

Broken Neural Scaling Laws in Learning the Optical Properties of Solids

Max Großmann¹,* Malte Grunert², and Erich Runge¹

Institute of Physics and Institute of Micro- and Nanotechnologies, Technische Universität Ilmenau, 98693 Ilmenau, Germany



(Received 26 March 2026; accepted 15 July 2026; published 13 August 2026)

In materials science, data are scarce and expensive to generate, whether computationally or experimentally. Therefore, it is crucial to identify how model performance scales with dataset size and model capacity to distinguish between data- and model-limited regimes. Neural scaling laws provide a framework for quantifying this behavior and guide the design of materials datasets and machine learning architectures. Here, we investigate neural scaling laws for predicting the optical properties of solids, using over 200 000 dielectric functions and Drude frequencies of metals from high-throughput *ab initio* calculations. We study the scaling of three multiobjective graph neural network architectures trained to simultaneously predict the frequency-dependent complex interband dielectric function as a continuous spectral response and the Drude frequency as a scalar quantity. Across two rotationally invariant graph neural networks that differ in their effective body order and an equivariant graph neural network evaluated at different maximum angular degrees, we consistently observe broken neural scaling laws with respect to dataset size. In contrast, scaling with respect to the number of model parameters is well described, in all cases, by saturating power laws.

DOI: [10.1103/wv8r-7zr1](https://doi.org/10.1103/wv8r-7zr1)

I. INTRODUCTION

Neural scaling laws (NSLs) quantify how the predictive performance of machine learning (ML) models depends on the resources used for training, primarily the amount of data, the number of model parameters, and the available compute. Their conceptual roots lie in statistical learning theory and early neural network research, where learning curves were studied theoretically and numerically to relate the generalization error to the number of training examples and model capacity [1–5]. While much of the early work relied on idealized assumptions or on datasets and models that are small by today's standards, modern NSL studies [6–10] extend these ideas empirically to large-scale ML settings. Although the theoretical origins of these empirical power laws remain an active area of research [11,12], NSLs are now widely used to guide model development across domains, including computer vision and large language models.

In many large-scale studies [6–10], NSLs are reported to follow simple power-law behavior over multiple orders of magnitude with respect to dataset size, model size, and the amount of compute used for training. However, this behavior is not universal. Earlier work on learning curves already discussed distinct asymptotic regimes, crossovers, and phase-transition-like changes in the scaling of the generalization error [1–4]. In a more modern setting, deviations from simple power laws, i.e., crossovers between scaling regimes with

different scaling exponents, have been reported for a number of tasks and learning settings and are often referred to as broken neural scaling laws (BNSLs), a term introduced, to the best of our knowledge, by Caballero *et al.* [13].

NSL analyses are also well established in chemistry and materials science, where they are used to assess the scaling behavior of ML models for tasks such as the prediction of molecular atomization energies [14], the generation of small molecules [15], and the training of machine-learned interatomic potentials (MLIPs) [16–24]. However, these studies have largely focused on molecular datasets or on material properties related to the potential energy surface (PES) of solids, such as energies, forces, and stresses. NSL studies for ML models that learn excited-state properties of solids beyond the PES remain limited, largely because such properties are difficult to compute systematically at large scale. A recent study on screening photovoltaic materials by Walker and Butler [25] provided an initial small-data analysis of how surrogate models for band gaps, absorption spectra, and derived efficiency metrics improve with increasing dataset size, but was limited to datasets of only a few thousand materials and did not aim to establish robust large-scale NSLs. It therefore remains unclear how ML models for excited-state properties of solids scale with increasing data availability.

To address this question, we investigate NSLs for the optical properties of metals as a representative class of excited-state properties of solids beyond the PES. Specifically, we focus on the complex-valued dielectric function $\epsilon(\omega)$ of metals, i.e., a physical response that encodes how solids interact with light and underpins a wide range of optical technologies. From an ML perspective, learning $\epsilon(\omega)$ for metals is challenging, as it constitutes a multiobjective task requiring the prediction of both the interband dielectric function and the intraband response, which is characterized by a Drude frequency [26]. In addition, this task provides a controlled

*Contact author: max.grossmann@tu-ilmenau.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

setting for studying the scaling behavior of multiobjective ML models in materials science [24], because the interband contribution defines a continuous spectral target whereas the Drude frequency defines a scalar target.

Previous work by us [27–29] and others [30–32] has shown that graph neural networks (GNNs) can efficiently learn optical spectra of semiconductors and insulators. However, the corresponding training datasets remain limited in size and accuracy, largely due to the computational cost of climbing the Jacob’s Ladder of optoelectronic properties [28]. Metals, in contrast, provide a physically well-controlled regime in which it is feasible to generate accurate optical data on a large scale using comparatively lower-level *ab initio* methods. This is facilitated by strong electronic screening in metals, which suppresses excitonic effects and often allows for quantitative agreement with experiments within the independent-particle approximation (IPA) augmented by a Drude term [33–35]. Consequently, the optical properties of metals combine technological relevance with the computational tractability needed for accurate large-scale data generation, making them well suited for probing NSLs under the data availability and generation constraints typical of materials science.

Based on these considerations, and to resolve data-scaling behavior over a broad range of dataset sizes, D , we generated a dataset containing 201 361 dielectric functions of metals using high-throughput *ab initio* calculations. This dataset exceeds existing datasets [27–32] for dielectric functions of semiconductors and insulators by about an order of magnitude.

To probe the scaling behavior with respect to model parameter count, N , and the role of architectural inductive biases, we employ three multiobjective GNNs. At this point, we note that throughout this study, the words “invariant” and “equivariant” refer to symmetries with respect to the special Euclidean group $SE(3)$. We use two invariant architectures based on our previous work [27–29], which were deliberately streamlined to minimize architectural degrees of freedom and enable controlled variation of N . These models differ in their use of two- and three-body interactions, allowing us to assess how increasing the effective body order affects both the predictive performance and NSLs. To test whether the NSLs change when transitioning from an invariant to an equivariant model architecture, we include an equivariant GNN based on the universal models for atoms (UMA) backbone, which builds on the ESEN architecture [20,36], and evaluate its performance for different maximum angular degrees $\ell_{\max}$. Here, $\ell_{\max}$ corresponds to $L_{\max}$ in the original UMA notation [20], with ℓ used to avoid confusion with the loss function L .

Together, the size of our dataset and the selected ML models allow us to investigate neural scaling behavior across a wide range of dataset sizes and test whether the observed scaling behavior is specific to a particular invariant architecture or persists across both invariant and equivariant representations. Across the three model architectures, we consistently observe BNSLs with respect to dataset size, whereas scaling with respect to the number of model parameters is well described by saturating power laws. This finding is significant for ML in materials science, where data generation is often the limiting factor. Furthermore, by resolving how performance changes with dataset size, model capacity, and architectural induc-

tive bias, the presented NSLs help identify whether further improvements are more likely to come from additional data, larger models, or more expressive architectures.

II. RESULTS

In the following, we analyze the scaling behavior of GNNs that predict $\varepsilon(\omega)$ of metals from structural information alone. As detailed in Sec. IV, the dielectric response can be separated into a complex interband contribution, $\varepsilon_{\text{inter}}(\omega)$, and a Drude term proportional to the square of the Drude frequency, ω_D . Both are in general tensors [32], but we restrict ourselves to rotationally invariant averages [27,28,37] (Greek indices refer to the Cartesian components): $\bar{\varepsilon}_{\text{inter}}(\omega) = \text{Tr}[\varepsilon_{\text{inter},\alpha\beta}(\omega)]/3$ and $\bar{\omega}_D^2 = \text{Tr}[\omega_{D,\alpha\beta}^2]/3$ (consistent with the quadratic appearance of ω_D in the intraband contributions), respectively.

A. Dataset

Extracting reliable NSLs requires large and robust datasets, which are scarce in materials science, particularly for optical properties. Even the largest existing datasets contain fewer than 30 000 spectra [29], which is insufficient to systematically probe scaling behavior across multiple orders of magnitude in dataset size.

To address this limitation, we generated a large-scale dataset of dielectric functions for 205 224 intermetallic compounds using high-throughput *ab initio* calculations based on structures filtered from the ALEXANDRIA database [38,39] (downloaded March 2024). All calculations were automated using custom in-house workflows to ensure full reproducibility. Converged calculations for $\bar{\varepsilon}_{\text{inter}}(\omega)$ and $\bar{\omega}_D$ were obtained for 201 361 materials, while for the remaining 3863 materials the IPA part of the workflow could not be converged within the available computational budget. The structural filtering criteria, calculation methodology, and design of the high-throughput workflow are described in Sec. IV. Supplemental Note 1 of the Supplemental Material [40] provides statistics on the distribution of elements and the distribution of the number of atoms per unit cell across the dataset. The generated dataset and workflow code are available through the repositories provided in the Data Availability statement.

To validate the quality of the employed calculation methodology and, consequently, of the resulting dataset, we compared the dielectric functions of 27 elemental metals computed with our workflow against experimental data (see Supplemental Note 2 of the Supplemental Material [40]). Overall, we find good agreement between the calculated and experimental spectra in both peak positions and spectral shape, consistent with the results of Prandini *et al.* [34].

B. Model setup

We investigate NSLs using three GNN architectures: two invariant architectures and one equivariant architecture derived from a state-of-the-art MLIP.

For the invariant models, we build upon our previous OPTIMATE models for the optical spectra of semiconductors and insulators [27–29], adapting the architecture to the present multiobjective task and introducing two variants, OPTIMETAL2B and OPTIMETAL3B. OPTIMETAL2B

follows the original two-body message-passing formulation of OPTIMATE, whereas OPTIMETAL3B extends this framework by explicitly incorporating three-body interactions, building on our recent OPTIMATE3B developments [29]. Both models share the same design philosophy and training protocol, differing only in their effective body order (two-body versus three-body interactions) and, for selected variants, their message-passing formulation (crystal graph convolution [79], CGC, versus transformer convolution [80], TC). The basis for both OPTIMETAL2B and OPTIMETAL3B is a set of invariant input features (atom types and bond lengths for both models, with bond angles additionally included for OPTIMETAL3B), which enables us to isolate the effect of body order on model performance and scaling behavior. To investigate NSLs in a controlled setting, we deliberately constrained the architectural degrees of freedom of OPTIMETAL2B and OPTIMETAL3B, whose architectures are described in detail in Supplemental Note 3 of the Supplemental Material [40]. Prior to the NSL analysis, the architectural components and training hyperparameters of the OPTIMETAL models were systematically optimized at a fixed model width (see Supplemental Note 4 of the Supplemental Material [40]), ensuring that the resulting scaling laws reflect near-optimal model performance rather than architectural inefficiencies.

For the equivariant architecture, we use an adapted UMA model, whose backbone builds on the equivariant ESEN architecture [20,36]. We refer to Fu *et al.* [36] and Wood *et al.* [20] for architectural details and describe our task-specific modifications in Sec. IV. For the NSL analysis presented below, the adapted UMA backbone was kept fixed, with the maximum angular degree $\ell_{\max} \in \{0, 1, 2, 3\}$ as the only additional architectural variable besides model width. In line with the OPTIMETAL models, the training hyperparameters were optimized at a fixed model width using the workflow described in Supplemental Note 4 of the Supplemental Material [40].

All models trained on the full training set accurately predict $\bar{\epsilon}_{\text{inter}}(\omega)$, $\bar{\omega}_{\text{D}}$, and material colors on the held-out test set. Detailed performance metrics and representative predictions are reported in Supplemental Note 5 of the Supplemental Material [40].

C. Neural scaling laws

Next, we quantify how the predictive performance of the GNNs scales with dataset size, D , and parameter count, N , across model architectures. For data scaling, we randomly subsampled the training set into seven subsets containing 2500, 5000, 10 000, 20 000, 40 000, 80 000, and 160 000 materials, respectively, and used these subsets consistently throughout the subsequent NSL analysis. To assess whether these subsets remained representative, we compared the distribution of elements and the distribution of the number of atoms per unit cell, finding no substantial deviations (see Supplemental Note 6 of the Supplemental Material [40]). The parameter count was controlled by varying the hidden dimension d_h (cf. Supplemental Note 3 of the Supplemental Material [40]) in powers of 2 from 16 to 1024 for the OPTIMETAL models, corresponding to models with approximately $N \approx 100\text{ k} - 100\text{ M}$ trainable parameters. For the adapted UMA model, d_h was varied in powers of 2 from 16 to 256, resulting

in models with approximately $N \approx 100\text{ k} - 50\text{ M}$ trainable parameters, depending on $\ell_{\max}$. The upper end of this range was limited by the higher memory requirements of the equivariant architecture, since the adapted UMA model with $d_h \geq 512$ exceeded the 40 GB memory of the NVIDIA A100 graphics processing units (GPUs) under the fixed training setup and batch size used for all NSL experiments (see Sec. IV).

First, we analyze one-dimensional (1D) NSLs, i.e., we investigate $L(D)$ while holding N constant and vice versa. The loss function L is defined as the sum of mean absolute errors (MAE) for the interband dielectric function and the Drude frequency (see Sec. IV). For data scaling, i.e., $L(D)$, we use a fixed hidden dimension $d_h = 256$ for all model architectures, which corresponds to $N \approx 10\text{ M}$ trainable parameters and matches d_h used in the OPTIMETAL architecture optimization described in Supplemental Note 4 of the Supplemental Material [40]. For parameter scaling, i.e., $L(N)$, we fix the dataset size to $D = 20\,000$.

Before presenting the 1D NSLs, we describe how the functional forms used to fit the validation loss, L_{val} , were chosen. For data scaling, this choice is central to identifying whether the observed scaling behavior follows a conventional power law or a BNSL. Inspired by previous studies on NSLs [6,7,9,13,15,22], we evaluated four candidate functions for the data-scaling analysis [see Eqs. (8)–(11) in Sec. IV]: a simple power law, a simple power law with a saturation floor, and two smoothly broken power laws, one without and one with an adjustable amplitude. To determine the function that best describes the data-scaling behavior, we compared all fitted functional forms using the Akaike information criterion corrected (AICc) for small sample sizes [81]. The AICc balances fit quality with the number of parameters used to achieve it by penalizing overly complex functions that would otherwise overfit the data. The fit with the lowest AICc value was selected as the preferred functional form. In contrast, for parameter scaling, we do not perform an AICc-based functional-form selection but instead use the expected simple power law with a saturation floor, i.e., Eq. (9), reflecting the expected scaling behavior when increasing N at fixed D : Once the model capacity becomes sufficient for the available data, further increases in N yield diminishing returns and the validation loss approaches an effective finite-data irreducible loss. Accordingly, we interpret the parameter-scaling results in terms of finite-data saturation rather than broken scaling. The functional forms of the 1D NSLs are defined in Sec. IV, and Supplemental Note 7 of the Supplemental Material [40] reports the AICc values for all tested data-scaling fits.

Figure 1 summarizes the resulting 1D NSLs for data scaling, $L_{\text{val}}(D)$, and parameter scaling, $L_{\text{val}}(N)$, shown in the top and bottom panels, respectively, across the OPTIMETAL architectures and one representative adapted UMA model with $\ell_{\max} = 2$. The corresponding fit parameters are given in Table I.

To assess the sensitivity of NSLs to the message-passing operator, we compare OPTIMETAL2B variants with CGC [79] and TC [80] message passing in Fig. 1, as both yield nearly identical performance after the architecture optimization (see Supplemental Note 4 of the Supplemental Material [40]). Furthermore, we compare OPTIMETAL2B and OPTIMETAL3B to assess how increasing the effective body order from two-

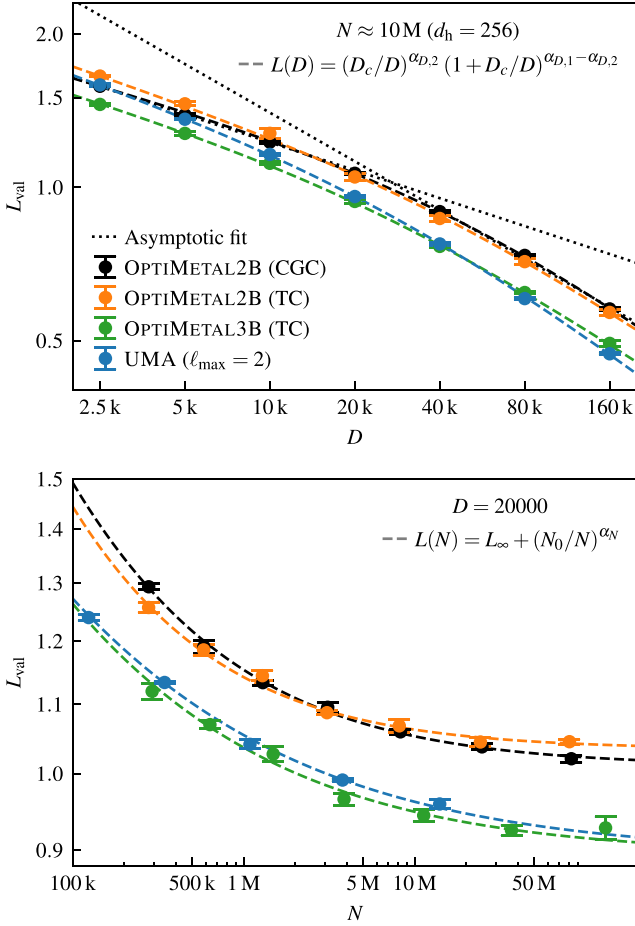


FIG. 1. 1D NSLs for dataset size and parameter scaling across model architectures. The top panel shows the validation loss, L_{val} , as a function of dataset size, D , at fixed parameter count, N . The bottom panel shows L_{val} as a function of N at fixed D . Dots represent L_{val} averaged over three random model initializations, and error bars indicate the standard deviation. The dashed lines show the corresponding NSL fits, whose functional form is given in the panel legends (see Sec. IV for details). The corresponding fit parameters are given in Table I. In the top panel, we show two conventional NSLs (simple power laws) fitted to the first and last three data points of OPTIMETAL2B (CGC) (dotted black lines, annotated as “Asymptotic fit”), to highlight the “brokenness” of the neural scaling behavior.

body to three-body interactions affects the NSLs. Finally, we include the adapted UMA model with $\ell_{\text{max}} = 2$ to determine whether the use of an equivariant model architecture affects the scaling behavior of the invariant prediction task considered here.

Based on the AICc analysis described above, we found that 1D data scaling across all models shown in the top panel of Fig. 1 is best described by a smoothly broken power law without an adjustable amplitude. OPTIMETAL2B (CGC and TC), OPTIMETAL3B, and the adapted UMA model with $\ell_{\text{max}} = 2$ exhibit similar scaling exponents in the low-data regime, with $\alpha_{D,1} \approx 0.15$ – 0.19 , suggesting that all models extract information from small datasets with comparable efficiency. Above the model-dependent crossover point $D_c = 26.6$ – 52.4 k, the scaling slope steepens significantly to $\alpha_{D,2} =$

TABLE I. Fit parameters of the 1D NSLs across model architectures. The reported parameters are obtained from nonlinear least-squares fits to validation losses averaged over three random model initializations. For data scaling, we use the smoothly broken power law in Eq. (10), following the AICc-based functional-form selection described in the text. This fit yields the low- and high-data exponents $\alpha_{D,1}$ and $\alpha_{D,2}$, as well as the crossover point D_c . For parameter scaling, we use the expected saturating power law in Eq. (9), characterized by the exponent α_N , the characteristic parameter count N_0 , and the irreducible loss L_∞ . The corresponding 1D NSL fits are shown in Figs. 1 and 2.

Data scaling	$\alpha_{D,1}$	$\alpha_{D,2}$	D_c
OPTIMETAL2B (CGC)	0.15	0.42	52.4 k
OPTIMETAL2B (TC)	0.18	0.39	42.1 k
OPTIMETAL3B (TC)	0.17	0.38	26.6 k
UMA ($\ell_{\text{max}} = 0$)	0.16	0.50	96.9 k
UMA ($\ell_{\text{max}} = 1$)	0.18	0.46	49.3 k
UMA ($\ell_{\text{max}} = 2$)	0.19	0.43	30.0 k
UMA ($\ell_{\text{max}} = 3$)	0.19	0.41	25.1 k
Parameter scaling	α_N	N_0	L_∞
OPTIMETAL2B (CGC)	0.53	25.1 k	1.01
OPTIMETAL2B (TC)	0.58	21.5 k	1.03
OPTIMETAL3B (TC)	0.41	9.0 k	0.89
UMA ($\ell_{\text{max}} = 0$)	0.27	689	1.12
UMA ($\ell_{\text{max}} = 1$)	0.40	6.4 k	1.00
UMA ($\ell_{\text{max}} = 2$)	0.37	7.4 k	0.89
UMA ($\ell_{\text{max}} = 3$)	0.66	22.6 k	0.92

0.38–0.43, indicating a shift to a regime in which additional data yields disproportionately larger performance gains. In this high-data regime, the scaling behavior appears to be largely model independent, with similar scaling exponents across all models. Consistent with this, the data scaling of the TC- and CGC-based variants of OPTIMETAL2B is nearly identical. The CGC-based variant performs slightly better in the low-data regime, although it has a slightly lower scaling exponent $\alpha_{D,1}$. The addition of three-body interactions in OPTIMETAL3B improves performance through a constant shift of the scaling curve toward lower validation losses. However, within the 1D NSL analysis, three-body interactions have no significant impact on the scaling exponents, a conclusion that is revised in the subsequent 2D NSL analysis. Since the equivariant adapted UMA model exhibits similar qualitative and quantitative BNSL behavior to the invariant OPTIMETAL models, this suggests that the equivariant representation does not measurably improve the 1D data-scaling behavior for the invariant prediction task considered here. However, as for the effect of three-body interactions, this conclusion is revised in the subsequent 2D NSL analysis.

Looking at the bottom panel of Fig. 1, we observe that parameter scaling follows the expected behavior. All models follow a simple power-law trend with exponents $\alpha_N = 0.37$ – 0.58 up to approximately 5 M parameters, beyond which the validation loss begins to saturate. This behavior is consistent with finite-data saturation, in which the available training data, rather than model capacity, limits further improvement at a fixed D . The validation loss with respect to the number

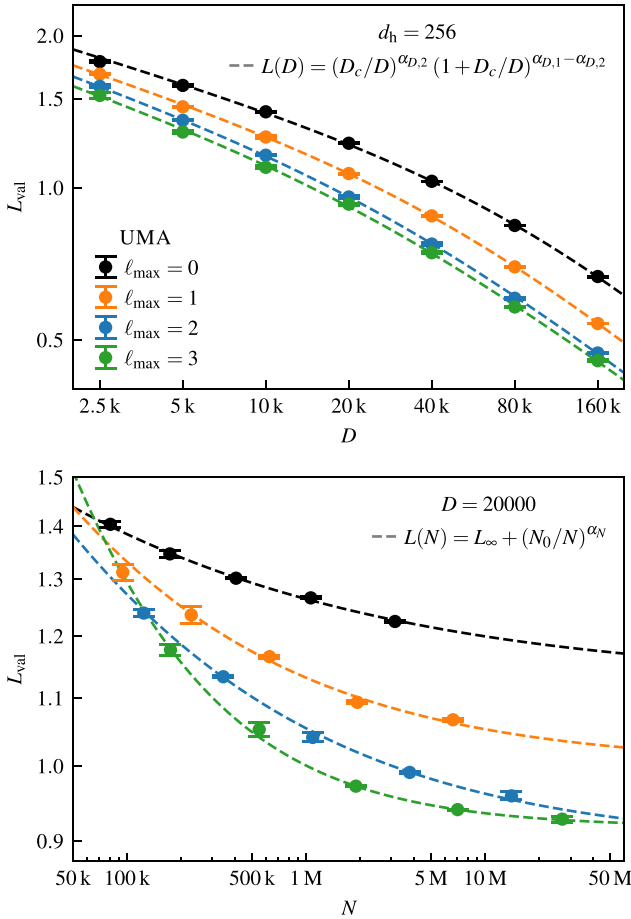


FIG. 2. 1D NSLs for dataset size and parameter scaling across the adapted UMA model at different $\ell_{\max}$. The top panel shows the validation loss, L_{val} , as a function of dataset size, D , at fixed parameter count, N . The bottom panel shows L_{val} as a function of N at fixed D . Dots represent L_{val} averaged over three random model initializations, and error bars indicate the standard deviation. The dashed lines show the corresponding NSL fits, whose functional form is given in the panel legends (see Sec. IV for details). The corresponding fit parameters are given in Table I.

of model parameters saturates more slowly for OPTIMETAL3B with $\alpha_N = 0.41$ and the adapted $\ell_{\max} = 2$ UMA model with $\alpha_N = 0.37$ compared to the OPTIMETAL2B models with $\alpha_N = 0.53$ – 0.58 . As with data scaling, the parameter scaling of the TC- and CGC-based variants of OPTIMETAL2B is quite similar. However, the TC-based variant now performs slightly better in the low-parameter regime, though it saturates at a higher validation loss as N increases. The adapted $\ell_{\max} = 2$ UMA model reaches the same fitted irreducible loss as OPTIMETAL3B, with $L_\infty = 0.89$ for both models, suggesting that the equivariant architecture does not reduce the finite-data saturation limit relative to the invariant three-body model in this 1D parameter-scaling analysis.

To assess the effect of $\ell_{\max}$ on the scaling behavior of the adapted UMA architecture, Fig. 2 shows the corresponding 1D NSLs for the adapted UMA model with $\ell_{\max} \in \{0, 1, 2, 3\}$. The corresponding fit parameters are also included in Table I. Consistent with the preceding 1D NSL analysis across dif-

ferent model architectures, the adapted UMA model exhibits broken data scaling and saturating parameter scaling for all $\ell_{\max}$. In the following, we focus on how increasing $\ell_{\max}$ affects the NSLs.

Examining the top panel of Fig. 2, we find that the low-data exponents remain in a narrow range of $\alpha_{D,1} = 0.16$ – 0.19 , while increasing slightly with $\ell_{\max}$. Increasing $\ell_{\max}$ also improves performance by shifting the scaling curves toward lower validation losses. This improvement is accompanied by a significant reduction in the crossover point, from $D_c = 96.9$ k for $\ell_{\max} = 0$ to 25.1 k for $\ell_{\max} = 3$, indicating that increasing the maximum angular degree reduces the amount of training data required to enter the favorable high-data regime. At the same time, the high-data exponent decreases from $\alpha_{D,2} = 0.50$ for $\ell_{\max} = 0$ to $\alpha_{D,2} = 0.41$ for $\ell_{\max} = 3$, approaching the values found for the invariant OPTIMETAL models. The observed improvements in performance, the increase in $\alpha_{D,1}$, and the earlier onset of the high-data regime largely saturate beyond $\ell_{\max} = 2$, as the scaling curves and fit parameters for $\ell_{\max} = 2$ and $\ell_{\max} = 3$ are very similar.

The 1D parameter scaling in the bottom panel of Fig. 2 shows how the finite-data saturation limit depends on $\ell_{\max}$. Increasing $\ell_{\max}$ from 0 to 2 lowers the fitted irreducible loss from $L_\infty = 1.12$ to $L_\infty = 0.89$, showing that higher angular degrees improve the finite-data saturation limit of the adapted UMA architecture. The adapted $\ell_{\max} = 2$ model reaches the same fitted irreducible loss as OPTIMETAL3B, with $L_\infty = 0.89$. For $\ell_{\max} = 3$, the fitted irreducible loss is slightly higher at $L_\infty = 0.92$, although this small difference in the extrapolation should be interpreted cautiously, as larger $\ell_{\max} = 3$ models would be required to determine the finite-data asymptotic limit more robustly. Together with the data-scaling results, this suggests that the benefit of increasing $\ell_{\max}$ likely diminishes for larger $\ell_{\max}$, with only limited additional gains expected beyond $\ell_{\max} = 3$ for the invariant optical-property prediction task considered here.

In addition to the 1D NSLs for the total multiobjective validation loss, we verified that its two components exhibit the same qualitative scaling behavior when analyzed separately using the presented multiobjective models. The loss contributions for both the continuous interband dielectric function and the scalar Drude frequency exhibit broken data scaling and saturating parameter scaling, demonstrating that the observed NSLs are not dominated by one part of the multiobjective loss alone (see Supplemental Note 8 of the Supplemental Material [40]).

Since the hidden dimension was set to $d_h = 256$ (i.e., $N \approx 10$ M parameters) for the 1D data-scaling experiments, the models operate in a strongly overparametrized regime across all dataset sizes. In the low-data regime, however, the limited training data may not uniquely constrain the structure-property relationship, possibly allowing many parameter configurations to fit the data equally well. As a result, the models may exhibit a type of “best-guess” behavior, capturing broad, global trends correlated with the optical response. Meanwhile, finer spectral features remain inadequately constrained, which is consistent with the small-data-scaling exponent observed in the low-data regime.

In contrast, the increase of the data-scaling exponent above D_c suggests a transition to a data-efficient regime, in which

larger datasets sample a broader range of structural and electronic environments, enabling the learning of more detailed and complex structure-property correlations. In this regime, each additional material may provide increasingly informative constraints, resulting in the second, steeper branch of the scaling curve.

This interpretation raises the question of whether the observed BNSLs persist when both D and N are varied simultaneously. If the BNSLs were primarily an artifact of performing data scaling with overparametrized models, then models whose capacity is well matched to the available data should exhibit simple power-law scaling. To test this, we analyze two-dimensional (2D) NSL maps, i.e., $L(D, N)$, which allow us to disentangle the interplay between dataset size and model capacity, thereby determining whether the observed BNSLs in data scaling are intrinsic to the learning task or specific to the fixed-model-size setting.

Since the CGC- and TC-based variants of OPTIMETAL2B exhibit nearly identical scaling behavior in the 1D NSLs, we focus the 2D NSL analysis on the TC-based architectures of OPTIMETAL2B and OPTIMETAL3B. This choice isolates the effect of increasing the effective body order from two to three. To assess whether the 2D NSLs change when transitioning from invariant to equivariant representations, we additionally analyze the adapted UMA model with $\ell_{\max} \in \{0, 1, 2, 3\}$.

To determine the functional form of the 2D NSL maps, we used the AICc to compare two candidate forms: a smoothly interpolating form inspired by Kaplan *et al.* [7] and an additive form following Hoffmann *et al.* [9] [see Eqs. (14) and (15)]. Based on this comparison, the Kaplan-type formulation was selected for the analysis presented here to provide a consistent description of $L(D, N)$ across all model architectures. The full AICc comparison, including the sensitivity of the adapted UMA fits to the smallest-width models, is reported in Supplemental Note 7 of the Supplemental Material [40].

The resulting Kaplan-type 2D NSL maps based on Eq. (14) are shown in Fig. 3. For clarity, the main figure shows OPTIMETAL2B, OPTIMETAL3B, and the adapted UMA model with $\ell_{\max} = 2$ and $\ell_{\max} = 3$, while the corresponding 2D NSL maps for $\ell_{\max} = 0$ and $\ell_{\max} = 1$ are shown in Supplemental Note 9 of the Supplemental Material [40]. In each subfigure, the top panel shows $L(N)$ at different dataset sizes D , whereas the bottom panel shows $L(D)$ at different parameter counts N . The corresponding fit parameters for all models are given in Table II.

Looking at the top panel of each subfigure of Fig. 3, we observe that, across all investigated model architectures, the performance improvement with increasing N slows down once $N \gtrsim 10$ M, a trend that appears to be largely independent of dataset size. However, the extent to which increasing N improves performance prior to saturation strongly depends on the amount of training data: Small datasets exhibit weak parameter scaling, whereas larger datasets benefit substantially more from increased model capacity before performance gains diminish. Compared with the 1D parameter-scaling exponents given in Table I, the 2D NSL maps yield smaller effective parameter-scaling exponents, indicating that the apparent benefit of increasing N is reduced when the dataset-size dependence of the scaling is taken into account. For the OPTIMETAL models, the 2D NSL analysis yields simi-

TABLE II. Fit parameters of the 2D NSL maps across model architectures. The reported parameters are obtained from nonlinear least-squares fits to validation losses averaged over three random model initializations. The Kaplan-type 2D NSL in Eq. (14) was used to describe $L(D, N)$ across all model architectures, following the AICc-based functional-form selection described in the text. This functional form combines the 1D data- and parameter-scaling laws shown in Figs. 1 and 2, but omits the irreducible loss term L_∞ in the parameter NSL, as explained in Sec. IV. This fit yields the low- and high-data exponents $\alpha_{D,1}$ and $\alpha_{D,2}$, as well as the crossover point D_c , along with the parameter-scaling exponent α_N and the characteristic parameter count N_0 . The corresponding 2D NSL maps are shown in Fig. 3 and Supplemental Note 9 of the Supplemental Material [40].

Model	α_N	$\alpha_{D,1}$	$\alpha_{D,2}$	N_0	D_c
OPTIMETAL2B (TC)	0.30	0.20	0.37	226.7 k	32.6 k
OPTIMETAL3B (TC)	0.33	0.16	0.46	137.4 k	29.4 k
UMA ($\ell_{\max} = 0$)	0.34	0.16	0.49	61.0 k	86.1 k
UMA ($\ell_{\max} = 1$)	0.27	0.16	0.60	37.1 k	51.2 k
UMA ($\ell_{\max} = 2$)	0.23	0.17	0.61	27.1 k	35.4 k
UMA ($\ell_{\max} = 3$)	0.25	0.17	0.53	31.5 k	30.3 k

lar parameter-scaling exponents of $\alpha_N = 0.30$ and 0.33 for OPTIMETAL2B and OPTIMETAL3B, respectively, indicating that increasing the effective body order does not significantly affect parameter scaling. For the adapted UMA models with $\ell_{\max} = 2$ and $\ell_{\max} = 3$, the corresponding exponents are slightly smaller, with $\alpha_N = 0.23$ and 0.25 , respectively. This suggests that, once $\ell_{\max}$ is fixed, increasing model width yields smaller performance gains for the equivariant architecture than for the invariant OPTIMETAL models within the 2D NSL analysis.

Turning to data scaling, shown in the bottom panel of each subfigure in Fig. 3, we observe a pronounced dependence on model capacity. For the smallest models, $L(D)$ initially follows a simple power law in the small-data regime but begins to saturate above $D = 40$ k, reflecting the limited capacity of the model in the high-data, low-parameter limit. This saturation is more pronounced for the invariant OPTIMETAL models than for the adapted UMA model. In the intermediate-parameter regime ($N \approx 1$ M), this bottleneck is lifted: All architectures exhibit simple power-law data-scaling behavior, indicating that model capacity and available data are well matched. As the parameter count N increases further, the data-scaling slope steepens, and for sufficiently overparametrized models, BNSLs emerge. This effect is more pronounced for OPTIMETAL3B than for OPTIMETAL2B, consistent with the larger separation between its fitted exponents ($\alpha_{D,1} = 0.16$ and $\alpha_{D,2} = 0.46$) compared to OPTIMETAL2B ($\alpha_{D,1} = 0.20$ and $\alpha_{D,2} = 0.37$). In contrast to the 1D NSLs, this suggests that increasing the effective body order improves data scaling in the high-data regime. For the adapted UMA model, the separation between the low- and high-data exponents is even larger, with $\alpha_{D,1} = 0.17$ and $\alpha_{D,2} = 0.61$ for $\ell_{\max} = 2$, and $\alpha_{D,1} = 0.17$ and $\alpha_{D,2} = 0.53$ for $\ell_{\max} = 3$. Thus, in contrast to the 1D data-scaling analysis, the 2D NSL analysis suggests that the adapted equivariant UMA architecture improves data scaling more strongly in the high-data regime than the invariant OPTIMETAL architectures. This is reflected by

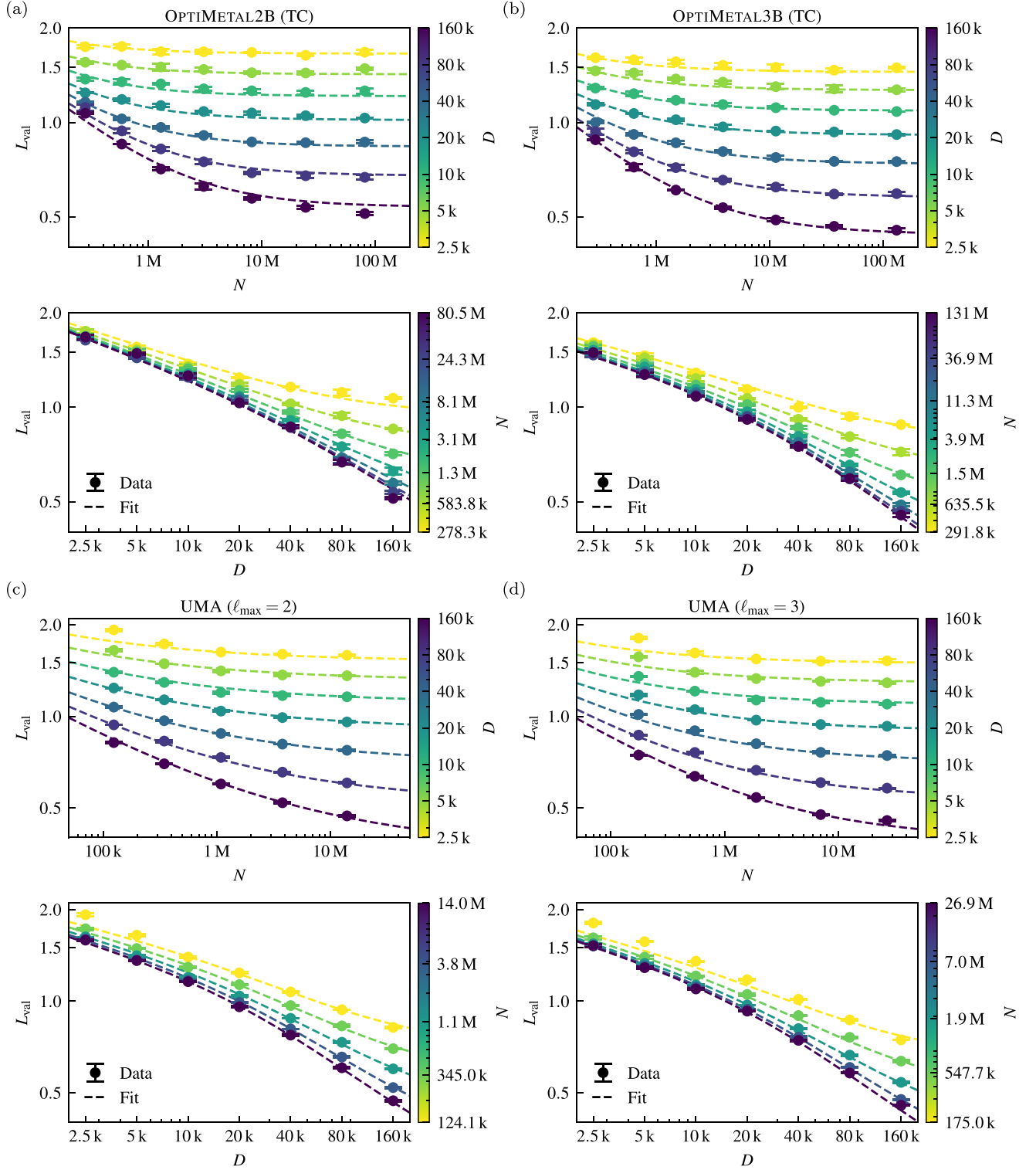


FIG. 3. 2D NSL maps across model architectures. The validation loss, L_{val} , is shown as a function of dataset size, D , and parameter count, N , for (a) OPTiMETAL2B (TC), (b) OPTiMETAL3B (TC), and the adapted UMA model with (c) $\ell_{\text{max}} = 2$ and (d) $\ell_{\text{max}} = 3$. Dots represent L_{val} averaged over three random model initializations, and error bars indicate the standard deviation. The dashed lines show the corresponding Kaplan-type 2D NSL fits based on Eq. (14) (see Sec. IV for details). Each subfigure provides two complementary views of the scaling behavior: The top panel shows the validation loss for models with different dataset sizes as a function of parameter count, as indicated by the color bar. The bottom panel swaps the roles of dataset size and parameter count. The corresponding fit parameters are given in Table II.

the larger high-data exponents of $\alpha_{D,2} = 0.61$ and 0.53 for $\ell_{\max} = 2$ and $\ell_{\max} = 3$, respectively, compared to $\alpha_{D,2} = 0.46$ for OPTIMETAL3B. Consistent with the 1D NSL analysis, the crossover point decreases with increasing $\ell_{\max}$, from $D_c = 86.1$ k for $\ell_{\max} = 0$ to $D_c = 30.3$ k for $\ell_{\max} = 3$, indicating that higher angular degrees reduce the amount of data required to enter the high-data-scaling regime.

Together, the 1D and 2D NSL analyses demonstrate that the broken data scaling observed in the 1D setting persists when the dataset size and parameter count are varied jointly, suggesting genuine multiregime learning behavior. The 2D NSL maps reveal that simple power-law data scaling emerges when dataset size and model capacity are well matched, whereas BNSLs arise in the overparametrized regime. Overall, the global scaling behavior of the invariant and equivariant GNNs trained to predict the optical properties of metals is best described by a combination of smoothly broken data scaling and saturating parameter scaling, with the sharpness of the data-scaling transition depending on the architectural inductive bias. The 2D NSL analysis further shows that architectural changes, such as increasing the effective body order or transitioning from invariant to equivariant architectures, affect data scaling in the high-data regime. In contrast, the low-data regime and parameter scaling are comparatively insensitive to both body order and equivariant representations. This distinction only becomes apparent when D and N are varied simultaneously.

III. DISCUSSION

In this study, we examine neural scaling behavior across invariant and equivariant GNN architectures using over 200 000 *ab initio* dielectric functions of metals as a representative task for learning the excited-state properties of solids beyond the PES.

Our key finding is the emergence of BNSLs in data scaling across model architectures. Rather than following a simple power law, model performance exhibits distinct scaling regimes, with a pronounced steepening of the data-scaling slope above a characteristic dataset size. In contrast, parameter scaling follows a saturating power law, indicating that, in the data-scarce regime of materials science, increasing model size is only beneficial up to a point, after which it yields progressively diminishing returns. By comparing invariant GNNs that differ only in their effective body order and message-passing operator, we further isolate the influence of the model architecture on the neural scaling behavior. Variations in the message-passing operator have little influence on the observed scaling trends, whereas increasing the effective body order improves data scaling in the high-data regime while leaving parameter scaling largely unchanged. Extending this comparison to the adapted UMA model shows that the equivariant architecture can further improve high-data scaling, while exhibiting similar saturating parameter scaling. Notably, the dependence of the NSLs on architectural inductive bias becomes apparent only in 2D NSL analyses that jointly consider dataset size and model capacity.

These findings are particularly relevant for materials science, where training data are scarce and costly to generate, yet model sizes remain modest compared to those in other

ML domains. Although increasing model capacity or adopting more expressive architectural inductive biases can provide rapid performance gains for a fixed dataset, this strategy may be impractical for many applications due to increased training cost and slower inference, especially for MLIPs [20,82,83]. Therefore, looking ahead, data availability and generation are likely to remain the primary bottlenecks for improving the predictive performance of ML models in materials science, motivating strategies that improve data efficiency.

In this context, two particularly active and competitive areas of research stand out. The first is the development of architectures with improved physical and geometric inductive biases, including higher-body-order message passing [84], equivariant neural networks [85,86], and architectures that more explicitly capture long-range interactions [87,88]. The second is the strategic use of Δ -ML [89,90] and transfer learning [28,91–94]. These approaches use correlations between related tasks or levels of theory to reduce the amount of high-fidelity target data needed for accurate predictions. From the perspective of the present NSL analysis, such approaches can be viewed as attempts to improve the effective data scaling or to move the crossover point to smaller high-fidelity dataset sizes. This makes them particularly attractive in domains where high-fidelity observables are computationally or experimentally expensive to obtain, such as optical spectra.

Finally, while this study focuses on the optical properties of metals, applying the same NSL analysis to other learning objectives in materials science remains an important direction for future research. Of particular interest are, in our opinion, multiobjective tasks [24] that combine different target domains, such as band gaps and other electronic properties, with properties related to the PES, such as energies, forces, and stresses. In such multidomain training settings, different combinations of task complexity, data availability, target noise, and model expressivity may lead to qualitatively different scaling regimes. Understanding these regimes could ultimately help guide the development of more general foundation models for materials science.

IV. METHODS

A. Dielectric functions of metals

The linear dielectric response of a homogeneous solid in the optical limit can be expressed via the frequency-dependent dielectric tensor $\varepsilon_{\alpha\beta}(\omega)$. Within the IPA, and neglecting microscopic local-field effects [33], the total response, $\varepsilon_{\alpha\beta}$, is usually separated into interband, $\varepsilon_{\text{inter},\alpha\beta}(\omega)$, and intraband, $\varepsilon_{\text{intra},\alpha\beta}(\omega)$, contributions.

$$\varepsilon_{\alpha\beta}(\omega) = \varepsilon_{\text{inter},\alpha\beta}(\omega) + \varepsilon_{\text{intra},\alpha\beta}(\omega). \quad (1)$$

Within the abovementioned limits and approximations, the interband term describes transitions between different bands and is given by [34]

$$\varepsilon_{\text{inter},\alpha\beta}(\omega) = \delta_{\alpha\beta} - \frac{4\pi}{V} \sum_{\mathbf{k}} \sum_{\substack{n,n' \\ n \neq n'}} \frac{M_{n\mathbf{k},\alpha}^* M_{n'\mathbf{k},\beta}}{(E_{n'\mathbf{k}} - E_{n\mathbf{k}})^2} \times \frac{f_{n\mathbf{k}} - f_{n'\mathbf{k}}}{\omega - (E_{n'\mathbf{k}} - E_{n\mathbf{k}}) + i\gamma_{\text{inter}}}, \quad (2)$$

where $E_{n\mathbf{k}}$ are band energies and $f_{n\mathbf{k}}$ are Fermi-Dirac occupations of Bloch states $\Psi_{n\mathbf{k}}$ with band index n at wave vector $\mathbf{k}$ in the Brillouin zone. V is the unit-cell volume and γ_{inter} is an *ad hoc* numerical broadening parameter. $M_{m\nu\mathbf{k},\alpha} = \langle \Psi_{m\mathbf{k}} | v_\alpha | \Psi_{n\mathbf{k}} \rangle$ are the matrix elements of the velocity operator $v_\alpha = -i[r_\alpha, H_{\text{KS}}]$ along Cartesian direction α , where H_{KS} is the Kohn-Sham (KS) Hamiltonian.

The intraband contribution arises from carriers at the Fermi surface and is approximated by a Drude model [34]

$$\varepsilon_{\text{intra},\alpha\beta}(\omega) = -\frac{\omega_{\text{D},\alpha\beta}^2}{\omega(\omega + i\gamma_{\text{D}})}, \quad (3)$$

with the Drude frequency tensor

$$\omega_{\text{D},\alpha\beta}^2 = \frac{4\pi}{V} \sum_{\mathbf{k},n} \left(-\frac{\partial f_{n\mathbf{k}}}{\partial E_{n\mathbf{k}}} \right) M_{m\mathbf{k},\alpha}^* M_{m\mathbf{k},\beta}, \quad (4)$$

where γ_{D} is the Drude damping, an *ad hoc* broadening parameter similar to γ_{inter} . For semiconductors and insulators at $T = 0$, the Drude frequency vanishes, i.e., $\omega_{\text{D},\alpha\beta} = 0$, leaving only the interband contribution. Therefore, datasets of dielectric functions for semiconductors and insulators only contain the interband contribution $\varepsilon_{\text{inter},\alpha\beta}(\omega)$.

B. *Ab initio* calculations

The density functional theory (DFT) calculations were performed with the plane-wave code QUANTUM ESPRESSO [95,96] using optimized norm-conserving Vanderbilt pseudopotentials from the SG15 library [97] and PBE [98–100] as the exchange-correlation functional. The interband dielectric tensor and the Drude frequency tensor in the IPA were calculated using the SIMPLE code [101], which is part of the QUANTUM ESPRESSO distribution. We selected the SIMPLE code because it implements the optimal basis (OB) method introduced by Shirley [102] for the periodic part of the Bloch wavefunctions, thus enabling their interpolation to arbitrary k points [103]. Such interpolation is particularly valuable for optical calculations, as these usually require dense sampling of the Brillouin zone to converge. This is especially true for metals, as their optical response depends sensitively on the precise electronic structure at the Fermi surface, requiring extremely fine k -point grids (see Supplemental Note 10 of the Supplemental Material [40] for an example). Thus, the interpolation scheme implemented in SIMPLE allows for the efficient evaluation of optical properties on dense k -point grids. However, even with this interpolation scheme, calculating and converging the dielectric functions for the 201 361 intermetallic compounds required 2.3 million CPU hours using the workflow described below.

The crystal structures for the calculations were taken from the ALEXANDRIA database [38,39] (downloaded March 2024) of theoretically stable crystals and filtered by the following criteria: compounds containing only metallic elements from the first six rows of the periodic table (excluding lanthanides), ten or fewer atoms in the unit cell, a space group number greater than or equal to 75, a distance from the convex hull of less than 250 meV, and a nonmagnetic ground state. Following these criteria, we kept up to the five most stable crystal structures for each unique composition. After applying these filters,

205 224 intermetallic compounds remained, which were then reduced to their primitive standard structure using SPGLIB [104].

For every material, we performed the following three-step high-throughput workflow to calculate the diagonal components of the interband dielectric tensor, $\varepsilon_{\text{inter},\alpha\alpha}(\omega)$, and the diagonal components of the Drude frequency tensor, $\omega_{\text{D},\alpha\alpha}$. All k -point grids ($k_x \times k_y \times k_z$) were generated from a structure-independent reciprocal density $n_{\mathbf{k}}$ per atom, as defined in PYMATGEN [105], using Monkhorst-Pack grids (i.e., the automatic k -point grid mode of QUANTUM ESPRESSO). The number of subdivisions along each reciprocal-space direction was chosen to be even, such that any odd subdivision was increased to the next (even) integer.

First, we converged the DFT plane-wave cutoff energy and the k -point grid, employing the same strategy as in our previous G_0W_0 benchmark workflow [106]. Here, we adopted a looser convergence threshold of 1 kcal/mol (0.04 eV per atom) for the total energy, rather than the 1 meV per atom used in the benchmark setting, in order to reduce computational cost. For all calculations, we used a “cold” smearing [107] of 300 meV to improve convergence for the metallic systems considered here. The resulting converged plane-wave cutoff energy and converged k -point grid were used in all subsequent steps, unless otherwise noted.

Second, following the converged ground state calculation, we performed a non-self-consistent DFT calculation on a uniform $2 \times 2 \times 2$ k -point grid without symmetry reduction, as required by SIMPLE for constructing the OB [34,101]. Here, the number of bands included in the KS Hamiltonian was chosen such that all transitions up to 20 eV above the Fermi energy were included. Based on this calculation, we constructed the OB with the threshold for the Gram-Schmidt orthonormalization algorithm set to 10^{-2} a.u. (input variable `s_bands` in SIMPLE [101]).

Before describing the final step of the workflow, we briefly outline the settings used to compute optical properties in the IPA in SIMPLE. To ensure consistent Fermi energies in SIMPLE, we again used a smearing of 300 meV (`simpleip_in%fermi_ngauss = -1`, `simpleip_in%fermi_degauss = 0.022 05 Ry`). When evaluating the derivative of the Fermi-Dirac distribution in Eq. (4), SIMPLE employs a Gaussian smearing to improve convergence of the Drude frequency with respect to the k -point grid [101], which we set to 100 meV (`simpleip_in%drude_degauss = 0.007 35 Ry`). Dielectric functions were evaluated on a uniform frequency grid of 2001 points between 0 and 20 eV with $\gamma_{\text{inter}} = 300$ meV (`simpleip_in%inter_broadening = 0.022 05 Ry`) and $\gamma_{\text{D}} = 100$ meV (`simpleip_in%intra_broadening = 0.007 35 Ry`) [cf. Eqs. (2) and (3)].

In the third and final step, we converged $\varepsilon_{\text{inter},\alpha\alpha}(\omega)$ and $\omega_{\text{D},\alpha\alpha}$ with respect to the k -point grid. We initialized the convergence loop with a k -point grid corresponding to $8 \times n_{\mathbf{k}}$, effectively doubling the converged DFT k -point grid. Since individual calculations are efficient due to interpolation, this dense starting point reduces the number of iterations required to achieve convergence. Here, we increased $n_{\mathbf{k}}$ such that the number of subdivisions in at least two directions of the next grid exceeded those of the previous grid, i.e., $k_i^{n+1} > k_i^n$ for

at least two directions $i \in \{x, y, z\}$. For each k -point grid, we computed $\varepsilon_{\text{inter},\alpha\alpha}(\omega)$ and $\omega_{\text{D},\alpha\alpha}$. The procedure was repeated until (1) the change in $\omega_{\text{D},xx}$ between successive grids was below 100 meV, and (2) the similarity coefficient (SC) [27]:

$$\text{SC}[\tilde{\varepsilon}(\cdot); \varepsilon(\cdot)] = 1 - \frac{\int |\tilde{\varepsilon}(\omega) - \varepsilon(\omega)| d\omega}{\int |\varepsilon(\omega)| d\omega} \quad (5)$$

between the rotationally invariant average of consecutive total dielectric tensors, i.e., $\bar{\varepsilon}(\omega) = \text{Tr}[\varepsilon_{\alpha\beta}(\omega)]/3$, evaluated from 1 to 20 eV, exceeded 0.98. In Eq. (5), $\tilde{\varepsilon}(\omega)$ refers to the dielectric function computed on the current k -point grid, and $\varepsilon(\omega)$ refers to that computed on the preceding, less dense k -point grid. The k -point grid convergence loop terminated once both criteria were satisfied or after 20 iterations. In the latter case, the material was discarded from the dataset.

For each material, we stored the crystal structure, along with all information relating to the convergence of the total energy and optical properties, the converged $\varepsilon_{\text{inter},\alpha\alpha}(\omega)$, and $\omega_{\text{D},\alpha\alpha}$ in the form of a `ComputedStructureEntry` from `PY-MATGEN` [105] (essentially a JSON file).

C. Machine learning

All ML models were implemented using `PYTORCH` and `PYTORCH GEOMETRIC`. For training and evaluating the models, the dataset was split by unique chemical composition [27,108] into fixed training, validation, and test sets in an 80:10:10 ratio, each containing 160 728, 20 179, and 20 454 materials, respectively. Supplemental Note 11 of the Supplemental Material [40] provides statistics on the distribution of elements and the distribution of the number of atoms per unit cell across the training, validation, and test sets. As stated at the beginning of Sec. II, for each material we used rotationally invariant averages of the interband dielectric tensor, $\bar{\varepsilon}_{\text{inter}}(\omega) = \text{Tr}[\varepsilon_{\text{inter},\alpha\beta}(\omega)]/3$, and Drude frequency tensor, $\bar{\omega}_{\text{D}}^2 = \text{Tr}[\omega_{\text{D},\alpha\beta}^2]/3$, as targets for all GNNs. For each material, the primary graph was created by converting crystal structures into multigraphs [79]. Each node in the primary graph corresponds to an atom in the unit cell, and edges are created between nodes if the distance between the corresponding atoms is less than or equal to $r_c^{2\text{B}} = 5.5 \text{ \AA}$, taking into account periodic boundary conditions. The line graph, required for OPTIMETAL3B, was then obtained by connecting edges whose lengths are less than or equal to $r_c^{3\text{B}} = 4.0 \text{ \AA}$, in line with M3GNET [109] and MATTERSIM [110].

For the equivariant models, we adapted the UMA architecture [20], which builds on the ESEN backbone [36], to the present optical-property prediction task. Because charge-state, spin-state, and DFT-task labels are not included in the input-target mapping for our task, we removed the corresponding input embeddings, as well as the mixture-of-linear-experts routing used in the original UMA model [20]. The UMA backbone, configured with two message-passing layers, was then used to construct atom-wise equivariant latent features, from which we retained the $\ell = 0$ channel for graph-level predictions. These scalar node features were pooled using mean pooling into a global latent representation and passed to two output multilayer perceptrons (MLPs), analogous to those used in the OPTIMETAL models described in Supplemental Note 3 of the Supplemental Material [40]. While the OPTI-

METAL output MLPs used batch normalization and rectified linear unit activations, consistent with the other MLPs in the OPTIMETAL architectures, the output MLPs in the adapted UMA model followed the MLP design of the original UMA implementation, using layer normalization and sigmoid linear unit activations in the hidden layers. Apart from these modifications, the UMA backbone was left unchanged.

The loss function used for training the multiobjective GNNs is defined as the sum of the MAEs

$$\text{MAE}[Y; \tilde{Y}] = \frac{1}{D} \sum_{i=1}^D |Y_i - \tilde{Y}_i| \quad (6)$$

for $\bar{\varepsilon}_{\text{inter}}(\omega)$ and $\bar{\omega}_{\text{D}}$:

$$L = \text{MAE}[\bar{\varepsilon}_{\text{inter}}^{\text{ML}}(\omega); \bar{\varepsilon}_{\text{inter}}^{\text{DFT}}(\omega)] + \text{MAE}[\bar{\omega}_{\text{D}}^{\text{ML}}; \bar{\omega}_{\text{D}}^{\text{DFT}}] \quad (7)$$

following our previous finding that MAE-based training performs better than MSE-based training for optical spectra [27].

All models were trained for 500 epochs with `ADAMW` [111], employing a cosine annealing learning-rate schedule. For the OPTIMETAL models, each training run began with a 5-epoch warm-up phase, during which the learning rate increased linearly from zero to the initial maximum learning rate, η_{max} , of the cosine annealing schedule. Note that the warm-up epochs were included in the total training epoch count. After the warm-up phase, the learning rate gradually decayed to zero over the remaining epochs using a half-cosine cycle. For the adapted UMA model, no warm-up phase was used, and instead, the learning rate followed a half-cosine decay from η_{max} to zero starting at the first epoch. We applied gradient clipping with a gradient norm threshold of 100 and used a fixed batch size of 256 for all training runs. Each model was trained with three different random initializations to ensure robust performance metrics. All training and inference runs for the OPTIMETAL models were performed on NVIDIA A100 GPUs in mixed-precision mode using `bfloat16` to reduce memory usage and accelerate computation. The adapted UMA model was trained using `float32`, following the numerical-precision considerations reported in Ref. [20].

For the OPTIMETAL models, all architectural choices and optimizer hyperparameters were fixed to the values identified during the architecture optimization described in Supplemental Note 4 of the Supplemental Material [40], which was performed at $d_h = 256$ using the 20 000-sample training subset. However, we found that training became unstable for the TC-based OPTIMETAL models at $d_h > 256$ when using η_{max} optimized at $d_h = 256$. To address this issue, we scaled η_{max} as a function of N using a simple power-law relation, thereby avoiding width-specific learning-rate reoptimization. Based on additional experiments reported in Supplemental Note 12 of the Supplemental Material [40], we found that a fitted learning-rate scaling exponent equal to one yielded the best overall stability and performance. For the adapted UMA model, optimizer hyperparameters were optimized using the first stage of the workflow described in Supplemental Note 4 of the Supplemental Material [40] at $d_h = 256$ and with the 20 000-sample training subset.

A total of 9707 GPU hours were used to optimize the architecture of the OPTIMETAL models and train all models for the NSL analysis.

D. Functional forms of neural scaling laws

Guided by previous studies on NSLs [6,7,9,13,15,22], we evaluated four candidate functional forms for the 1D NSL analysis:

$$L(X) = \left(\frac{X_0}{X}\right)^\alpha, \quad (8)$$

$$L(X) = L_\infty + \left(\frac{X_0}{X}\right)^\alpha, \quad (9)$$

$$L(X) = \left(\frac{X_c}{X}\right)^{\alpha_2} \left[1 + \left(\frac{X_c}{X}\right)^{\alpha_1 - \alpha_2}\right], \quad (10)$$

$$L(X) = A \left(\frac{X_c}{X}\right)^{\alpha_2} \left[1 + \left(\frac{X_c}{X}\right)^{\alpha_1 - \alpha_2}\right]. \quad (11)$$

Here, $X \in \{D, N\}$ denotes either the dataset size, D , or the number of trainable parameters, N . In the simple NSLs in Eqs. (8) and (9), X_0 sets the point where $L(X_0) = 1$ and $L(X_0) = L_\infty + 1$, respectively, while in the BNSLs in Eqs. (10) and (11), X_c marks the crossover point where the scaling exponent changes. The parameters α , α_1 , and α_2 control the decay of the loss with increasing X , while L_∞ represents a possible irreducible loss and A an optional multiplicative amplitude. In learning settings based on *ab initio* data, we generally expect that, in the limit of sufficiently large D and N , $L_\infty = 0$, since no residual entropy or irreducible uncertainty is expected. In practice, however, finite numerical accuracy can produce an effective irreducible loss. Each of the four functions was fitted, using nonlinear least-squares regression, to the validation loss averaged over three random model initializations.

To compare the descriptive quality of these functions, we used the AIC, which balances how well a function fits the data with the number of parameters used for the fit [81]. For least-squares fits, the AIC is given by

$$\text{AIC} = 2k + n \ln \left(\frac{1}{n} \sum_{i=1}^n [L(X_i) - L_{\text{val}}(X_i)]^2 \right), \quad (12)$$

where k is the number of fit parameters and n is the number of data points used for the fit. Since the 1D NSL fits are based on only a small number of data points, we applied the finite-sample correction to obtain the AICc

$$\text{AICc} = \text{AIC} + \frac{2k^2 + 2k}{n - k - 1}, \quad (13)$$

which prevents overly flexible models from being favored when n is small relative to k [81].

After determining the 1D NSL forms that best describe the data scaling, i.e., Eq. (10), and observing that the parameter scaling follows the expected saturating power law, i.e., Eq. (9), we constructed 2D NSL maps $L(D, N)$ by combining their respective dependencies. Inspired by the work of

Kaplan *et al.* [7], we tested the smoothly interpolating form

$$L(D, N) = \left[\left(\frac{N_0}{N} \right)^{\frac{\alpha_N}{\alpha_{D,2}}} + \left(\frac{D_c}{D} \right) \left(1 + \frac{D_c}{D} \right)^{\left(\frac{\alpha_{D,1}}{\alpha_{D,2}} - 1 \right)} \right]^{\alpha_{D,2}}. \quad (14)$$

In contrast to the 1D parameter-scaling fits, the Kaplan-type 2D formulation does not include an irreducible loss term L_∞ , as it already encodes saturation with respect to N at small D . As an alternative formulation, following Hoffmann *et al.* [9], we also tested the additive description

$$L(D, N) = \left(\frac{N_0}{N} \right)^{\alpha_N} + \left(\frac{D_c}{D} \right)^{\alpha_{D,2}} \left(1 + \frac{D_c}{D} \right)^{\alpha_{D,1} - \alpha_{D,2}}. \quad (15)$$

For the additive Hoffmann-type formulation, we also tested an additional irreducible loss term L_∞ , but the nonlinear least-squares regression converged to $L_\infty = 0$ for the fitted 2D NSL maps.

ACKNOWLEDGMENTS

We thank the staff of the Compute Center of the Technische Universität Ilmenau, especially Mr. Henning Schwanbeck, for providing an excellent research environment. M. Großmann thanks C. Dreßler, J. Hänseroth, and A. Flötotto for their valuable feedback on the first draft of the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft DFG (Project No. 537033066) and the Carl-Zeiss-Stiftung (Funding Code P2023-02-008).

M.G. and M.G. conceived the idea; M. Großmann wrote the high-throughput workflow, performed all *ab initio* calculations, analyzed the data, trained and optimized the ML models, performed the scaling law analysis, visualized all results, and wrote the first draft of the manuscript. E.R. supervised the work and contributed to the interpretation. All authors revised and approved the manuscript.

The authors declare no competing interests.

DATA AVAILABILITY

The dataset containing the crystal structures, interband dielectric functions, and Drude frequencies of 201 361 intermetallic compounds is available in Refs. [112,113]. TENSORBOARD log files of all training runs, along with the training, validation, and test splits of the dataset in compressed HDF5 format, and the derived graph representations, are available in Ref. [114]. The weights of all models evaluated in Supplemental Note 5 of the Supplemental Material [40] are publicly available: Those of the invariant OPTIMETAL2B and OPTIMETAL3B models are available in Ref. [115], whereas those of the adapted equivariant UMA models are available in Ref. [114] because of their larger file sizes. The third-party electronic structure codes QUANTUM ESPRESSO and SIMPLE are available in Ref. [116]. The high-throughput workflow used in this study, including the filtered input structures derived from the ALEXANDRIA database, is available in Ref. [117]. All code and scripts used to implement, train, evaluate, and optimize the architecture of the machine learning models, as well as those used to obtain the neural scaling laws, are available in Ref. [115].

- [1] S. Amari, N. Fujita, and S. Shinomoto, Four types of learning curves, *Neural Comput.* **4**, 605 (1992).
- [2] S. Amari, A universal theorem on learning curves, *Neural Netw.* **6**, 161 (1993).
- [3] S. Amari and N. Murata, Statistical theory of learning curves under entropic loss criterion, *Neural Comput.* **5**, 140 (1993).
- [4] K.-R. Müller, M. Finke, N. Murata, K. Schulten, and S. Amari, A numerical study on learning curves in stochastic multilayer feedforward networks, *Neural Comput.* **8**, 1085 (1996).
- [5] V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 2000).
- [6] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, Deep learning scaling is predictable, empirically, [arXiv:1712.00409](https://arxiv.org/abs/1712.00409).
- [7] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, Scaling laws for neural language models, [arXiv:2001.08361](https://arxiv.org/abs/2001.08361).
- [8] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, Scaling vision transformers, in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, Los Alamitos, CA, USA, 2022), pp. 1204–1213.
- [9] J. Hoffmann *et al.*, An empirical analysis of compute-optimal large language model training, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., Red Hook, NY, USA, 2022), Vol. 35, pp. 30016–30030.
- [10] A. Grattafiori *et al.*, The Llama 3 herd of models, [arXiv:2407.21783](https://arxiv.org/abs/2407.21783).
- [11] B. Bordelon, A. Atanasov, and C. Pehlevan, A dynamical model of neural scaling laws, in *Proceedings of the 41st International Conference on Machine Learning*, edited by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Proceedings of Machine Learning Research Vol. 235 (PMLR, Vienna, Austria, 2024), pp. 4345–4382.
- [12] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma, Explaining neural scaling laws, *Proc. Natl. Acad. Sci. USA* **121**, e2311878121 (2024).
- [13] E. Caballero, K. Gupta, I. Rish, and D. Krueger, Broken neural scaling laws, in *International Conference on Learning Representations, Kigali, Rwanda* (ICLR, 2023).
- [14] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Phys. Rev. Lett.* **108**, 058301 (2012).
- [15] N. C. Frey, R. Soklaski, S. Axelrod, S. Samsi, R. Gómez-Bombarelli, C. W. Coley, and V. Gadepally, Neural scaling of deep chemical models, *Nat. Mach. Intell.* **5**, 1297 (2023).
- [16] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, in *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., Red Hook, NY, USA, 2022), Vol. 35, pp. 11423–11436.
- [17] X. Ji, Z. Wang, Z. Gao, H. Zheng, L. Zhang, G. Ke, and W. E, Exploring molecular pretraining model at scale, in *Advances in Neural Information Processing Systems*, edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Curran Associates, Inc., Red Hook, NY, USA, 2024), Vol. 37, pp. 46956–46978.
- [18] E. Qu and A. S. Krishnapriyan, The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains, in *Advances in Neural Information Processing Systems*, edited by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Curran Associates, Inc., Red Hook, NY, USA, 2024), Vol. 37, pp. 139030–139053.
- [19] C. Li, Z. Ye, M. L. Pasini, J. Y. Choi, C. Wan, and P. Balaprakash, Scaling laws of graph neural networks for atomistic materials modeling, in *Proceedings of the 62nd Annual ACM/IEEE Design Automation Conference, DAC'25* (IEEE Press, New York, NY, USA, 2025).
- [20] B. Wood, M. Dzamba, X. Fu, M. Gao, M. Shuaibi, L. Barroso-Luque, K. Abdelmaqsoud, V. Gharakhanyan, J. Kitchin, D. Levine, K. Michel, A. Sriram, T. Cohen, A. Das, S. Sahoo, A. Rizvi, Z. Ulissi, and L. Zitnick, UMA: A family of universal models for atoms, in *Advances in Neural Information Processing Systems*, edited by D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (Curran Associates, Inc., Red Hook, NY, USA, 2025), Vol. 38, pp. 129391–129427.
- [21] A. Trikha, K. Chu, A. Gosai, P. Szachta, and E. Weiner, Scaling laws for neural material models, [arXiv:2509.21811](https://arxiv.org/abs/2509.21811).
- [22] K. Ngo and S. Ravanbakhsh, Scaling laws and symmetry, evidence from neural force fields, in *The Fourteenth International Conference on Learning Representations, Rio de Janeiro, Brazil* (ICLR, 2026).
- [23] A. Mazitov, F. Bigi, M. Kellner, P. Pegolo, D. Tisi, G. Fraux, S. Pozdnyakov, P. Loche, and M. Ceriotti, PET-MAD as a lightweight universal interatomic potential for advanced materials modeling, *Nat. Commun.* **16**, 10653 (2025).
- [24] H. Yang *et al.*, MatterSim-MT: A multi-task foundation model for in silico materials characterization, [arXiv:2605.07927](https://arxiv.org/abs/2605.07927).
- [25] M. Walker and K. T. Butler, The carbon cost of materials discovery: Can machine learning really accelerate the discovery of new photovoltaics? *Mater. Horiz.* **13**, 1993 (2026).
- [26] M. Dressel and G. Grüner, *Electrodynamics of Solids: Optical Properties of Electrons in Matter* (Cambridge University Press, Cambridge, England, 2002).
- [27] M. Grunert, M. Großmann, and E. Runge, Deep learning of spectra: Predicting the dielectric function of semiconductors, *Phys. Rev. Mater.* **8**, L122201 (2024).
- [28] M. Grunert, M. Großmann, and E. Runge, Machine learning climbs the Jacob's ladder of optoelectronic properties, *Nat. Commun.* **16**, 8142 (2025).
- [29] M. Grunert, M. Großmann, and E. Runge, Optical spectra prediction using three-body information, *Appl. Phys. Lett.* **128**, 053302 (2026).
- [30] A. Ibrahim and C. Ataca, Prediction of frequency-dependent optical spectrum for solid materials: A multioutput and multifidelity machine learning approach, *ACS Appl. Mater. Interfaces* **16**, 41145 (2024).
- [31] N. T. Hung, R. Okabe, A. Chottrattanapituk, and M. Li, Universal ensemble-embedding graph neural network for direct prediction of optical spectra from crystal structures, *Adv. Mater.* **36**, 2409175 (2024).

- [32] T.-W. Hsu, Z. Fang, A. Bansil, and Q. Yan, Accurate prediction of tensorial spectra using equivariant graph neural network, *Nat. Commun.* **17**, 3330 (2026).
- [33] A. Marini, G. Onida, and R. Del Sole, Plane-wave DFT-LDA calculation of the electronic structure and absorption spectrum of copper, *Phys. Rev. B* **64**, 195125 (2001).
- [34] G. Prandini, G.-M. Rignanese, and N. Marzari, Photorealistic modelling of metals from first principles, *npj Comput. Mater.* **5**, 129 (2019).
- [35] B. A. Barker, J. Deslippe, J. Lischner, M. Jain, O. V. Yazyev, D. A. Strubbe, and S. G. Louie, Spinor GW/Bethe-Salpeter calculations in BerkeleyGW: Implementation, symmetries, benchmarking, and performance, *Phys. Rev. B* **106**, 115127 (2022).
- [36] X. Fu, B. M. Wood, L. Barroso-Luque, D. S. Levine, M. Gao, M. Dzamba, and C. L. Zitnick, Learning smooth and expressive interatomic potentials for physical property prediction, in *Proceedings of the 42nd International Conference on Machine Learning*, edited by A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Proceedings of Machine Learning Research Vol. 267 (PMLR, Vancouver, Canada, 2025), pp. 17875–17893.
- [37] P. J. M. A. Carriço, M. Ferreira, T. F. T. Cerqueira, F. Nogueira, and P. Borlido, High-refractive-index materials screening from machine learning and *ab initio* methods, *Phys. Rev. Mater.* **8**, 015201 (2024).
- [38] J. Schmidt, N. Hoffmann, H. Wang, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti, and M. A. L. Marques, Machine-learning-assisted determination of the global zero-temperature phase diagram of materials, *Adv. Mater.* **35**, 2210788 (2023).
- [39] J. Schmidt, T. F. Cerqueira, A. H. Romero, A. Loew, F. Jäger, H.-C. Wang, S. Botti, and M. A. Marques, Improving machine-learning models in materials science through large datasets, *Mater. Today Phys.* **48**, 101560 (2024).
- [40] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/wv8r-7zr1> for additional details on the first-principles calculations and dataset, model architectures, model architecture optimization, detailed model performance metrics, and neural- and learning-rate scaling analyses, which includes Refs. [41–78].
- [41] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.* **1**, 011002 (2013).
- [42] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, The materials application programming interface (API): A simple, flexible and efficient API for materials data based on representational state transfer (REST) principles, *Comput. Mater. Sci.* **97**, 209 (2015).
- [43] S. Babar and J. H. Weaver, Optical constants of Cu, Ag, and Au revisited, *Appl. Opt.* **54**, 477 (2015).
- [44] F. Cheng, P.-H. Su, J. Choi, S. Gwo, X. Li, and C.-K. Shih, Epitaxial growth of atomically smooth aluminum on silicon and its intrinsic optical properties, *ACS Nano* **10**, 9852 (2016).
- [45] *Handbook of Optical Constants of Solids*, edited by E. D. Palik (Academic Press, San Diego, CA, USA, 1985), Vol. 1.
- [46] *Handbook of Optical Constants of Solids*, edited by E. D. Palik (Academic Press, San Diego, CA, USA, 1991), Vol. 2.
- [47] *Handbook of Optical Constants of Solids*, edited by E. D. Palik (Academic Press, San Diego, CA, USA, 1998), Vol. 3.
- [48] A. G. Mathewson and H. P. Myers, Absolute values of the optical constants of some pure metals, *Phys. Scr.* **4**, 291 (1971).
- [49] N. V. Smith, Optical constants of rubidium and cesium from 0.5 to 4.0 eV, *Phys. Rev. B* **2**, 2840 (1970).
- [50] P. Schmitt, N. Felde, T. Döhring, M. Stollenwerk, I. Uschmann, K. Hanemann, M. Siegler, G. Klemm, N. Gratzke, A. Tünnermann, S. Schwinde, S. Schröder, and A. Szeghalmi, Optical, structural, and functional properties of highly reflective and stable iridium mirror coatings for infrared applications, *Opt. Mater. Express* **12**, 545 (2022).
- [51] N. V. Smith, Optical constants of sodium and potassium from 0.5 to 4.0 eV by split-beam ellipsometry, *Phys. Rev.* **183**, 634 (1969).
- [52] T. Inagaki, L. C. Emerson, E. T. Arakawa, and M. W. Williams, Optical properties of solid Na and Li between 0.6 and 3.8 eV, *Phys. Rev. B* **13**, 2305 (1976).
- [53] K. J. Palm, J. B. Murray, T. C. Narayan, and J. N. Munday, Dynamic optical properties of metal hydrides, *ACS Photon.* **5**, 4677 (2018).
- [54] J. H. Weaver, D. W. Lynch, and C. G. Olson, Optical properties of niobium from 0.1 to 36.4 eV, *Phys. Rev. B* **7**, 4311 (1973).
- [55] P. B. Johnson and R. W. Christy, Optical constants of transition metals: Ti, V, Cr, Mn, Fe, Co, Ni, and Pd, *Phys. Rev. B* **9**, 5056 (1974).
- [56] J. H. Weaver, C. G. Olson, and D. W. Lynch, Optical investigation of the electronic structure of bulk Rh and Ir, *Phys. Rev. B* **15**, 4115 (1977).
- [57] J. H. Weaver, C. G. Olson, and D. W. Lynch, Optical properties of crystalline tungsten, *Phys. Rev. B* **12**, 1293 (1975).
- [58] M. R. Querry, Optical constants of minerals and other materials from the millimeter to the ultraviolet, Technical Report 88009, Chemical Research, Development and Engineering Center, U.S. Army Armament Munitions Chemical Command, 1987, DTIC Identifier ADA192210.
- [59] P. Fulde, *Electron Correlations in Molecules and Solids* (Springer, Berlin, 1995).
- [60] E. Runge and G. Zwicknagl, Electronic structure calculations and strong correlations: A model study, *Ann. Phys.* **508**, 333 (1996).
- [61] E. Pavarini, E. Koch, F. Anders, and M. Jarrell, *Correlated Electrons: From Models to Materials* (Verlag des Forschungszentrum Jülich, Jülich, Germany, 2012).
- [62] T. Kotani, M. van Schilfgaarde, and S. V. Faleev, Quasiparticle self-consistent GW method: A basis for the independent-particle approximation, *Phys. Rev. B* **76**, 165106 (2007).
- [63] D. Pashov, S. Acharya, W. R. Lambrecht, J. Jackson, K. D. Belashchenko, A. Chantis, F. Jamet, and M. van Schilfgaarde, Questaal: A package of electronic structure methods based on the linear muffin-tin orbital technique, *Comput. Phys. Commun.* **249**, 107065 (2020).
- [64] M. Großmann, M. Grunert, and E. Runge, A robust, simple, and efficient convergence workflow for GW calculations, *npj Comput. Mater.* **10**, 135 (2024).
- [65] P. E. Blöchl, O. Jepsen, and O. K. Andersen, Improved tetrahedron method for Brillouin-zone integrations, *Phys. Rev. B* **49**, 16223 (1994).

- [66] K. Choudhary and B. DeCost, Atomistic line graph neural network for improved materials property predictions, *npj Comput. Mater.* **7**, 185 (2021).
- [67] R. Ruff, P. Reiser, J. Stühmer, and P. Friederich, Connectivity optimized nested line graph networks for crystal structures, *Digital Discovery* **3**, 594 (2024).
- [68] J. Gasteiger, J. Groß, and S. Günnemann, Directional message passing for molecular graphs, in *International Conference on Learning Representations, Virtual Conference (ICLR, 2020)*.
- [69] S. Brody, U. Alon, and E. Yahav, How attentive are graph attention networks?, in *International Conference on Learning Representations, Virtual Conference (ICLR, 2022)*.
- [70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, USA, 2017), Vol. 30.
- [71] Y. Li, C. Gu, T. Dullien, O. Vinyals, and P. Kohli, Graph matching networks for learning the similarity of graph structured objects, in *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov, Proceedings of Machine Learning Research Vol. 97 (PMLR, Long Beach, CA, USA, 2019), pp. 3835–3845.
- [72] O. Vinyals, S. Bengio, and M. Kudlur, Order matters: Sequence to sequence for sets, in *International Conference on Learning Representations, San Juan, Puerto Rico (ICLR, 2016)*.
- [73] M. Grunert, M. Großmann, and E. Runge, Discovery of sustainable energy materials via the machine-learned material space, *Small* **22**, 2412519 (2026).
- [74] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (ACM Press, New York, NY, USA, 2019)*, pp. 2623–2631.
- [75] S. Watanabe, Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance, *arXiv:2304.11127*.
- [76] Commission Internationale de l’Eclairage, CIE standard illuminant D65, Technical Report, CIE, 2019.
- [77] Commission Internationale de l’Eclairage, Colour-matching functions of CIE 1931 standard colorimetric observer, Technical Report, CIE, 2019.
- [78] G. Sharma and H. Trussell, Digital color imaging, *IEEE Trans. Image Process.* **6**, 901 (1997).
- [79] T. Xie and J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, *Phys. Rev. Lett.* **120**, 145301 (2018).
- [80] K. K. Thekumparampil, C. Wang, S. Oh, and L.-J. Li, Attention-based graph neural network for semi-supervised learning, *arXiv:1803.03735*.
- [81] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference* (Springer, New York, NY, USA, 2002).
- [82] J. T. Frank, O. T. Unke, K.-R. Müller, and S. Chmiela, A Euclidean transformer for fast and stable machine learned force fields, *Nat. Commun.* **15**, 6539 (2024).
- [83] J. Riebesell, R. E. A. Goodall, P. Benner, Y. Chiang, B. Deng, G. Ceder, M. Asta, A. A. Lee, A. Jain, and K. A. Persson, A framework to evaluate machine learning crystal stability predictions, *Nat. Mach. Intell.* **7**, 836 (2025).
- [84] A. Bochkarev, Y. Lysogorskiy, and R. Drautz, Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing, *Phys. Rev. X* **14**, 021036 (2024).
- [85] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nat. Commun.* **13**, 2453 (2022).
- [86] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, The design space of E(3)-equivariant atom-centred interatomic potentials, *Nat. Mach. Intell.* **7**, 56 (2025).
- [87] J. T. Frank, S. Chmiela, K.-R. Müller, and O. T. Unke, Machine learning global atomic representations with Euclidean fast attention, *Nat. Mach. Intell.* **8**, 388 (2026).
- [88] E. Qu, B. M. Wood, A. S. Krishnapriyan, and Z. W. Ulissi, A recipe for scalable attention-based MLIPs: Unlocking long-range accuracy with all-to-all node attention, *arXiv:2603.06567*.
- [89] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Big data meets quantum chemistry approximations: The Δ -machine learning approach, *J. Chem. Theory Comput.* **11**, 2087 (2015).
- [90] C. Qu, A. Nandi, P. L. Houston, Q. Yu, R. Conte, and J. M. Bowman, “Gold-standard” Δ -machine learned transferable potential for linear alkanes, *J. Phys. Chem. Lett.* **16**, 12393 (2025).
- [91] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, *Nat. Commun.* **10**, 2903 (2019).
- [92] N. Hoffmann, J. Schmidt, S. Botti, and M. A. L. Marques, Transfer learning on large datasets for the accurate prediction of material properties, *Digital Discovery* **2**, 1368 (2023).
- [93] D. Baum, A. Förster, and L. Visscher, Transfer learning of GW Bethe–Salpeter equation excitation energies, *Chem. Sci.* **17**, 8090 (2026).
- [94] M. L. Pasini, J. Y. Choi, K. Mehta, R. Messerly, R. Weaver, L. Ungerboeck, I. Lyngaas, B. Stump, A. M. Aji, K. W. Schulz, and J. Polo, Exascale multi-task graph foundation models for imbalanced, multi-fidelity atomistic data, *arXiv:2604.15380*.
- [95] P. Giannozzi *et al.*, QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials, *J. Phys.: Condens. Matter* **21**, 395502 (2009).
- [96] P. Giannozzi *et al.*, Advanced capabilities for materials modelling with QUANTUM ESPRESSO, *J. Phys.: Condens. Matter* **29**, 465901 (2017).
- [97] D. R. Hamann, Optimized norm-conserving Vanderbilt pseudopotentials, *Phys. Rev. B* **88**, 085117 (2013).
- [98] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [99] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple [Phys. Rev. Lett. **77**, 3865 (1996)], *Phys. Rev. Lett.* **78**, 1396(E) (1997).

- [100] M. A. Marques, M. J. Oliveira, and T. Burnus, Libxc: A library of exchange and correlation functionals for density functional theory, *Comput. Phys. Commun.* **183**, 2272 (2012).
- [101] G. Prandini, M. Galante, N. Marzari, and P. Umari, SIMPLE code: Optical properties with optimal basis functions, *Comput. Phys. Commun.* **240**, 106 (2019).
- [102] E. L. Shirley, Optimal basis sets for detailed Brillouin-zone integrations, *Phys. Rev. B* **54**, 16464 (1996).
- [103] D. Prendergast and S. G. Louie, Bloch-state-based interpolation: An efficient generalization of the Shirley approach to interpolating electronic structure, *Phys. Rev. B* **80**, 235126 (2009).
- [104] A. Togo, K. Shinohara, and I. Tanaka, Spglib: A software library for crystal symmetry search, *Sci. Technol. Adv. Mater. Methods* **4**, 2384822 (2024).
- [105] S. P. Ong *et al.*, Python Materials Genomics (pymatgen): A robust, open-source Python library for materials analysis, *Comput. Mater. Sci.* **68**, 314 (2013).
- [106] M. Großmann, M. Thieme, M. Grunert, and E. Runge, Many-body perturbation theory vs. density functional theory: A systematic benchmark for band gaps of solids, *npj Comput. Mater.* **12**, 25 (2026).
- [107] N. Marzari, D. Vanderbilt, A. De Vita, and M. C. Payne, Thermal contraction and disordering of the Al(110) surface, *Phys. Rev. Lett.* **82**, 3296 (1999).
- [108] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Scaling deep learning for materials discovery, *Nature (London)* **624**, 80 (2023).
- [109] C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, *Nat. Comput. Sci.* **2**, 718 (2022).
- [110] H. Yang *et al.*, MatterSim: A deep learning atomistic model across elements, temperatures and pressures, [arXiv:2405.04967](https://arxiv.org/abs/2405.04967).
- [111] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *International Conference on Learning Representations, New Orleans, LA, USA (ICLR, 2019)*.
- [112] M. Großmann, M. Grunert, and E. Runge, Dielectric functions of intermetallic compounds (Part 1). figshare. Dataset, 2026, <https://doi.org/10.6084/m9.figshare.31111798.v1>.
- [113] M. Großmann, M. Grunert, and E. Runge, Dielectric functions of intermetallic compounds (Part 2). figshare. Dataset, 2026, <https://doi.org/10.6084/m9.figshare.31112491.v1>.
- [114] M. Großmann, M. Grunert, and E. Runge, Machine learning dataset, model weights, and training logs for the publication “Broken neural scaling laws in learning the optical properties of solids”. figshare. Dataset, 2026, <https://doi.org/10.6084/m9.figshare.31112554.v2>.
- [115] <https://github.com/MaxGrossmann/optimetal>.
- [116] <https://quantum-espresso.org>.
- [117] https://github.com/MaxGrossmann/ht_metals.