SupportData. The MatterVial open-source Python package developed in this study is available at https://github.com/rogeriog/MatterVial.

Received: 2 September 2025; Accepted: 16 December 2025;
Published online: 15 January 2026

References

1. Rodrigues, J. F., Florea, L., De Oliveira, M. C. F., Diamond, D. & Oliveira, O. N. Big data and machine learning for materials science. *Discov. Mater.* **1**, 12 (2021).
2. Dey, A. et al. State of the art and prospects for halide perovskite nanocrystals. *ACS Nano* **15**, 10775–10981 (2021).
3. Guo, K., Yang, Z., Yu, C.-H. & Buehler, M. J. Artificial intelligence and machine learning in design of mechanical materials. *Mater. Horiz.* **8**, 1153–1172 (2021).
4. De Breuck, P.-P., Hautier, G. & Rignanese, G. M. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. *npj Comput. Mater.* **7**, 1–8 (2021).
5. Zhang, B., Zhou, M., Wu, J. & Gao, F. Predicting the materials properties using a 3D graph neural network with invariant representation. *IEEE Access* **10**, 62440–62449 (2022).
6. Tawfik, S. A. & Russo, S. P. Naturally-meaningful and efficient descriptors: machine learning of material properties based on robust one-shot ab initio descriptors. *J. Cheminform.* **14**, 78 (2022).
7. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
8. Pretto, T., Baum, F., Gouvêa, R. A., Brolo, A. G. & Santos, M. J. L. Optimizing the synthesis parameters of double perovskites with machine learning using a multioutput regression model. *J. Phys. Chem. C.* **128**, 7041–7052 (2024).
9. Liu, J. et al. Toward excellence of electrocatalyst design by emerging descriptor-oriented machine learning. *Adv. Funct. Mater.* **32**, 2110748 (2022).
10. Kim, G.-H., Lee, C., Kim, K. & Ko, D.-H. Novel structural feature-descriptor platform for machine learning to accelerate the development of organic photovoltaics. *Nano Energy* **106**, 108108 (2023).
11. Li, S. et al. Encoding the atomic structure for machine learning in materials science. *WIREs Comput. Mol. Sci.* **12**, e1558 (2022).
12. Dunn, A. MatBench Leaderboard. https://matbench.materialsproject.org/ (2024).
13. Riebesell, J. et al. A framework to evaluate machine learning crystal stability predictions. *Nat. Mach. Intell.* **7**, 836–847 (2025).
14. Lam Pham, T. et al. Machine learning reveals orbital interaction in materials. *Sci. Technol. Adv. Mater.* **18**, 756–765 (2017).
15. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
16. Kanter, J. M. & Veeramachaneni, K. Deep feature synthesis: towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* 1–10 (IEEE, Campus des Cordeliers, Paris, France, 2015).
17. Arik, S. Ö. & Pfister, T. TabNet: Attentive Interpretable Tabular Learning. AAAI **35**, 6679–6687 (2021).
18. Ruff, R., Reiser, P., Stühmer, J. & Friederich, P. Connectivity optimized nested graph networks for crystal structures. Preprint at http://arxiv.org/abs/2302.14102 (2023).
19. Ko, T. W. et al. Materials Graph Library (MatGL), an open-source graph deep learning library for materials science and chemistry. *npj Comput. Mater.* **11**, 253 (2025).
20. Goodall, R. E. A. & Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* **11**, 6280 (2020).
21. Rhodes, B. et al. Orb-v3: atomistic simulation at scale. Preprint at https://doi.org/10.48550/arXiv.2504.06231 (2025).
22. Shiota, T., Ishihara, K. & Mizukami, W. Universal neural network potentials as descriptors: towards scalable chemical property prediction using quantum and classical computers. *Digit. Discov.* **3**, 1714–1728 (2024).
23. El-Samman, A. M. et al. Global geometry of chemical graph neural network representations in terms of chemical moieties. *Digit. Discov.* **3**, 544–557 (2024).
24. El-Samman, A. M., De Castro, S., Morton, B. & De Baerdemacker, S. Transfer learning graph representations of molecules for pKa[13], C-NMR, and solubility. *Can. J. Chem.* **102**, 275–288 (2024).
25. Elijošius, R. et al. Zero shot molecular generation via similarity kernels. *Nat. Commun.* **16**, 5991 (2025).
26. Kim, S. Y., Park, Y. J. & Li, J. Leveraging neural network interatomic potentials for a foundation model of chemistry. Preprint at https://doi.org/10.48550/arXiv.2506.18497 (2025).
27. Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T. Interpretable and explainable machine learning for materials science and chemistry. *Acc. Mater. Res.* **3**, 597–607 (2022).
28. Pilania, G. Machine learning in materials science: from explainable predictions to autonomous design. *Comput. Mater. Sci.* **193**, 110360 (2021).
29. Abolhasani, M. & Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2**, 483–492 (2023).
30. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **6**, 1–10 (2020).

31. De Breuck, P.-P., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *J. Phys.: Condens. Matter* **33**, 404002 (2021).
32. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 11002 (2013).
33. Shen, J. et al. Reflections on one million compounds in the open quantum materials database (OQMD). *J. Phys. Mater.* **5**, 031001 (2022).
34. Schmidt, J. et al. Machine-learning-assisted determination of the global zero-temperature phase diagram of materials. *Adv. Mater.* **35**, 2210788 (2023).
35. Choudhary, K. et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).
36. Barroso-Luque, L. et al. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models. Preprint at https://doi.org/10.48550/arXiv.2410.12771 (2024).
37. matbench_v0.1: MegNet (kgcnn v2.1.0) - MatBench. *Materialsproject.org* https://matbench.materialsproject.org/Full%20Benchmark%20Data/matbench_v0.1_MegNet_kgcnn_v2.1.0/ (2020).
38. Xu, Y. & Qian, Q. i-SISSO: mutual information-based improved sure independent screening and sparsifying operator algorithm. *Eng. Appl. Artif. Intell.* **116**, 105442 (2022).
39. Jiang, X., Liu, G., Xie, J. & Hu, Z. Boosting SISSO performance on small sample datasets by using random forests prescreening for complex feature selection. Preprint at https://doi.org/10.48550/ARXIV.2409.19209 (2024).
40. Foppa, L., Purcell, T. A. R., Levchenko, S. V., Scheffler, M. & Ghiringhelli, L. M. Hierarchical symbolic regression for identifying key physical parameters correlated with bulk properties of perovskites. *Phys. Rev. Lett.* **129**, 055301 (2022).
41. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, San Francisco, CA, USA, 2016).
42. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
43. Purcell, T. A. R., Scheffler, M. & Ghiringhelli, L. M. Recent advances in the SISSO method and their implementation in the SISSO++ code. *J. Chem. Phys.* **159**, 114110 (2023).
44. Dunn, A. et al. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **6**, 138 (2020). Data originally curated by Citrine Informatics (https://citrination.com/datasets/153092/).
45. Choudhary, K., Kalish, I., Beams, R. & Tavazza, F. High-throughput identification and characterization of two-dimensional materials using density functional theory. *Sci. Rep.* **7**, 5179 (2017).
46. Petretto, G. et al. High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 180065 (2018).
47. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
48. Petousis, I. et al. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Sci. Data* **4**, 160134 (2017).
49. Kawazoe, Y., Masumoto, T., Tsai, A.-P., Yu, J.-Z. & Aihara, T. Jr. 1 Introduction. In *Nonequilibrium Phase Diagrams of Ternary Amorphous Alloys* (eds Kawazoe, Y. et al.) 37A 1–32 (Springer-Verlag, Berlin/Heidelberg, 1997).
50. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
51. De Jong, M. et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 150009 (2015).
52. Castelli, I. E. et al. New cubic perovskites for one- and two-photon water splitting using the computational materials repository. *Energy Environ. Sci.* **5**, 9034 (2012).

Acknowledgements

We acknowledge the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) for computational resources. This work also used Lucia, the Tier-1 supercomputer of the Walloon Region. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 001 and Université catholique de Louvain (Ref. No. ARH/MKK/01155839).

Author contributions

R.G., P.D.B., and G.M.R. conceptualized the study. R.G. performed the experiments and data analysis, and wrote the main manuscript. P.D.B. and G.M.R. supervised the investigation. R.G., P.D.B., T.P., G.M.R., and M.J.L.S. reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-025-01938-2.

Correspondence and requests for materials should be addressed to Rogério Almeida Gouvêa or Gian-Marco Rignanese.