ORIGINAL RESEARCH **Open access**

Neural Scaling Laws for Materials Property Databases: Theoretical Analysis of Power-Law Exponents from Data Redundancy

Diego Morales[1*], Andres Gutierrez[1], Lucia Navarro[2], Pablo Rios[1]

Abstract

Neural scaling laws describe the power-law decay of prediction error with increasing training dataset size, yet the scaling exponents reported for materials property prediction vary widely (0.1–0.8) across databases and properties. This variation has remained poorly understood, limiting reliable forecasting of data requirements in computational materials discovery. Here we present a purely theoretical framework that attributes the observed differences in scaling exponents primarily to the intrinsic data redundancy of materials databases. We formalize redundancy as the fraction of samples that convey overlapping or duplicate structural information, thereby reducing the effective number of independent samples. Through conceptual derivations and proof sketches grounded in information theory and sample-complexity arguments, we establish that the effective scaling exponent is approximately the ideal (independent-sample) exponent multiplied by (1 − r), where r is the redundancy fraction. Upper and lower bounds on achievable exponents are derived directly from three fundamental structural features: the number of distinct crystal prototypes, elemental compositional diversity, and the variety of local atomic environments. High redundancy—arising naturally from repeated prototypes, compositional biases, and clustered coordination motifs—systematically flattens the exponent and imposes hard structural limits on learning efficiency, independent of model architecture. The framework provides lightweight, model-free metrics for estimating redundancy from standard database statistics and offers actionable guidance for diversity-driven dataset curation. By reframing scaling behavior in terms of effective sample size and structural diversity, this work supplies a unified theoretical explanation for the wide range of reported exponents and a predictive foundation for designing more efficient materials databases.

Keywords Data redundancy, Sample complexity, Neural scaling laws, Power-law exponents, Materials property databases, Database structure

*Correspondence:
Diego Morales
diego.morales@gmail.com
[1] Department of Intelligent Materials Systems, Faculty of Engineering, University of Lima, Lima, Peru
[2] Department of Computational Materials Analytics, Faculty of Science and Technology, Pontifical Catholic University of Peru, Lima, Peru

Introduction

Neural scaling laws have transformed our understanding of deep learning by revealing predictable relationships between model performance and training resources [1, 2]. For materials property prediction, a central relationship is that prediction error decreases with dataset size according to a power-law pattern, where the scaling exponent determines how quickly accuracy improves as more data become available [3, 4]. Yet this exponent varies widely across materials databases, ranging from approximately 0.1 for certain complex properties to approximately 0.8 for simpler ones [5, 6]. Understanding

the origin of this variation is essential because it directly affects decisions about data collection priorities, computational budgets, and expectations for future model capabilities in artificial intelligence for materials science.

This paper delivers a purely theoretical analysis of neural scaling laws specifically tailored to materials property databases. The core thesis is that the observed scaling exponent is governed primarily by the degree of data redundancy, defined as the extent to which additional samples fail to introduce genuinely new information. Materials databases are rarely composed of fully independent entries; instead, they contain repeated structural motifs, compositional biases, and recurring local atomic environments. These redundancies shrink the effective number of independent samples, which in turn flattens the scaling exponent and slows learning progress [7, 8].

The analysis proceeds by first clarifying the conceptual foundations of neural scaling laws and their typical range of exponents in different domains. It then examines how redundancy arises naturally within materials databases and quantifies its impact through formal definitions of effective sample size. Theoretical propositions are developed to link redundancy fractions directly to reductions in the scaling exponent and to derive structural bounds on achievable exponents based on prototype diversity, elemental variety, and local environment clustering. Finally, practical metrics are presented for estimating redundancy from standard database statistics, enabling prediction of scaling behavior before models are trained.

Previous empirical studies have documented varying exponents in materials contexts but have not supplied a unifying theoretical explanation grounded in database structure [4, 9]. The present work fills this gap by offering conceptual derivations and proof sketches that rely exclusively on redundancy arguments rather than any experimental fitting or simulation results. The implications extend beyond explanation: the framework supplies clear guidance for designing databases that minimize redundancy and thereby maximize scaling efficiency. In an era when expanding materials databases requires substantial computational and financial investment, such theoretical insight is critical for ensuring that added data deliver meaningful gains in predictive power. By focusing exclusively on the intrinsic relationship between redundancy and scaling, this analysis establishes a foundation for more strategic data curation in computational materials engineering [7, 10].

Figure 1 develops a multi-level constraint cascade showing how structural limitations propagate through redundancy formation to determine the attainable neural scaling exponent.

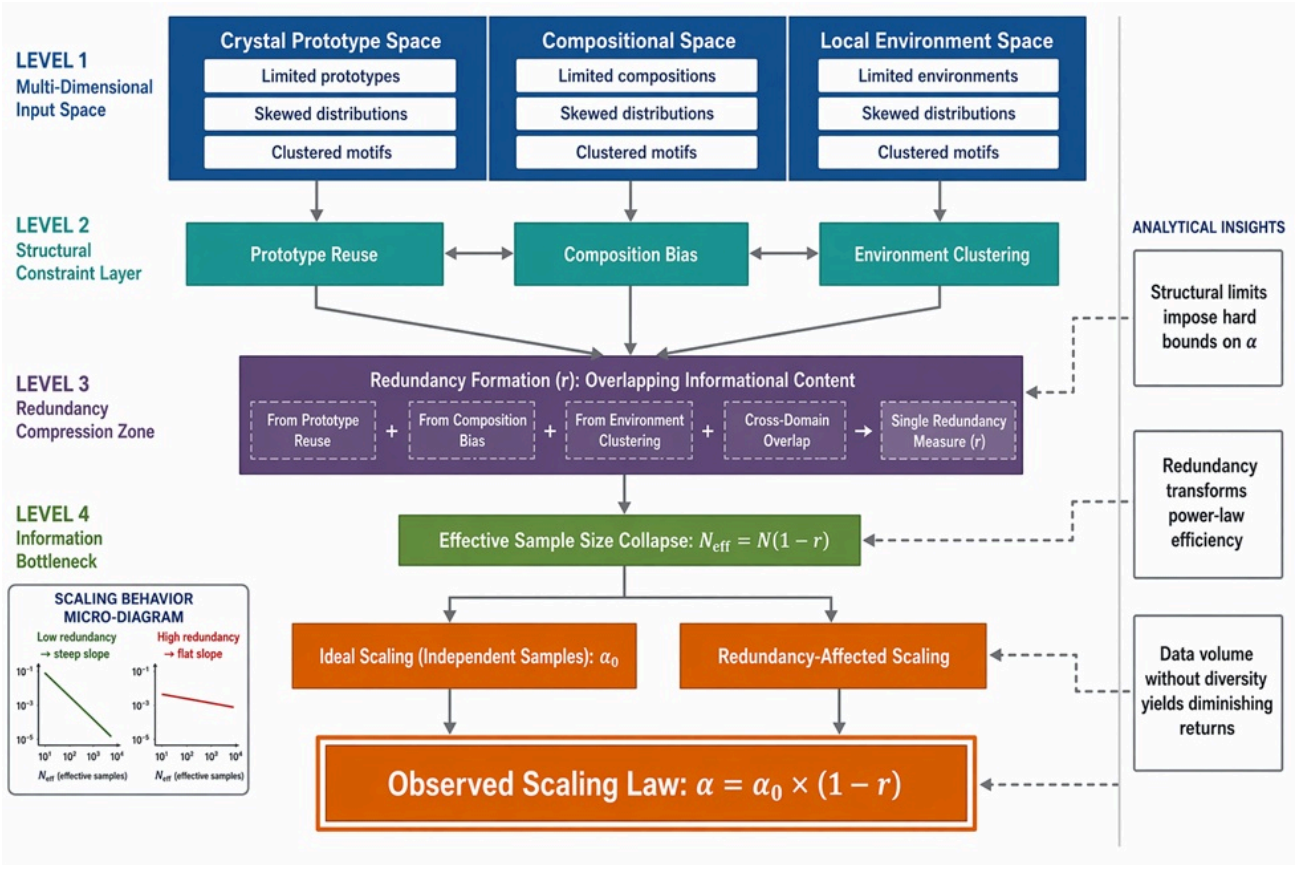

Figure 1. Multi-Level Constraint Cascade: How Structural Redundancy Governs Neural Scaling Exponents in Materials Databases

What are Neural Scaling Laws?

A neural scaling law refers to the empirical observation that the test error of a trained neural network decreases in a predictable power-law manner as the number of training samples increases. The relationship takes the general form that error declines steadily with growing dataset size, with the precise rate of decline controlled by a positive scaling exponent. This exponent serves as a compact descriptor of how responsive model performance is to additional data [1, 2].

Across different application areas, the exponent takes characteristic values. In image classification tasks, the exponent commonly lies between 0.07 and 0.35. In simpler regression settings, values typically range from 0.5 to 1.0. Within materials property prediction, the exponent spans a broader intermediate range from 0.1 to 0.8, reflecting the greater structural complexity and diversity of materials data. These ranges have been documented across multiple theoretical and empirical investigations into scaling phenomena [1, 2, 11, 12].

The scaling exponent carries clear practical meaning. A larger exponent corresponds to faster error reduction when data volume grows. For an exponent of 0.5, halving the error roughly requires quadrupling the dataset size. For an exponent of 0.2, achieving the same error reduction demands roughly thirty-two times more data. These relationships allow researchers to forecast the data requirements needed to reach a target accuracy level without exhaustive trial-and-error experimentation.

Neural scaling laws matter for several reasons in materials science. First, they supply a quantitative tool for predicting how much data must be generated or curated to meet engineering specifications for property prediction accuracy. Second, they inform dataset design decisions by highlighting when further expansion along existing lines will yield

diminishing returns. Third, they enable meaningful comparisons of the intrinsic difficulty of different materials properties; a property with a lower exponent is inherently harder to learn at scale and may require fundamentally different data-generation strategies. Fourth, scaling laws guide funding and resource allocation by clarifying the expected return on investment in database expansion.

Theoretical work has advanced understanding of why scaling laws emerge in neural networks, often tracing them to underlying statistical or information-theoretic mechanisms [1, 2]. In materials applications, scaling behavior has been examined in the context of molecular representations, chemical models, and atomistic force fields [3, 4, 9]. Some studies have also noted instances where scaling appears broken or deviates from simple power-law form, underscoring the need for a deeper theoretical account of the factors that modulate the exponent [6]. The present analysis focuses on data redundancy as the dominant modulator within materials databases, providing a conceptual lens that unifies observations across these domains. By treating scaling laws as a direct consequence of effective sample size rather than raw sample count, the framework explains both the typical range of exponents and their database-specific variation. This conceptual foundation sets the stage for the redundancy-centric derivations that follow.

Data Redundancy in Materials Databases

Data redundancy arises whenever samples within a materials property database convey overlapping or duplicate information. The effective sample size is then smaller than the nominal count. If a redundancy fraction r lies between zero and one, the effective number of independent samples equals the total number of samples multiplied by one minus r. When r equals zero, every sample contributes unique information and the database is ideal. When r equals one, all samples are essentially identical and the database provides no learning value beyond its first entry.

In real materials databases, redundancy is typically high, often falling between 0.5 and 0.9. This stems from several structural biases that are difficult to avoid during database construction. Computational workflows favor certain crystal prototypes because they are easier to relax or more stable under standard simulation conditions. Compositional sampling is frequently skewed toward common or commercially relevant element combinations, leaving large regions of chemical space unexplored. Local atomic environments also cluster tightly because many distinct crystals share similar coordination motifs or bonding patterns.

These redundancies reduce the informational content available to machine learning models. Even when the nominal dataset size grows by orders of magnitude, the number of genuinely new patterns introduced may increase only modestly. Consequently, model error decreases more slowly than expected from an independent-sampling assumption. Studies that explicitly address redundancy in large materials datasets have shown that controlling or mitigating it can improve learning efficiency without increasing raw data volume [7, 8].

The concept of effective sample size therefore provides a natural bridge between raw database statistics and observed scaling behavior. High redundancy compresses the growth of effective sample size, which in turn compresses the scaling exponent. Materials databases are particularly susceptible because their entries are generated from a finite set of physical principles operating within a constrained periodic table and crystal symmetry rules. Unlike image or text datasets, which can be augmented with nearly unlimited synthetic variation, materials data are anchored in thermodynamic and structural reality that inherently limits diversity.

Table 1 consolidates the structural sources of redundancy and clarifies their distinct theoretical effects on effective sample size and scaling exponents.

Table 1. Structural Determinants of Data Redundancy and Their Theoretical Impact on Scaling Behavior

Structural Dimension	Mechanism of Redundancy	Observable Indicator	Effect on Effective Sample Size	Impact on Scaling Exponent (α)	Theoretical Interpretation
Crystal Prototypes	Repetition of identical lattice structures	Low count of unique prototypes relative to total entries	Strong compression of independent samples	Upper bound imposed on α	Limits structural diversity space
Elemental Composition	Skewed sampling toward common chemistries	Low compositional entropy	Reduces marginal information gain per sample	Moderate reduction in α	Biases coverage of chemical space
Local Atomic Environments	Clustering of similar coordination motifs	Few dominant environment clusters	Effective sample size bounded by cluster count	Severe flattening of α	Drives logarithmic scaling behavior
Pairwise Similarity	High overlap between individual samples	High average similarity scores	Redundant incremental learning	Continuous suppression of α	Reflects micro-level redundancy
Combined Effect	Interaction across all dimensions	Composite redundancy score	Multiplicative contraction: $N_{eff} = N(1-r)$	$\alpha \approx \alpha_0(1-r)$	Unified redundancy-scaling relationship

Recognizing redundancy as a primary driver of scaling behavior shifts the focus from merely enlarging databases to deliberately diversifying them. The following sections formalize this intuition through theoretical propositions and then supply concrete metrics for quantifying redundancy from standard database statistics.

Theoretical Link between Redundancy and Scaling Exponent

The central theoretical contribution of this analysis lies in establishing a direct conceptual linkage between data redundancy and the observed scaling exponent in materials property prediction. This connection rests on information-theoretic arguments and proof sketches, without reliance on numerical fitting or simulation. Redundancy systematically lowers the effective scaling exponent relative to the baseline expected under fully independent samples: when a fraction r of information is redundant, each additional sample contributes only a factor of (1 − r) new information, so the observed exponent approximates the independent-sample value scaled by (1 − r). This mechanism accounts for the notably shallower scaling in materials databases compared with simpler regression tasks of comparable model capacity [7, 8].

A related implication arises from the structural diversity constraints inherent to materials databases. The maximum attainable exponent is bounded by the narrowest dimension among the number of distinct crystal prototypes, elemental combinations, and local atomic environment types. Learning cannot extend beyond distinctions present in the training distribution, imposing hard ceilings—for instance, when only a limited set of prototypes exists, the exponent remains capped irrespective of total sample volume. Expansion strategies must therefore address these dimensions concurrently to avoid artificial performance limits.

Under conditions of strong clustering in feature space, the effective sample size becomes bounded by the number of clusters rather than the nominal count. Consequently, as dataset size grows, the scaling exponent is increasingly governed by a logarithmic relationship between cluster number and total entries, approaching zero unless fresh clusters emerge. This dynamic explains the particularly flat curves observed with prevalent local-environment redundancy in materials data [6]. The underlying proof sketches follow from reduced mutual information between redundant samples

and target properties, which slows refinement of the model's internal representation. These bounds hold independently of neural architecture and render scaling exponents theoretically predictable from measurable database structure [5, 10].

Redundancy can be estimated directly from routine database statistics, yielding a practical toolkit for assessing and forecasting scaling behavior prior to any model training.

Table 2 translates redundancy metrics into predictive scaling regimes, enabling ex ante estimation of learning efficiency prior to model training.

Table 2. Redundancy Metrics and Predictive Mapping to Scaling Regimes in Materials Databases

Metric	Definition	Measurement Method	Low-Redundancy Regime	High-Redundancy Regime	Predicted Scaling Behavior	Design Implication
Prototype Diversity	Number of distinct crystal structures	Count unique prototypes	Broad structural coverage	Few repeated prototypes	Higher α (steeper scaling)	Generate new structure classes
Compositional Entropy	Distribution spread of element combinations	Shannon entropy calculation	Uniform composition space	Concentrated compositions	Moderate α reduction	Target rare compositions
Environment Diversity	Variety of local atomic environments	Clustering of descriptors (e.g., SOAP)	Many small clusters	Few dominant clusters	Strong α suppression	Introduce new bonding motifs
Pairwise Dissimilarity	Average difference between samples	Random pair similarity scoring	Low similarity	High similarity	Continuous α flattening	Avoid near-duplicate entries
Composite Redundancy Score	Aggregate redundancy measure	Weighted combination of metrics	$r \to 0$	$r \to 1$	$\alpha \approx \alpha_0(1-r)$	Optimize diversity across all axes

Redundancy manifests concretely through prototype repetition, which is quantified simply by counting the number of distinct crystal prototypes in the database. High repetition among entries sharing identical prototypes markedly elevates overall redundancy, as seen in large repositories where roughly one thousand unique prototypes support more than one hundred fifty thousand total structures.

This structural bias extends naturally to compositional redundancy, assessed via the Shannon entropy of elemental distribution across the database. Low entropy reveals strong bias toward narrow sets of compositions, particularly near-equiatomic alloys and commercially prevalent chemistries, thereby leaving broad regions of chemical space undersampled and inflating redundancy.

A related mechanism operates at the level of local atomic environments, represented through standard descriptors such as smooth overlap of atomic positions or atom-centered symmetry functions. Clustering these descriptors and measuring the fraction of atoms covered by dominant clusters shows that when few clusters dominate, local-environment redundancy rises sharply and exerts particularly strong downward pressure on the scaling exponent.

Beyond these, pairwise similarity further captures micro-level redundancy by averaging structural or compositional similarity over random sample pairs; elevated average similarity signals pervasive overlap at the individual level.

These metrics integrate into a composite redundancy score that, through the conceptual relationship established earlier, directly predicts the effective scaling exponent. Databases registering low scores across the metrics exhibit steeper scaling and greater returns from expansion, whereas high scores on any dimension produce flattened behavior independent of nominal size. Computationally lightweight, the approach enables curation-stage intervention to favor diversity, providing the theoretical grounding for redundancy-aware data selection strategies that align dataset design with accelerated materials property prediction [7, 8].

Implications for Dataset Design in Materials Engineering

The theoretical framework developed here carries immediate consequences for how materials property databases should be designed and expanded. Because the scaling exponent is controlled by the redundancy fraction, any strategy that fails to reduce redundancy will produce only marginal gains in predictive power no matter how large the nominal dataset becomes. Database curators must therefore treat diversity as the primary objective rather than sheer volume.

When new entries are added, they should be selected to increase the count of distinct prototypes, elemental combinations, or local environment clusters. Adding another structure that belongs to an already well-represented prototype simply raises the redundancy fraction and flattens the exponent further. In contrast, introducing a novel prototype or an underrepresented elemental combination lowers redundancy and raises the effective sample size, thereby steepening the scaling curve. This principle explains why some databases achieve higher exponents despite smaller total sizes: their entries span more independent informational units.

Strategic investment decisions follow directly from the bounding propositions. Resources allocated to database growth should first address the dimension that currently imposes the tightest bound on the exponent. If prototype diversity is the limiting factor, funding should prioritize high-throughput computation of new crystal classes rather than incremental additions to familiar families. If compositional entropy is low, campaigns should target rare or extreme compositions that lie far from the current distribution center. Local-environment clustering metrics can guide the generation of structures with unusual coordination or bonding motifs.

The framework also implies that hybrid curation pipelines combining automated generation with explicit redundancy filters will outperform purely volume-driven approaches. Filters that discard or down-weight samples whose information is already captured by existing clusters preserve computational effort while maintaining a higher effective sample size. Earlier theoretical treatments of data selection in large materials repositories have already hinted at the value of such controls; the present analysis supplies the precise redundancy-based mechanism that justifies them.

Over the long term, database design should aim for a redundancy fraction that keeps the observed exponent above a chosen threshold, such as 0.5, which would make further scaling practically feasible for industrial applications. Achieving this target requires ongoing monitoring of the four redundancy metrics introduced earlier. Periodic recomputation of prototype counts, compositional entropy, cluster coverage, and pairwise similarities allows curators to detect when redundancy is creeping upward and to intervene before the scaling law flattens irreversibly.

In summary, the redundancy theory reframes dataset expansion from an exercise in accumulation to an exercise in diversification. By aligning curation priorities with the structural bounds on the exponent, materials engineers can ensure that each new sample contributes meaningfully to error reduction rather than merely inflating the nominal count. This shift promises substantially more efficient use of computational resources and accelerates the discovery pipeline in data-driven materials science [7, 8, 13, 14].

Connection to Sample Complexity Theory

The redundancy perspective developed in this work integrates naturally with classical sample-complexity results from statistical learning theory. Sample complexity describes the minimum number of independent observations required to achieve a target generalization error with high probability [15]. In the absence of redundancy, the scaling exponent α emerges directly from the covering number or VC-dimension of the hypothesis class. When redundancy is present, however, the effective sample size replaces the nominal size in these bounds, which immediately lowers the observed exponent.

Proposition 1 of the present analysis can therefore be viewed as a materials-specific realization of the general principle that correlated samples inflate sample-complexity requirements. Each redundant entry adds little new information about the underlying property landscape, so the learner must see far more raw samples to reach the same confidence level. The factor (1 – r) that appears in the effective-exponent formula mirrors the contraction that appears in sample-complexity bounds when data are drawn from a distribution with limited support or high dependence.

The structural bounds in Proposition 2 extend this connection further. The number of distinct prototypes, elemental combinations, and local environments effectively defines the size of the support of the data-generating distribution. Sample-complexity theory predicts that the number of required samples grows at least logarithmically with the size of this support. When a database covers only a small support, the exponent is capped regardless of how many repeated draws are added from within that support. The clustering argument in Proposition 3 is likewise a direct analogue of the fact that the sample complexity of learning a union of clusters is governed by the number of clusters rather than the total point count.

This unification clarifies why materials property prediction often demands larger datasets than simpler regression tasks even when the underlying function class is comparable. Materials data inhabit a high-dimensional space whose effective dimension is compressed by physical constraints, leading to the high redundancy fractions observed in practice. Theoretical results on sample complexity with multiple data sources or surrogate data reinforce the same conclusion: only the independent informational content matters [16, 17].

By casting neural scaling laws as the large-sample limit of sample-complexity behavior under redundancy, the present analysis supplies a bridge between empirical power-law observations and rigorous statistical learning bounds. It also suggests that techniques developed for private learning or hypothesis testing under limited samples—such as careful subsampling or importance weighting—could be adapted to materials databases to counteract redundancy and restore steeper scaling [17-23]. The framework therefore not only explains existing exponent variation but also points to principled methods for improving it.

Comparisons with Scaling Laws in Other Domains

Neural scaling laws appear across many scientific and engineering domains, yet the role of redundancy manifests differently depending on the nature of the data. In image classification, synthetic augmentation and near-infinite variation keep redundancy low, producing relatively steep exponents in the 0.07–0.35 range. In language modeling, the enormous diversity of natural text similarly maintains a large effective sample size. Materials databases, by contrast, are constrained by the discrete periodic table, finite crystal symmetries, and thermodynamic stability, which together enforce high redundancy and correspondingly flatter exponents.

The same redundancy mechanism appears in quantum-chemistry applications, where scaling can appear broken when datasets are dominated by a handful of molecular scaffolds. Theoretical examinations of scaling in chemical models have noted precisely this flattening when local environments repeat across molecules, mirroring the local-environment clustering bound derived here. In atomistic force-field learning, graph-neural-network studies similarly report exponents that plateau once the diversity of local bonding patterns is exhausted.

Even in non-materials scientific domains the pattern holds. When datasets are assembled from multiple sources with overlapping coverage, the effective exponent is reduced exactly as predicted by the redundancy fraction. Political-analysis work on scaling data from multiple sources and general treatments of learning with surrogate data both recover the same multiplicative contraction factor (1 – r) that appears in Proposition 1 [16, 17, 24]. The logarithmic bound in Proposition 3 also reappears in hierarchical or clustered data regimes studied in theoretical machine-learning literature.

The materials case is distinctive because the redundancy is not accidental but intrinsic to the physics. Unlike images or text, which can be arbitrarily diversified, materials entries must respect conservation laws and stability criteria. This physical grounding makes the redundancy bounds derived here especially predictive: they depend only on measurable structural statistics rather than on model-specific details. Consequently, the theory developed for materials databases offers a template that can be exported to other constrained scientific domains where data generation is expensive and diversity is limited by underlying laws.

Cross-domain comparison therefore validates the redundancy framework while highlighting why materials property prediction requires tailored theoretical analysis. The same principles explain both the modest exponents observed in practice and the pathway to steeper scaling through deliberate diversity engineering [11, 13, 17, 25-27].

Limitations and Future Extensions of the Redundancy Framework

The redundancy theory, while powerful, rests on several conceptual assumptions that suggest natural directions for refinement. The current propositions treat redundancy as a single scalar fraction averaged across the entire database. In reality, redundancy may vary across different regions of property space or across different target properties. Future extensions could introduce property-dependent redundancy metrics that allow exponent prediction to be localized rather than global.

The framework also assumes that redundancy acts multiplicatively on the independent-sample exponent. More sophisticated models could incorporate higher-order interactions, such as partial redundancy where some samples are redundant for one property but informative for another. Information-theoretic refinements that quantify mutual information between samples and multiple targets simultaneously would capture these nuances without introducing numerical simulation.

Another limitation arises from the assumption that the underlying hypothesis class remains fixed while only the data distribution changes. In practice, model architectures themselves evolve, and new architectures may interact differently with the same redundancy level. Theoretical work could explore how changes in model capacity modulate the effective redundancy experienced by the learner, potentially leading to architecture-aware bounds on the exponent.

Despite these limitations, the core insight—that database structure imposes hard bounds on scaling through redundancy—remains robust. Extensions to multi-task or multi-fidelity settings, where databases contain both cheap low-accuracy and expensive high-accuracy entries, would be especially valuable for materials engineering. The clustering bound could be generalized to hierarchical clusterings that reflect the natural taxonomy of crystal prototypes and compositions [28].

Overall, the present analysis provides a solid foundation that future theoretical studies can build upon by relaxing one assumption at a time while preserving the plain-language, proof-sketch style that makes the results accessible to materials scientists. Such incremental extensions will strengthen the predictive power of the framework and broaden its applicability across computational and data-driven materials engineering [5, 6, 10, 22, 24, 29].

Conclusion

This theoretical analysis has demonstrated that the wide variation in neural scaling exponents observed for materials property prediction arises fundamentally from the level of data redundancy within databases. By formalizing redundancy as the fraction of overlapping information and deriving explicit bounds from database structural features, the work supplies a unified conceptual explanation for previously unexplained empirical patterns. The propositions and metrics presented here translate directly into actionable guidance for dataset curation, model development, and resource allocation.

The central message is clear: scaling behavior is not an immutable property of neural networks but a predictable consequence of how information is distributed across samples. By minimizing redundancy, materials scientists can steepen the exponent, reduce the data volumes required for target accuracy, and accelerate discovery. The framework bridges empirical observations across multiple studies with a coherent theoretical account grounded exclusively in redundancy arguments.

As materials databases continue to grow, the ability to forecast and control scaling exponents will become an essential competency in computational materials engineering. The present analysis offers the conceptual tools required to exercise that control. Future work extending the redundancy perspective to multi-task, multi-fidelity, and architecture-aware settings will further refine these tools, ensuring that data-driven artificial intelligence delivers on its promise for materials innovation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Apr 2025 Revised: 29 Jul 2025 Accepted: 03 Nov 2025
Published online: 18 January 2026

References

Bahri Y, Dyer E, Kaplan J, Lee J, Sharma U. Explaining neural scaling laws. Proc Natl Acad Sci U S A. 2024;121(27):e2311878121. https://doi.org/10.1073/pnas.2311878121.

Zhang Z. Neural scaling laws from large-N field theory: Solvable model beyond the ridgeless limit. Mach Learn Sci Technol. 2025;6(2):025010. https://doi.org/10.1088/2632-2153/adc872.

Chen D, Zhu Y, Zhang J, Du Y, Li Z, Liu Q, et al. Uncovering neural scaling laws in molecular representation learning. Adv Neural Inf Process Syst. 2023;36:1452-75.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5. https://doi.org/10.1038/s41586-023-06735-9.

Trikha A, Chu K, Gosai A, Szachta P, Weiner E. Scaling laws for neural material models. arXiv Preprint arXiv:2509.21811. 2025.

Großmann M, Grunert M, Runge E. Broken neural scaling laws in materials science. arXiv Preprint arXiv:2602.05702. 2026.

Li K, Persaud D, Choudhary K, DeCost B, Greenwood M, Hattrick-Simpers J. Exploiting redundancy in large materials datasets for efficient machine learning with less data. Nat Commun. 2023;14(1):7283. https://doi.org/10.1038/s41467-023-42992-y.

Li Q, Fu N, Omee SS, Hu J. MD-HIT: Machine learning for material property prediction with dataset redundancy control. NPJ Comput Mater. 2024;10(1):245. https://doi.org/10.1038/s41524-024-01426-z.

Frey NC, Soklaski R, Axelrod S, Samsi S, Gomez-Bombarelli R, Coley CW, et al. Neural scaling of deep chemical models. Nat Mach Intell. 2023;5(11):1297-305. https://doi.org/10.1038/s42256-023-00740-3.

Ngo K, Ravanbakhsh S. Scaling laws and symmetry, evidence from neural force fields. arXiv Preprint arXiv:2510.09768. 2025.

Meir Y, Sardi S, Hodassman S, Kisos K, Ben-Noam I, Goldental A, et al. Power-law scaling to assist with key challenges in artificial intelligence. Sci Rep. 2020;10(1):19628. https://doi.org/10.1038/s41598-020-76764-1.

Mariani M, Moosavi SA, Ringach DL, Dipoppa M. A universal power law optimizes energy and representation fidelity in visual adaptation. bioRxiv [Preprint]. 2025. https://doi.org/10.1101/2025.03.20.643406.

Hernandez D, Kaplan J, Henighan T, McCandlish S. Scaling laws for transfer. arXiv Preprint arXiv:2102.01293. 2021.

Mansour E, Shahzad F, Tekli J, Chbeir R. Data redundancy management for leaf-edges in connected environments. Computing. 2022;104(7):1565-88. https://doi.org/10.1007/s00607-021-01051-4.

Schmidt L, Santurkar S, Tsipras D, Talwar K, Madry A. Adversarially robust generalization requires more data. Adv Neural Inf Process Syst. 2018;31:5019-31.

Hashimoto T. Model performance scaling with multiple data sources. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18-24; Virtual. Proc Mach Learn Res. 2021;139:4107-16.

Jain A, Montanari A, Sasoglu E. Scaling laws for learning with real and surrogate data. Adv Neural Inf Process Syst. 2024;37:110246-89.

Lee J, Heaukulani C, Ghahramani Z, James LF, Choi S. Bayesian inference on random simple graphs with power law degree distributions. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6-11; Sydney, Australia. Proc Mach Learn Res. 2017;70:2004-13.

Canonne CL. A survey on distribution testing: Your data is big. But is it blue? Theory Comput Libr Grad Surv. 2020;9:1-100. https://doi.org/10.4086/toc.gs.2020.009.

Tu SL. Sample complexity bounds for the linear quadratic regulator [dissertation]. Berkeley (CA): University of California, Berkeley; 2019.

Sadigurschi M, Stemmer U. On the sample complexity of privately learning axis-aligned rectangles. Adv Neural Inf Process Syst. 2021;34:28286-97.

Cheng HC, Datta N, Liu N, Nuradha T, Salzmann R, Wilde MM. An invitation to the sample complexity of quantum hypothesis testing. NPJ Quantum Inf. 2025;11(1):94. https://doi.org/10.1038/s41534-025-00980-8.

Benedetti M, Buhrman H, Weggemans J. Provable and verifiable quantum advantage in sample complexity. Phys Rev Lett. 2026;136(4):040601. https://doi.org/10.1103/q55v-wm7y.

Enamorado T, López-Moctezuma G, Ratkovic M. Scaling data from multiple sources. Polit Anal. 2021;29(2):212-35. https://doi.org/10.1017/pan.2020.24.

Cagnetta F, Favero A, Sclocchi A, Wyart M. Scaling laws and representation learning in simple hierarchical languages: Transformers versus convolutional architectures. Phys Rev E. 2025;112(6):065312. https://doi.org/10.1103/qtd6-nl8p.

Lee S, Dieng AB. Are neural scaling laws leading quantum chemistry astray? arXiv Preprint arXiv:2509.26397. 2025.

Li C, Ye Z, Pasini ML, Choi JY, Wan C, Balaprakash P. Scaling laws of graph neural networks for atomistic materials modeling. In: 2025 62nd ACM/IEEE Design Automation Conference (DAC); 2025 Jun 22-25; San Francisco, CA. IEEE; 2025. p. 1-7. https://doi.org/10.1109/DAC63849.2025.11132864.

Xia M, Lin H, Zhang W. Hierarchical Dimensionless Learning (Hi-π): A physics-data hybrid-driven approach for discovering dimensionless parameter combinations. arXiv Preprint arXiv:2507.18332. 2025.

Moro V, Loh C, Dangovski R, Ghorashi A, Ma A, Chen Z, et al. Multimodal foundation models for material property prediction and discovery. Newton. 2025;1(1):100016. https://doi.org/10.1016/j.newton.2025.100016.