HEABENCH: A MULTIMODAL DATASET AND BENCHMARK FOR HIGH-ENTROPY ALLOY UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-entropy alloys (HEAs) offer a rich compositional space for materials design, yet their heterogeneous atomic environments make learning structure–property relationships challenging. Progress is constrained in part by limited paired atomic and electronic data and by the lack of consistent protocols for evaluating representations across spatial scales. We introduce **HEABench**, a multimodal dataset and benchmark of 3,109 quinary, 100-atom configurations spanning 60 elements, with paired structures, total electron charge densities, and property targets, complemented by magnetic labels. HEABench supports global property prediction, local structure–charge matching, and site-resolved magnetic transfer under limited supervision. Experiments show that representing atomic positions as spatial fields via site-centered Gaussians recovers much of the electronic-target fusion benefit, whereas additional charge-density gains depend on the target and evaluation metric. Local matching under configuration and elemental identity controls reveals encoder differences obscured by near-saturated cross-configuration retrieval. With 50 magnetically labeled configurations, structure–charge pretraining reduces site-moment MAE by 10.9% relative to random initialization, using only atomic structures during inference. These findings highlight charge density as a source of transferable supervision beyond its target-dependent benefits as an additional prediction input, and establish HEABench as a testbed for learning structural and electronic representations in chemically disordered alloys. Project page: `https://anonymous.4open.science/r/HEABENCH-F6E1`.

1 INTRODUCTION

High-entropy alloys (HEAs) offer a vast chemical space for designing materials with tailored properties. Navigating this space requires data-driven models that capture both composition and atomic structure. Yet bulk composition alone cannot capture the diversity of local environments in these chemically complex alloys. Different atomic arrangements give rise to distinct local chemical environments and lattice distortions, as observed in HEAs (Moniri et al., 2023). These local environments also influence electronic behavior. First-principles calculations show that atoms of the same element can carry substantially different charges depending on their surrounding chemical environments (Ishibashi et al., 2020). This variability motivates joint learning from structural and electronic representations across global properties, local environments, and site-resolved responses.

Composition, atomic structure, and total charge density provide different representations of these systems. Composition summarizes elemental identities and fractions, the structure specifies atomic occupations and geometry, and the charge density reflects the spatial electronic response to local atomic environments. Their physical connection, however, does not mean that the encoded information is equally learnable or useful, which raises a central question: Can joint learning exploit information shared across these representations to improve prediction over learning from each alone?

Investigating this question requires paired structural and electronic data, together with evaluation protocols that distinguish the contributions of each representation. Existing datasets for HEAs and multi-principal-element alloys (MPEAs) support composition-based property prediction and modeling of explicit atomic configurations (Borg et al., 2020; Li et al., 2024; Lopanitsyna et al., 2023b). Charge-density benchmarks and multimodal materials frameworks further enable the study of electronic fields and their connections to other material representations (Chen et al., 2025; Xiang et al.,

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

2025; Moro et al., 2025). Nevertheless, resources that pair atomic structures with electron densities across diverse local chemical environments remain limited for chemically disordered alloys like HEAs, as do unified protocols for assessing their individual and combined predictive value. Addressing this gap requires a benchmark that couples paired structural and electronic data with controlled evaluation across different tasks.

Therefore, we introduce **HEABench**, a multimodal dataset and benchmark for evaluating structural and electronic representations in chemically disordered alloys. HEABench contains 3,109 periodic configurations, each comprising 100 atoms and five constituent elements, with 60 elements represented across the dataset. A common first-principles workflow links relaxed atomic structures, total electron densities, and associated property labels. Site-resolved magnetic labels further enable the study of representation transfer under limited supervision.

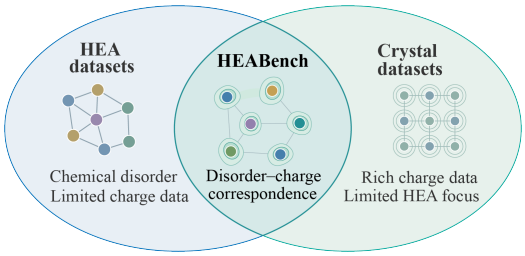


Figure 1: HEABench connects chemical disorder with paired structural and electronic representations, enabling joint study across diverse atomic environments.

These paired data support three complementary tasks that examine predictive utility, local correspondence, and transferability. **Global Multimodal Probing** measures the individual and combined predictive value of structural and electronic representations for configuration-level properties. **Local Structure-Charge Matching** tests whether models can identify corresponding local environments across the two representations and distinguish closely related environments. **Low-Label Magnetic Transfer** assesses whether pretrained representations improve site-resolved magnetic prediction when labeled data are scarce and only atomic structure is available at inference. Our results show that, beyond its target-dependent predictive benefits, charge density supports local correspondence and transferable structural representations, motivating broader representation-learning studies in HEAs and chemically disordered materials.

2 Background & Related Works

2.1 Structural and Electronic Representation Learning

Composition-based methods and crystal graph models provide established approaches to material-property learning (Ward et al., 2016; Wang et al., 2021; Xie & Grossman, 2018). Models such as ALIGNN and ComFormer incorporate geometric relations and periodicity into structural representations (Choudhary & DeCost, 2021; Yan et al., 2024), while electronic-property models extend prediction to structured outputs such as the density of states (DOS) (Ben Mahmoud et al., 2020; Kong et al., 2022). Charge-density learning bridges these directions: spatial electronic fields can serve as inputs for property prediction or as targets inferred from atomic structures (Zhao et al., 2020; Arora et al., 2022; Cheng & Peng, 2023; Kim & Ahn, 2024). Thus, field reconstruction and learning transferable field feature are related but distinct objectives for evaluating material representations.

Electronic-density benchmarks and multimodal frameworks further study these relationships. ECD evaluates crystal charge-density prediction and transfer from PBE to HSE data (Chen et al., 2025), whereas EDBench organizes density-centered prediction, retrieval, and generation for molecules (Xiang et al., 2025). CrysMMNet combines crystal structure and text for property learning (Das et al., 2023). More directly related to our setting, MultiMat aligns structure, DOS, charge density, and text during pretraining, evaluating both cross-modal retrieval and downstream property prediction (Moro et al., 2025). Its transferred encoders use structure alone at inference, showing that electronic information can contribute through pretraining without becoming a downstream input. These results motivate examining which aspects of cross-modal correspondence remain useful under chemical disorder. However, material-level retrieval alone cannot establish whether representations distinguish local environments that share elemental species and parent-configuration identity. Nor can it determine whether locally learned features transfer to site-resolved properties under limited downstream supervision. Addressing these questions requires paired local data and evaluating methods for fine-grained correspondence and site-resolved transfer in chemically disordered alloys.

Table 1: Representative HEA/MPEA resources. Source links open the cited publication or dataset. For prior work, ✓ denotes documented availability, including on request; × means not documented or not applicable. HEABench marks indicate inclusion in this work.

Datasets	**HEABench**	HEA25 2023	Li 2024	Mirzaee 2024	tabGAP 2021	SNAP 2020	Zhang 2022	Borg 2020	Detor 2022
Records n	**3,109**	25,628	83,797	1,550	2,329	5,905	7,086	1,545	86
Elements E	**60**	25	8	5	5	4	15	30	9
Five-element sets	**1,033**	3,563	56	1	1	0	0	78	22
Atoms N_A	**100**	36, 48	2–128	16	1–252	2–128	×	×	×
Cutoff (eV)	**550**	550	520	500	500	520	×	×	×
ϵ_{el} (eV/atom)	$\mathbf{10^{-7}}$	$10^{-6}/N_A$	10^{-5}	10^{-10}	×	$10^{-5}/N_A$	10^{-6}	×	×
f_{stop} (eV/Å)	**0.05**	×	$10^{-4\dagger}$	×	×	0.02	×	×	×
Relaxed structures	✓	×	✓	×	×	✓	×	×	×
3D charge grids	✓	×	×	✓	×	×	×	×	×
Electronic targets	✓	DOS/ eigenv.	×	×	×	×	×	×	×

ϵ_{el}: electronic convergence tolerance; f_{stop}: force threshold. Electronic targets: DOS, bands, or eigenvalues; named types indicate partial availability. †Li's relaxation criterion is 10^{-4} eV/atom, not a force threshold.

2.2 Data Resources for Chemically Disordered Alloys

General resources such as OQMD provide broad structure–energy data, while the Materials Project charge-density database links crystal structures to spatial electronic fields (Kirklin et al., 2015; Shen et al., 2022). In HEA and MPEA research, literature-curated and experimental collections relate composition and processing to observed phases and mechanical properties (Borg et al., 2020; Detor et al., 2022). The EMTO–CPA dataset of Zhang et al. (2022) enables composition-based elasticity prediction without explicit atomic occupations. In contrast, configuration-level resources retain the site occupations and atomic geometries needed for spatial modeling: NbMoTaW SNAP and MoNbTaVW tabGAP support interatomic-potential and defect studies (Li et al., 2020; Byggmästar et al., 2021), while HEA25 broadens chemical coverage for potential learning (Lopanitsyna et al., 2023b;a). The dataset of Li et al. (2024) further combines ordered and disordered alloys to study formation-energy prediction and generalization across structural and chemical regimes.

Despite these advances, explicit chemical disorder remains underrepresented in general materials databases, in part because of the cost of sampling and computing disordered configurations (Li et al., 2024). In alloys, electronic-field learning includes Mirzaee et al. (2024), who predict FCC alloy elastic constants from charge density, and Pathrudkar et al. (2025), who learn electron-density mappings across concentrated-alloy compositions. Table 1 summarizes the chemical coverage, configuration size, and electronic-data availability of HEA/MPEA resources. HEABench complements this evolving resource landscape with paired structures, total electron charge densities, and electronic targets across chemically disordered configurations, together with magnetic labels. Its controlled protocols enable complementary evaluations of predictive utility, local correspondence, and site-resolved transfer, thereby extending the data and benchmarks available for representation learning in HEAs.

3 Dataset

Initialization. The dataset contains 3,109 quinary, 100-atom periodic configurations. Initial structures are generated by random site substitution or special quasirandom structure (SQS) optimization, depending on the chemical subset, and retain explicit elemental occupations and atomic coordinates.

DFT workflow. Structural relaxation, static self-consistent calculation, and band-structure calculation are performed using VASP with PAW–PBE (Kresse & Furthmüller, 1996; Blöchl, 1994; Perdew et al., 1996). The high-throughput workflow associates each relaxed structure with its total electron charge density, formation energy, and electronic-structure targets. Dataset generation required over **2.5 million CPU core-hours**. Computational settings are provided in Appendix A.1.

Statistics. The configurations cover 60 elements and 1,033 distinct five-element combinations, with varied elemental fractions (Figure 2a,b). With explicit site occupations, this chemical diversity yields a broad range of local atomic environments. Meanwhile, each configuration is paired with a total charge-density field of median size 3.28×10^6 grid points (Figure 2c), providing spatially resolved electronic information alongside the atomic structure. Local-environment statistics, data validation, as well as sample visualizations are provided in detailed in Appendix A.3 for reference.

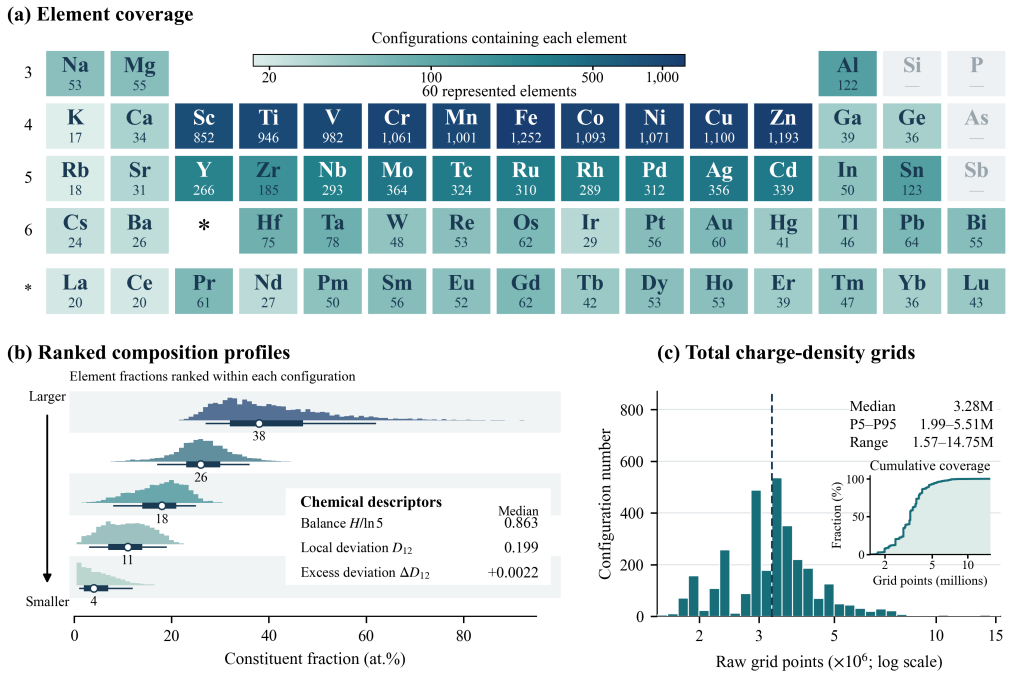


Figure 2: Chemical coverage, composition profiles, and charge-density grid sizes. **(a)** Configuration counts per element, with a logarithmic color scale; gray cells indicate zero coverage. Only periods containing represented elements and groups 1–15 are shown. **(b)** Ranked constituent fractions; labeled circles, thick bars, and thin bars mark medians, interquartile ranges, and 5th–95th percentiles. Profiles use 1-at.% bins and equal peak heights. Scalar summaries report configuration medians of composition balance $H/\ln 5$, local-composition deviation D_{12}, and its permutation-adjusted excess ΔD_{12} (Appendix A.3). **(c)** Raw total charge-density grid sizes for 3,109 valid structure–charge pairs, with logarithmic bins and horizontal axis. The dashed line marks the median, while the inset shows the empirical cumulative distribution of valid grid sizes, also on a logarithmic horizontal axis.

4 Benchmarking HEA Representation Learning Models

Global Multimodal Probing. We design this benchmark to examine the predictive utility of structure and charge density beyond composition. For each configuration c, the model receives one of four input settings: C_c, (C_c, S_c), (C_c, ρ_c), or (C_c, S_c, ρ_c), where C_c, S_c, and ρ_c denote composition, periodic atomic structure, and three-dimensional charge density. These inputs describe the same configuration at different levels, from elemental fractions and atomic arrangements to the electronic distribution obtained through DFT. Targets include formation energy E_{form}, near-Fermi scalar descriptors, and vector-valued DOS and band-structure targets. These targets test representation-dependent gains and whether structure–charge fusion improves prediction over either branch alone.

Local Structure–Charge Matching. To probe local correspondence in chemically disordered alloys, we formulate structure–charge matching as a bidirectional cross-modal retrieval task. Paired structure and charge-density patches describe the same local region, and each patch retrieves its counterpart among candidates from the other modality. By varying whether negative candidates share the same reference element pair or parent configuration, we test whether learned representations support fine-grained cross-modal matching beyond coarse chemical or configuration identity.

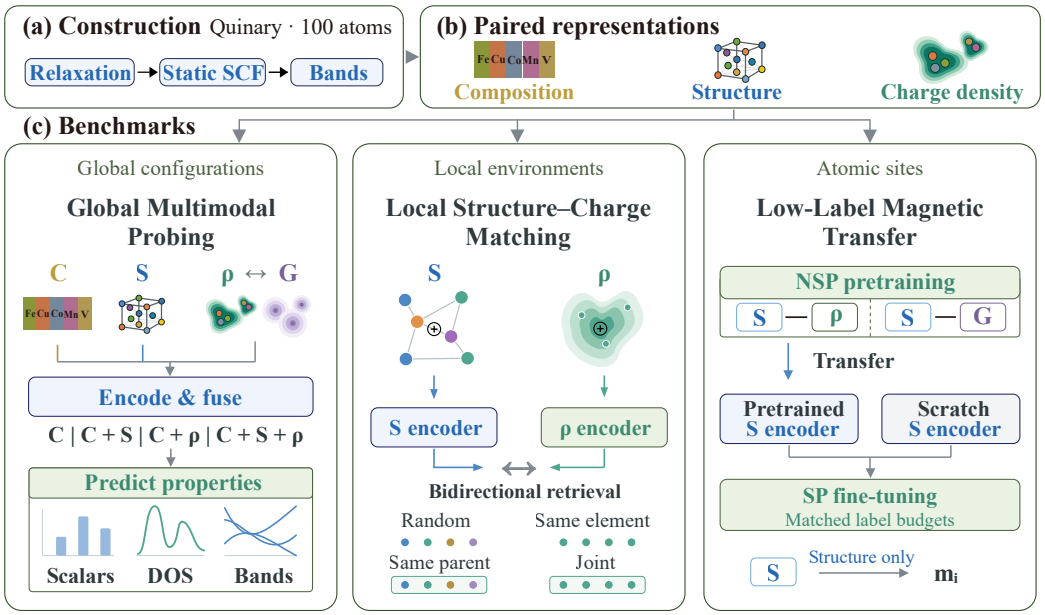


Figure 3: Overview of HEABench. **Figure (a)** Sampling chemically disordered HEA configurations and calculating through a shared DFT workflow. **Figure (b)** Paired composition C, structure S, and charge density ρ, with electronic properties and available magnetic labels. **Figure (c)** Benchmark for global multimodal probing, local structure–charge matching, and low-label magnetic transfer. G denotes a Gaussian encoding of atomic structure without self-consistent electronic response.

Low-Label Magnetic Transfer. Magnetic moments characterize spin-dependent electronic behavior and require additional spin-resolved calculations to obtain training labels. This motivates transferring representations learned from more abundant structure–charge pairs computed without distinguishing electron spins. We pretrain a structure encoder through local structure–charge matching, then fine-tune it with newly initialized heads to predict site and cell magnetic moments under limited supervision. The benchmark asks whether cross-modal pretraining yields transferable structural magnetic representations that improve magnetic label efficiency, using structure alone at inference.

4.1 Methods

Global multimodal probing. Per-modality inputs are encoded and projected into a shared latent space, $h_c^v = A_v(f_v(x_c^v))$, where $v \in \{C, S, F\}$ denotes composition, structure, or a spatial field. The composition branch uses a three-layer MLP on element and fraction features. Available structure and field representations update the composition representation through gated residual fusion:

$$h_c = \mathrm{LayerNorm}\left(h_c^C + \sum_{v \in \mathcal{V}_c} \alpha_c^v \odot h_c^v\right), \qquad \widehat{\widetilde{y}}_c = p_t(h_c). \tag{1}$$

Here $\mathcal{V}_c \subseteq \{S, F\}$ contains the available branches, α_c^v is a learned elementwise gate, and p_t predicts the training-standardized targets of separate family t, including formation energy, electronic scalars, DOS, and bands. For standardized residuals $r_{cj} = \widehat{\widetilde{y}}_{cj} - \widetilde{y}_{cj}$, we use the masked Smooth L1 objective

$$\mathcal{L}_t = \frac{\sum_{c,j} w_{cj}\, \ell_\delta(r_{cj})}{\sum_{c,j} w_{cj}}, \qquad \ell_\delta(r) = \begin{cases} r^2/(2\delta), & |r| < \delta, \\ |r| - \delta/2, & |r| \geq \delta, \end{cases} \tag{2}$$

where j indexes target components, $\delta > 0$ sets the quadratic-to-linear transition, and $w_{cj} \geq 0$ weights labels, with zero weight for missing values. To assess the added value of charge density, we replace ρ with a structure-derived periodic Gaussian field G under matched experimental conditions. Appendix B.3 details the fusion architecture, Gaussian-field construction, and input normalization.

Local structure–charge matching. Each example pairs a structure patch P_i^S with a charge-density patch P_i^ρ, both centered at the same periodic atom-pair mid-point. Then separate encoders map the patches to a shared latent space. The outputs are normalized as $h' = h/\|h\|_2$ and compute cosine similarities $s_{ij} = (h_i^{S\prime})^\top h_j^{\rho\prime}$. For a batch of B matched pairs, we minimize the symmetric contrastive objective

$$\mathcal{L}_{\text{match}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)} + \log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ji}/\tau)}\right], \tag{3}$$

where τ is the temperature and other batch pairs provide negatives in both retrieval directions. We combine four structure encoders with four charge-density encoders, including LSHCE, our implementation for local charge-density encoding, under a common training protocol.

For a non-learned reference beyond fixed spatial descriptors, we also construct a Gaussian radial reference. It pools the Gaussian field G and charge density ρ into concentric shells and compares their normalized radial profiles. This reference retains both modalities but replaces learned encoding with a fixed matching rule. Appendices B.2 and B.1 detail the encoders and radial construction.

Low-label magnetic transfer. We transfer the structure encoder from local structure–charge matching, discard the charge-density branch, and initialize site and cell prediction heads. Given site embeddings $h_{ci} = f_S(S_c)_i$, the heads predict $\widehat{m}_{ci} = q_{\text{site}}(h_{ci})$ and $\widehat{M}_c = q_{\text{cell}}(\overline{h}_c)$, where $\overline{h}_c$ denotes the mean site embedding. The encoder and both heads are jointly fine-tuned using structure alone. To accommodate global spin reversal, we select a single shared sign for all site and cell predictions in each configuration, yielding the joint objective:

$$\mathcal{L}_{\text{mag}} = \frac{1}{B}\sum_{c=1}^{B}\min_{s\in\{-1,+1\}}\left[\ell_{\text{site}}(s\widehat{\mathbf{m}}_c, \mathbf{m}_c) + \lambda\,\ell_{\text{cell}}(s\widehat{M}_c, M_c)\right], \tag{4}$$

where $\mathbf{m}_c$ and M_c denote reference site and cell moments, and B is the batch size. The site term averages Smooth L1 losses over sites, while λ weights the cell term. To examine the role of the pretraining field, we replace ρ with G during local matching on the same source configurations and transfer only the structure encoder. We compare charge-density pretraining with random initialization across label budgets, adding Gaussian-field pretraining as a low-label control,Appendix B.3

4.2 Evaluation

We evaluate prediction using MAE and RMSE in the original target scale, capturing average absolute error and the greater influence of large residuals, respectively. Retrieval uses R@1 and MRR, averaged over structure-to-charge and charge-to-structure directions. Within each protocol, comparisons use common evaluation samples and identical retrieval candidate lists for every model.

Table 2: Evaluation metrics.

Metric	Interpretation
MAE ↓	Mean absolute error
RMSE ↓	Greater weight on large errors
NLL ↓	Calibrated Gaussian score
R@1 ↑	Correct match ranked first
MRR ↑	Mean reciprocal match rank

For global prediction, we additionally report validation-calibrated Gaussian negative log-likelihood:

$$\text{NLL}_M = \left\langle \log\sigma_{Mj} + \frac{r_{Mcj}^2}{2\sigma_{Mj}^2} + \frac{1}{2}\log(2\pi)\right\rangle_{c,j}, \quad c: \text{configuration}, \quad j: \text{target component}. \tag{5}$$

Here r_{Mcj} is the residual standardized using training-target statistics, and σ_{Mj} is fitted on validation residuals and fixed during testing. We average over valid components within each configuration, then over configurations, giving each configuration equal weight in the resulting score. NLL therefore reflects both test errors and their agreement with the validation-fitted error scale, complementing MAE and RMSE across the input representations and highlighting the metric-dependent rankings.

For a lower-is-better score L, we define the gain over reference B as $g_L = L_B - L_M$, with positive values favoring model M. Relative MAE/RMSE reductions are $100 g_L/L_B\%$, whereas retrieval

score differences are expressed in percentage points rather than relative percentages. The corresponding figures and tables specify the reference condition and direction of each comparison. Where reported, pointwise 95% confidence intervals are estimated by paired bootstrap over configurations. Each resample uses the same configurations for both conditions and retains their associated sites or queries, preserving the pairing between methods and the dependence among observations within each configuration. Full evaluation protocols are provided in Appendices B.4.

5 EXPERIMENT

HEABench provides a challenging testbed for learning from atomic structure and electron charge density in HEAs. We conduct series of experiments to study predictive complementarity, fine-grained cross-modal correspondence, and label-efficient transfer, addressing three questions:

Q1: When does combining related modalities improve prediction over either modality alone?

Q2: Does cross-modal matching capture fine-grained correspondence beyond coarse-grained cues?

Q3: Can cross-modal pretraining improve representation transfer and downstream label efficiency?

5.1 Q1: GLOBAL MULTIMODAL PROBING

Table 3: **Global prediction with structure and spatial fields.** MAE / RMSE in the original target scale (lower is better). G: structure-derived Gaussian field; ρ: electron charge density. Target definitions: Appendix C.1.

Input	Formation (E_{form})	DOS at E_F (d_F)	Crossings (N_{cross})	Events (N_{event})	Slope (s_{cross})	DOS curve (**d**)	Bands (**b**)
C	0.0995 / 0.2431	0.1355 / 0.2148	1.1202 / 1.8023	3.5394 / 5.2230	0.0858 / 0.1274	0.1177 / 0.1998	0.0189 / 0.0266
$C+S$	**0.0455** / **0.0957**	0.1151 / 0.1919	0.8820 / 1.2462	2.7808 / 3.6825	0.0832 / 0.1446	0.0981 / 0.1682	0.0155 / 0.0234
$C+\rho$	0.0544 / 0.1103	0.1048 / 0.1467	**0.6811** / **0.9366**	**2.1564** / **2.7384**	0.0774 / 0.1147	0.0910 / 0.1523	0.0154 / 0.0224
$C+S+G$	0.0474 / 0.1365	0.0849 / 0.1181	0.7369 / 0.9586	2.2885 / 2.9044	**0.0754** / **0.1116**	0.0825 / 0.1359	**0.0148** / 0.0218
$C+S+\rho$	0.0515 / 0.1660	**0.0835** / **0.1171**	0.7306 / 0.9968	2.2148 / 2.8658	0.0781 / 0.1138	**0.0812** / **0.1317**	0.0149 / **0.0215**

Gaussian fields recover much of the fusion gain. Charge-density fusion reduces MAE and RMSE on all 6 electronic targets relative to $C+S$, but worsens formation energy (E_{form}) and trails $C+\rho$ on crossing and event counts (N_{cross}, N_{event}; Table 3). Gaussian fusion already achieves much of the electronic-target improvement: replacing the Gaussian with charge density gives MAE reductions of only -3.6% to $+3.2\%$. Thus, spatial-field representations account for much of the fusion benefit with charge density adds target-specific gains.

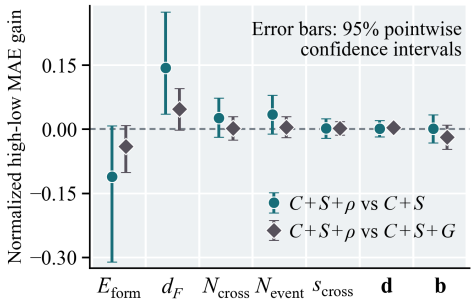
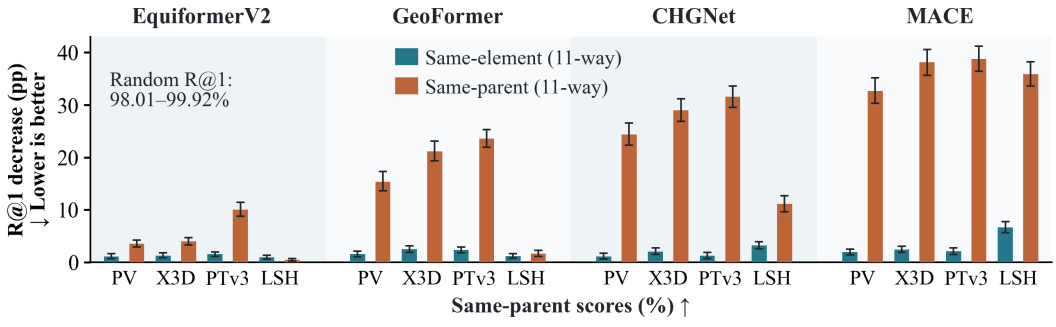


Figure 4: High–low MAE-gain contrasts.

Residual gains depend on the metric. For N_{cross}, charge-density fusion yields a 0.9% MAE improvement over Gaussian but a 4.0% RMSE degradation: smaller average absolute errors come at the cost of greater sensitivity to large residuals. The DOS vector (**d**) shows more consistent improvements: both errors decrease, and calibrated NLL improves by 0.029 (pointwise interval $[0.005, 0.062]$). For N_{event}, however, NLL favors Gaussian even though charge-density fusion yields lower point errors. Gaussian also achieves lower values on both error metrics for crossing slope (s_{cross}); for the band-energy vector (**b**), Gaussian has lower MAE while charge-density fusion has lower RMSE. Charge-density fusion shows its clearest advantage in DOS-curve prediction rather than a general improvement across electronic targets, and generally, the environment contrasts reveal no consistent cross-target trend (Figure 4; Appendix C.1). These global analyses quantify aggregate predictive contributions, but they do not directly probe whether structure and charge remain mutually consistent at the local level. We therefore next turn to a direct test of local structure–charge correspondence.

5.2 Q2: Local Structure–Charge Matching

Within-configuration retrieval reveals encoding differences. We compare two coarse identity controls: parent configuration (the source atomic structure) and reference element (the species at the designated atom-pair endpoint). Under the first control, different-parent retrieval is nearly saturated across all encoding combinations (98.01–99.92% R@1), leaving little room to distinguish encoders. Matching the reference element across different parents also has only a modest effect. In contrast, same-parent candidates reveal a much wider spread (60.27–99.47%; Figure 5), clearly separating local matching performance across encoding strategies.



	EquiformerV2				GeoFormer				CHGNet				MACE			
	PV	X3D	PTv3	LSH	PV	X3D	PTv3	LSH	PV	X3D	PTv3	LSH	PV	X3D	PTv3	LSH
R@1	96.2	95.4	89.5	99.5	84.1	77.9	75.3	98.1	75.0	70.6	68.0	88.5	66.7	61.2	60.3	62.1
MRR	97.9	97.5	94.3	99.7	90.7	87.0	85.8	98.9	85.0	82.3	80.7	93.4	79.3	75.6	75.0	76.1

Figure 5: 11-way bidirectional retrieval (n = 3,750). Bars: decreases from Random with paired pointwise 95% parent-bootstrap intervals. Rows: Same-parent scores (%). PV: PointVector; X3D: X-3D; LSH: LSHCE.

Accurate matching persists under joint identity controls. In Joint 4-way retrieval, candidates share both parent configuration and reference element. R@1 ranges from 72.29% to 99.75%, above the 25% chance level, with MRR following a similar pattern (Appendix C.2). These results show that local matching remains feasible even when parent and reference-element identity alone cannot distinguish candidates. For comparison, a fixed, untrained Gaussian radial reference matches normalized radial profiles of the structure-derived Gaussian field G and charge density ρ.

Table 4: Joint 4-way differences from Gaussian radial reference on identical queries (n = 2,777) and candidates. Bidirectional point estimates: Δ = learned − reference (pp); positive is better; **bold** denote the best.

Structure	Charge-density encoding							
Ref. R@1/MRR (%)	PointVector		X-3D		PTv3		LSHCE	
77.93 / 87.48	ΔR@1	ΔMRR	ΔR@1	ΔMRR	ΔR@1	ΔMRR	ΔR@1	ΔMRR
EquiformerV2	+19.23	+11.06	+19.43	+11.16	+17.73	+10.29	**+21.82**	**+12.39**
GeoFormer	+10.80	+6.45	+6.82	+4.28	+9.92	+6.11	**+21.05**	**+11.99**
CHGNet	+5.85	+3.70	+3.73	+2.53	+3.38	+2.34	**+13.83**	**+8.16**
MACE	**+0.40**	**+0.48**	−2.68	−1.27	−3.01	−1.46	−5.64	−3.02

Learned representations capture local distinctions beyond radial matching. The Gaussian radial reference provides a strong non-learned baseline, reaching 77.93% R@1 under the same joint identity constraints. Thirteen of the 16 learned combinations exceed it on both metrics in terms of point estimates, with gains of up to 21.82 and 12.39 percentage points in R@1 and MRR, respectively (Table 4). These gains reflect that learned representations capture local distinctions missed by fixed radial matching, even under joint identity constraints that leave candidates indistinguishable.

5.3 Q3: Low-Label Magnetic Transfer

Beyond the local correspondence examined in Q2, we test whether matching-based pretraining supports label-efficient magnetic prediction (Q3), using structure alone at downstream inference.

Charge-density pretraining improves site-level prediction most at the lowest label budget. Structure–ρ pretraining reduces site-level MAE and RMSE across all three budgets (Figure 6a,b).

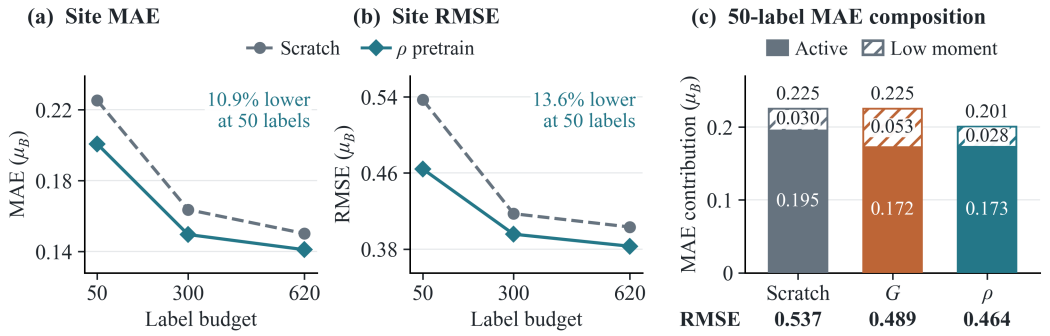


Figure 6: Magnetic transfer. (a,b) Site errors versus label budget (configurations). (c) At 50 labels, site-count-weighted contributions from active (larger reference moments) and low-moment (near-zero reference moments, $|m| \leq 0.1\,\mu_B$) sites sum to overall MAE. Overall RMSE is listed below; G and ρ denote pretraining fields.

Relative to SCRATCH (i.e., training from scratch), the gains reach 10.9% in MAE and 13.6% in RMSE at 50 labels, and decreasing to 6.1% and 5.0% at the full budget. These improvements appear in both average error and the larger residuals emphasized by RMSE. However, cell-level gains are less consistent: at the full budget, cell-level MAE is higher than that of SCRATCH (Appendix 11).

Relative to G, charge-density pretraining preserves active-site gains while improving low-moment prediction. The 50-label control identifies which site groups benefit (Figure 6c). Pretraining with the Gaussian field G already reduces active-site errors (larger reference moments) and improves overall RMSE by 8.9%. Its overall MAE remains nearly unchanged because these gains are offset by increased low-moment errors (near-zero reference moments). Charge-density pretraining retains the active-site benefit without this low-moment degradation. Compared with G, it lowers overall MAE and RMSE by 10.9% and 5.2%, respectively, with both low-moment errors reduced by about 48%; the active-site difference remains unresolved. Its additional value lies in accurate near-zero predictions alongside gains on larger moments, a distinction obscured by aggregate scores.

6 CONCLUSION

HEABench pairs structural and electronic data with controlled evaluation at global, local, and site levels in chemically disordered alloys. Spatial encoding accounts for much of the global fusion benefit, while additional charge-density gains vary by target and metric. Beyond global prediction, controlled local matching reveals encoder differences obscured by coarse retrieval, and structure–charge pretraining improves low-label magnetic prediction using structure alone at inference. Together, these findings highlight electronic data as both prediction inputs and supervision for local correspondence and transferable structural features. **HEABench supports extensions to new alloy chemistries, site-resolved properties, and pretraining objectives.** Its paired records, shared splits, and controlled protocols enable reproducible comparisons and community contributions toward understanding representation generalization in HEAs and other chemically disordered materials.

7 AI USE STATEMENT

In this work, we used generative AI tools to support computational workflows, research analysis, and manuscript preparation. For computational workflows and experiments, these tools assisted with scripting high-throughput job scheduling, sequential training runs, and large-scale data processing. During research planning and analysis, they supported brainstorming, literature identification and summarization, refinement of experimental analyses, and interpretation of preliminary results. For manuscript preparation, they assisted with drafting text and revising scientific figures. We reviewed the AI-assisted outputs, checked their consistency with the underlying experimental results, and take full responsibility for the final content, including the text, code, figures, and scientific claims.

REFERENCES

Gaurav Arora, Serveh Kamrava, Pejman Tahmasebi, and Dilpuneet S. Aidhy. Charge-density based convolutional neural networks for stacking fault energy prediction in concentrated alloys. *Materialia*, 26:101620, 2022. doi: 10.1016/j.mtla.2022.101620.

Chiheb Ben Mahmoud, Andrea Anelli, Gábor Csányi, and Michele Ceriotti. Learning the electronic density of states in condensed matter. *Physical Review B*, 102:235130, 2020. doi: 10.1103/PhysRevB.102.235130. URL https://arxiv.org/abs/2006.11803.

Peter E. Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953–17979, 1994. doi: 10.1103/PhysRevB.50.17953. URL https://doi.org/10.1103/PhysRevB.50.17953.

Christopher K. H. Borg, Carolina Frey, Jasper Moh, Tresa M. Pollock, Stéphane Gorsse, Daniel B. Miracle, Oleg N. Senkov, Bryce Meredig, and James E. Saal. Expanded dataset of mechanical properties and observed phases of multi-principal element alloys. *Scientific Data*, 7:430, 2020. doi: 10.1038/s41597-020-00768-9. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7722748/.

J. Byggmästar, K. Nordlund, and F. Djurabekova. Modeling refractory high-entropy alloys with efficient machine-learned interatomic potentials: Defects and segregation. *Physical Review B*, 104:104101, 2021. doi: 10.1103/physrevb.104.104101.

Pin Chen, Zexin Xu, Qing Mo, Hongjin Zhong, Fengyang Xu, and Yutong Lu. ECD: A machine learning benchmark for predicting enhanced-precision electronic charge density in crystalline inorganic materials. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0f4bbaaaf1e167f79134dd4cf11e3ed4-Abstract-Conference.html.

Chaoran Cheng and Jian Peng. Equivariant Neural Operator Learning with Graphon Convolution. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2311.10908.

Kamal Choudhary and Brian DeCost. Atomistic Line Graph Neural Network for improved materials property predictions. *npj Computational Materials*, 7:185, 2021. doi: 10.1038/s41524-021-00650-1.

Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. CrysMMNet: Multimodal Representation for Crystal Property Prediction. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 507–517, 2023. URL https://proceedings.mlr.press/v216/das23a.html.

Andrew Detor, Scott Oppenheimer, Rebecca Casey, and Cole Crawford. Refractory high entropy alloy dataset with room temperature ductility screening. *Data in Brief*, 45:108582, 2022. doi: 10.1016/j.dib.2022.108582.

Shoji Ishibashi, Yuji Ikeda, Fritz Körmann, Blazej Grabowski, and Jörg Neugebauer. Correlation analysis of strongly fluctuating atomic volumes, charges, and stresses in body-centered cubic refractory high-entropy alloys. *Physical Review Materials*, 4:023608, 2020. doi: 10.1103/PhysRevMaterials.4.023608. URL https://journals.aps.org/prmaterials/abstract/10.1103/PhysRevMaterials.4.023608.

Seongsu Kim and Sungsoo Ahn. Gaussian Plane-Wave Neural Operator for Electron Density Estimation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23805–23824, 2024. URL https://proceedings.mlr.press/v235/kim24b.html.

Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1:15010, 2015. doi: 10.1038/npjcompumats.2015.10. URL https://www.nature.com/articles/npjcompumats201510.

Shufeng Kong, Francesco Ricci, Dan Guevarra, Jeffrey B. Neaton, Carla P. Gomes, and John M. Gregoire. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nature Communications*, 13:949, 2022. doi: 10.1038/s41467-022-28543-x. URL https://www.nature.com/articles/s41467-022-28543-x.

Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996. doi: 10.1103/PhysRevB.54.11169. URL https://doi.org/10.1103/PhysRevB.54.11169.

Kangming Li, Kamal Choudhary, Brian DeCost, Michael Greenwood, and Jason Hattrick-Simpers. Efficient first principles based modeling via machine learning: from simple representations to high entropy materials. *Journal of Materials Chemistry A*, 12:12412–12422, 2024. doi: 10.1039/D4TA00982G. URL https://doi.org/10.1039/D4TA00982G.

Xiang-Guo Li, Chi Chen, Hui Zheng, Yunxing Zuo, and Shyue Ping Ong. Complex strengthening mechanisms in the NbMoTaW multi-principal element alloy. *npj Computational Materials*, 6:70, 2020. doi: 10.1038/s41524-020-0339-0.

Nataliya Lopanitsyna, Guillaume Fraux, Maximilian A. Springer, Sandip De, and Michele Ceriotti. Modeling high-entropy transition-metal alloys with alchemical compression: dataset HEA25, 2023a. URL https://archive.materialscloud.org/records/jkaf5-anv04. Version 1, dataset.

Nataliya Lopanitsyna, Guillaume Fraux, Maximilian A. Springer, Sandip De, and Michele Ceriotti. Modeling high-entropy transition metal alloys with alchemical compression. *Physical Review Materials*, 7(4):045802, 2023b. doi: 10.1103/PhysRevMaterials.7.045802. URL https://doi.org/10.1103/PhysRevMaterials.7.045802.

Hossein Mirzaee, Ramin Soltanmohammadi, Nathan Linton, Jacob Fischer, Serveh Kamrava, Pejman Tahmasebi, and Dilpuneet Aidhy. Elastic constants from charge density distribution in FCC high-entropy alloys using CNN and DFT. *APL Machine Learning*, 2:046104, 2024. doi: 10.1063/5.0229105.

Saman Moniri, Yao Yang, Jun Ding, Yakun Yuan, Jihan Zhou, Long Yang, Fan Zhu, Yuxuan Liao, Yonggang Yao, Liangbing Hu, Peter Ercius, and Jianwei Miao. Three-dimensional atomic structure and local chemical order of medium- and high-entropy nanoalloys. *Nature*, 624:564–569, 2023. doi: 10.1038/s41586-023-06785-z. URL https://www.nature.com/articles/s41586-023-06785-z.

Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Samuel Kim, Peter Y. Lu, Thomas Christensen, and Marin Soljačić. Multimodal foundation models for material property prediction and discovery. *Newton*, 2025. doi: 10.1016/j.newton.2025.100016. URL https://doi.org/10.1016/j.newton.2025.100016.

Shashank Pathrudkar, Stephanie Taylor, Abhishek Keripale, Abhijeet S. Gangan, Ponkrshnan Thiagarajan, Shivang Agarwal, Jaime Marian, Susanta Ghosh, and Amartya S. Banerjee. Electronic structure prediction of medium and high entropy alloys across composition space. *npj Computational Materials*, 11:378, 2025. doi: 10.1038/s41524-025-01856-3.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865. URL https://doi.org/10.1103/PhysRevLett.77.3865.

Jimmy-Xuan Shen, Jason M. Munro, Matthew K. Horton, Patrick Huck, Shyam Dwaraknath, and Kristin A. Persson. A representation-independent electronic charge density database for crystalline materials. *Scientific Data*, 9(1):661, 2022. doi: 10.1038/s41597-022-01746-z. URL https://doi.org/10.1038/s41597-022-01746-z.

Anthony Yu-Tung Wang, Steven K. Kauwe, Ryan J. Murdock, and Taylor D. Sparks. Compositionally restricted attention-based network for materials property predictions. *npj Computational Materials*, 7:77, 2021. doi: 10.1038/s41524-021-00545-1.

Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2:16028, 2016. doi: 10.1038/npjcompumats.2016.28.

Hongxin Xiang, Ke Li, Mingquan Liu, Zhixiang Cheng, Bin Yao, Wenjie Du, Jun Xia, Li Zeng, Xin Jin, and Xiangxiang Zeng. EDBench: Large-scale electron density data for molecular modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4820. URL https://papers.neurips.cc/paper_files/paper/2025/hash/d412ba5c76774f921762644e82515da3-Abstract-Datasets_and_Benchmarks_Track.html.

Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120:145301, 2018. doi: 10.1103/physrevlett.120.145301.

Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Complete and Efficient Graph Transformers for Crystal Material Property Prediction. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0ab51646ca369140c3c3ece011b66587-Abstract-Conference.html.

Jie Zhang, Chen Cai, George Kim, Yusu Wang, and Wei Chen. Composition design of high-entropy alloys with deep sets learning. *npj Computational Materials*, 8:89, 2022. doi: 10.1038/s41524-022-00779-7. URL https://www.nature.com/articles/s41524-022-00779-7.

Yong Zhao, Kunpeng Yuan, Yinqiao Liu, Steph-Yves Louis, Ming Hu, and Jianjun Hu. Predicting Elastic Properties of Materials from Electronic Charge Density Using 3D Deep Convolutional Neural Networks. *The Journal of Physical Chemistry C*, 124:17262–17273, 2020. doi: 10.1021/acs.jpcc.0c02348.

Alex Zunger, S.-H. Wei, L. G. Ferreira, and James E. Bernard. Special quasirandom structures. *Physical Review Letters*, 65(3):353–356, 1990. doi: 10.1103/PhysRevLett.65.353.

A Dataset

A.1 DFT Workflow and Computational Settings

Initial configurations contain five elements and 100 atoms. Depending on the chemical subset, site occupations are initialized by random substitution or special quasirandom structure (SQS) optimization (Zunger et al., 1990), with SQS construction performed using ICET. Each configuration undergoes structural relaxation, a static self-consistent calculation, and a subsequent band-structure calculation.

Calculations use VASP 6.3.2 with PAW potentials and the PBE functional (Kresse & Furthmüller, 1996; Blöchl, 1994; Perdew et al., 1996). Table 5 lists the key numerical settings.

Table 5: Key DFT settings used to construct HEABench.

Setting	Value
PAW datasets	PBE, version 5.4
Plane-wave cutoff	550 eV
Electronic convergence (EDIFF)	10^{-5} eV per cell
Ionic stopping setting (EDIFFG)	-0.05 eV/Å
Gaussian smearing	ISMEAR=0; SIGMA=0.1 eV

Table 6: Element-specific initial magnetic moments in μ_B per atom.

Element	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn
Initial moment	1.0	0.0	3.0	5.0	5.0	5.0	3.0	2.0	0.0	0.0
Element	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd
Initial moment	0.6	0.6	1.0	1.0	1.0	1.0	0.8	0.6	0.2	0.0

Unlisted elements use the configured default. Values are initial moments, not converged labels.

Magnetic initialization. For spin-polarized calculations, initial site moments are assigned by element through `MAGMOM`, as summarized in Table 6. These values initialize the self-consistent calculation; the magnetic labels are obtained from the converged solution.

Dataset generation required approximately 2.52×10^6 CPU core-hours: 1.78×10^6 for relaxation, 0.05×10^6 for static self-consistent calculations, and 0.69×10^6 for band calculations. Computations used 11 Linux nodes, each with 256 logical CPU cores and approximately 512 GiB RAM.

A.2 Sample Visualization

Figure 7 illustrates a single $Y_{26}Hg_{19}Gd_{19}Mg_{19}Pt_{17}$ configuration, showing its initial and relaxed structures alongside the paired charge-density field in two corresponding views.

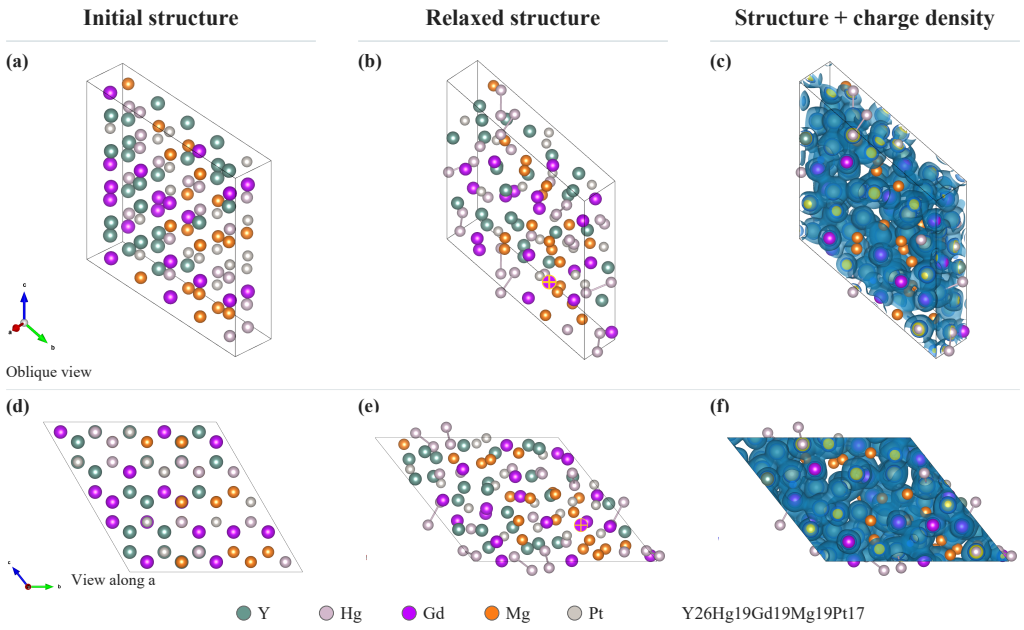


Figure 7: VESTA visualization of a paired sample. Columns show the initial structure, relaxed structure, and charge overlay; rows show oblique and a axis views. Atom colors identify elements. Both charge panels use positive isosurface levels of 0.019576 (teal), 0.10576 (blue), and 0.229576 (yellow) e/bohr^3. Yellow crosses are software atom-selection markers.

A.3 Statistic Metrics

Chemical coverage and composition profiles. Element occurrence counts the configurations containing each element, rather than the number of atoms. For each configuration, the five constituent fractions are sorted from largest to smallest; each rank is then summarized across configurations. The profiles have equal peak heights to facilitate shape comparison. Composition balance

is measured by normalized entropy,

$$\frac{H_c}{\ln 5} = -\frac{1}{\ln 5}\sum_{b=1}^{5} x_{cb}\ln x_{cb}, \tag{6}$$

where x_{cb} is the fraction of element b in configuration c. Values closer to one indicate more balanced fractions; this quantity does not depend on the spatial arrangement of atoms.

Local composition deviation. For site i, let $p_{cib}^{(k)}$ be the fraction of element b among its k nearest distinct neighboring sites under periodic boundary conditions. The corresponding parent reference excludes the central site: $q_{cib} = (n_{cb} - \mathbf{1}[z_{ci} = b])/(N_c - 1)$, where n_{cb} is the element count, N_c is the atom count, and z_{ci} is the central element. We define

$$D_{k,c} = \frac{1}{N_c}\sum_i \frac{1}{2}\sum_b \left| p_{cib}^{(k)} - q_{cib} \right|. \tag{7}$$

Larger values indicate stronger differences between local neighborhood compositions and the parent reference. Figure 2 reports $k = 12$.

Permutation-adjusted deviation. Finite neighborhoods can differ from the parent composition even under random site assignments. To account for this baseline, we randomly permute element labels 200 times within each configuration, preserving element counts and atomic positions:

$$\Delta D_{k,c} = D_{k,c} - \mathbb{E}_\pi\left[D_{k,c}^{\pi}\right]. \tag{8}$$

Positive values indicate greater local deviation than the permutation average; negative values indicate less. Values near zero indicate similarity to this reference, not identical environments at all sites. The scalar summaries in Figure 2 are medians across configurations.

Charge-density grid size. For a raw field with dimensions N_x, N_y, and N_z, the grid size is $N_{\text{grid}} = N_x N_y N_z$. The distribution uses 3,109 valid structure–charge pairs. Its cumulative curve gives the fraction of configurations whose fields contain no more than a given number of grid points.

B Model and Experimental Details

B.1 Gaussian Field Construction and Controls

For global prediction and magnetic-transfer pretraining, we compare electron charge density with a periodic structure-derived Gaussian field G:

$$G_c(\mathbf{r}) = \sum_{i,\mathbf{R}} Z_{ci}(2\pi\sigma_G^2)^{-3/2}\exp\left[-\frac{\|\mathbf{r} - \mathbf{r}_{ci} - \mathbf{R}\|^2}{2\sigma_G^2}\right], \tag{9}$$

where Z_{ci} is the POTCAR ZVAL, $\mathbf{R}$ runs over periodic cell translations, and $\sigma_G = 0.35$ Å. Atomic weights are deposited trilinearly onto a $56 \times 40 \times 40$ periodic grid and then Gaussian filtered.

We apply the same preprocessing form to G and electron charge density ρ. For each field source, 2,048 sampled values per training configuration define a scale a, followed by

$$u(x) = \text{sign}(x)\log(1 + |x|/a), \tag{10}$$

and standardization using the corresponding training statistics.

Because G is generated entirely from atomic structure and contains no self-consistent DFT electronic response, it serves as a structural control for ρ. Global multimodal probing compares $C+S+G$ with $C + S + \rho$, while low-label magnetic transfer compares structure–G and structure–ρ pretraining. In the downstream magnetic task, only the structure encoder is used, with neither G nor ρ provided as input.

Gaussian radial reference for local matching. For local structure–charge matching, we additionally construct a fixed non-learned Gaussian radial reference to assess whether learned representations capture information beyond simple radial field similarity.

The reference uses unit-weight atom-centered Gaussians:

$$G(\mathbf{x}) = \sum_a \exp\left(-\frac{\|\mathbf{x} - \mathbf{r}_a\|^2}{2\sigma^2}\right), \tag{11}$$

where $\sigma = 0.5$ Å. The Gaussian field and charge density ρ are summarized through normalized radial profiles within the local neighborhood, and their similarity is measured using cosine similarity.

This reference retains radial charge information while discarding learned angular representations, providing a fixed non-learned baseline for local structure–charge matching.

B.2 Modality-Specific Encoders

HEABench uses several modality-specific encoders to represent composition, atomic structure, and charge density. Global multimodal probing uses a composition MLP together with ALIGNN and MCDC for structure and spatial fields, respectively. Local structure–charge matching further considers multiple structure and charge-density encoders. All local matching branches are projected to a common 256-dimensional embedding space.

Composition and structure encoders.

- **Composition MLP.** The composition branch uses elemental identities and fractions as input to a three-layer MLP. It provides a global chemical representation without explicitly encoding atomic positions or local geometry.
- **ALIGNN.** ALIGNN represents periodic atomic structures using atomistic line graphs, incorporating both interatomic connectivity and bond-angle information. We use ALIGNN as the primary structure encoder for global multimodal probing.
- **EquiformerV2.** EquiformerV2 is an equivariant Transformer for three-dimensional atomic systems that preserves geometric information associated with rotations. It is used for local structure–charge matching and also serves as the structure backbone for low-label magnetic transfer.
- **GeoFormer.** GeoFormer is a geometry-aware Transformer encoder that learns local structural representations from atomic positions and their geometric relationships. It provides an alternative structure representation for cross-modal matching.
- **CHGNet.** CHGNet is a graph neural network designed for crystalline materials, representing atomic environments through interactions between atoms and their local neighborhoods. We use it as a crystal-graph-based structure encoder.
- **MACE.** MACE is an equivariant atomic graph model that represents local environments using higher-order geometric information. It provides another equivariant structure representation for the local matching benchmark.

Charge-density encoders.

- **MCDC.** MCDC is the spatial-field encoder used for global multimodal probing. It maps a three-dimensional charge-density or Gaussian field to a fixed-dimensional representation that can be fused with the composition and structure branches.
- **PointVector.** PointVector represents a local three-dimensional field as spatial samples with coordinates and field values and learns their local spatial relationships. Its output is projected to a 256-dimensional field embedding. The same type of field encoder is also used during magnetic-transfer pretraining.
- **X-3D.** X-3D operates directly on local three-dimensional charge-density patches and extracts a compact representation of the spatial electronic-density distribution.
- **PTv3.** PTv3 uses a Point Transformer V3 style architecture to represent local charge-density samples as a spatial point set, learning interactions among neighboring field samples.

- **LSHCE.** LSHCE is our local charge-density encoder based on spherical and radial expansions of the density field. It constructs rotation-invariant charge descriptors using angular contractions and charge-normalized Gram/power-spectrum features, together with additional local charge statistics, and projects the resulting descriptor to a 256-dimensional embedding.

Together, these encoders cover representations ranging from global chemical composition and periodic atomic graphs to dense three-dimensional electronic fields. Their use allows HEABench to compare complementary representation families rather than relying on a single encoder architecture.

B.3 Multimodal Fusion and Cross-Modal Representation Learning

HEABench uses different forms of multimodal interaction depending on the learning objective, including feature fusion, cross-modal alignment, and representation transfer.

For global multimodal probing, each available modality is first encoded and projected into a shared latent space. For $v \in \{C, S, F\}$, denoting composition, structure, and spatial field, respectively, we write

$$h_c^v = A_v(f_v(x_c^v)) \,. \tag{12}$$

The composition representation serves as the base representation, while available structure and field features are incorporated through gated residual fusion:

$$h_c = \mathrm{LayerNorm}\left(h_c^C + \sum_{v \in V_c} \alpha_c^v \odot h_c^v \right), \tag{13}$$

where $V_c \subseteq \{S, F\}$ contains the available auxiliary modalities and α_c^v is a learned elementwise gate. The fused representation is passed to target-specific prediction heads for formation energy, electronic scalars, DOS, and band targets.

For local structure–charge matching, paired structure and charge-density patches centered at the same periodic midpoint are aligned in a shared latent space. A structure patch P_i^S and its paired charge-density patch P_i^ρ are encoded as

$$h_i^S = f_S(P_i^S), \qquad h_i^\rho = f_\rho(P_i^\rho). \tag{14}$$

The 16 structure–charge encoder combinations are trained using bidirectional contrastive learning, where matched pairs are encouraged to have similar embeddings and other samples in the batch provide negatives.

For low-label magnetic transfer, cross-modal learning is used for representation initialization rather than downstream fusion. The structure encoder is first pretrained through structure–field matching and then transferred to magnetic prediction with newly initialized site- and cell-level heads. Only atomic structure is used during downstream training and inference; neither G nor ρ is provided as an additional input.

B.4 Training and Evaluation Details

Table 7 summarizes the dataset splits used for the three benchmark stages. Global prediction, local matching, and magnetic transfer are split at the parent-configuration level, with derived patches and sites inheriting their parent split. The first three rows use 3,109, 2,446, and 1,993 source configurations, while magnetic prediction uses 886 magnetically labeled configurations. NSP and SP denote non-spin-polarized and spin-polarized configurations, respectively.

For global multimodal probing, each target component is standardized using training-set statistics, and models are optimized with a masked Smooth L1 objective. Predictions are transformed back to their original physical scales for evaluation. We report MAE and RMSE as primary metrics, with validation-calibrated Gaussian negative log-likelihood (NLL) as a complementary metric for comparing modality and field conditions.

For local structure–charge matching, retrieval is evaluated using bidirectional R@1 and MRR. The retrieval gallery contains 3,904 matched pairs from 244 test parent configurations. The reference

Table 7: Task-specific data splits used in HEABench benchmarks. Counts are reported in the sample units used by each benchmark.

Benchmark	Sample unit	Train	Val	Test
Global Multimodal Probing	Configuration	2,487	311	311
Local Structure–Charge Matching	Matched patch pair	31,312	3,920	3,904
Low-Label Magnetic Transfer (structure–field pretraining)	Site-centered pair	159,500	19,900	19,900
Low-Label Magnetic Transfer (magnetic prediction)	Labeled site	62,000	13,300	13,300

element is defined as the species of the designated atom-pair endpoint (`center_element`), rather than an unordered element pair or the full local composition. Candidate controls include Random, Same-element, Same-parent, and Joint settings. The 11-way protocols use 3,750 queries, while Joint 4-way evaluation uses 2,777 queries after candidate eligibility filtering.

For low-label magnetic transfer, cross-modal pretraining uses non-spin-polarized configurations, while magnetic prediction uses magnetically labeled configurations with nested budgets of 50, 300, and 620 configurations. Checkpoints are selected based on validation site MAE. Site-level MAE is the primary metric, supplemented by site RMSE, active-site and low-moment errors, and cell-level MAE.

Statistical uncertainty is estimated using paired bootstrap resampling at the natural grouping level of each task: configurations for global prediction and parent configurations for local retrieval and magnetic transfer.

C Supplementary Results and Analyses

C.1 Global Prediction and Fusion Analyses

Target notation. The seven reported targets comprise formation energy E_{form}, DOS at the Fermi level d_F, band-crossing count N_{cross}, event count N_{event}, crossing-slope descriptor s_{cross}, the DOS vector $\mathbf{d}$, and the near-Fermi band-energy vector $\mathbf{b}$. Formation energy is reported in eV/atom, band energies in eV, and counts are dimensionless. Spectral and slope scores retain the stored target scales; no cross-target aggregate is reported.

Aggregation and support. The main-text global prediction results report MAE and RMSE in the original target scales. Table 8 summarizes the common test support and the training-target scales used for the supplementary gain analyses. For vector-valued targets, component-level and configuration-level aggregation are not identical. For the same $C + S$ DOS predictions, the component-micro MAE is 0.0980879 over 226,410 valid components, whereas the configuration-macro MAE is 0.0994986 over 311 configurations. This difference arises from aggregation weights, so the environment-conditioned analyses below therefore give configurations equal weight.

Table 8: Common test support and display scales. s_t is fitted on training targets; vector components are not independent configurations.

Target	Configurations	Valid components	s_t
E_{form}	300	300	0.284860
d_F	311	311	0.389425
N_{cross}	311	311	3.623054
N_{event}	311	311	11.861651
s_{cross}	307	307	1.047816
$\mathbf{d}$	311	226,410	0.442695
$\mathbf{b}$	311	398,080	0.045244

The DOS and band targets contain 1,001 and 1,280 components, respectively. Eleven formation-energy labels and four crossing-slope labels have non-finite references. Model comparisons use the same valid configurations and component masks.

Environment-conditioned fusion gains. To examine whether fusion gains vary with local chemical environments, we group test configurations using the permutation-adjusted local composition deviation ΔD_{12}. Fixed training-set tertiles -0.0027404040 and 0.0068907828 define the low, middle, and high groups. For configuration-level absolute errors a_c, we evaluate

$$g_c^{\rho|S} = a_c^{C+S} - a_c^{C+S+\rho}, \qquad g_c^{\rho|G} = a_c^{C+S+G} - a_c^{C+S+\rho}, \qquad g_c^{G|S} = g_c^{\rho|S} - g_c^{\rho|G}. \tag{15}$$

Positive values therefore indicate lower error for ρ relative to structure, ρ relative to the Gaussian field, and the Gaussian field relative to structure, respectively. Figure 8 reports the resulting MAE gains, with uncertainty estimated using 10,000 configuration-bootstrap replicates.

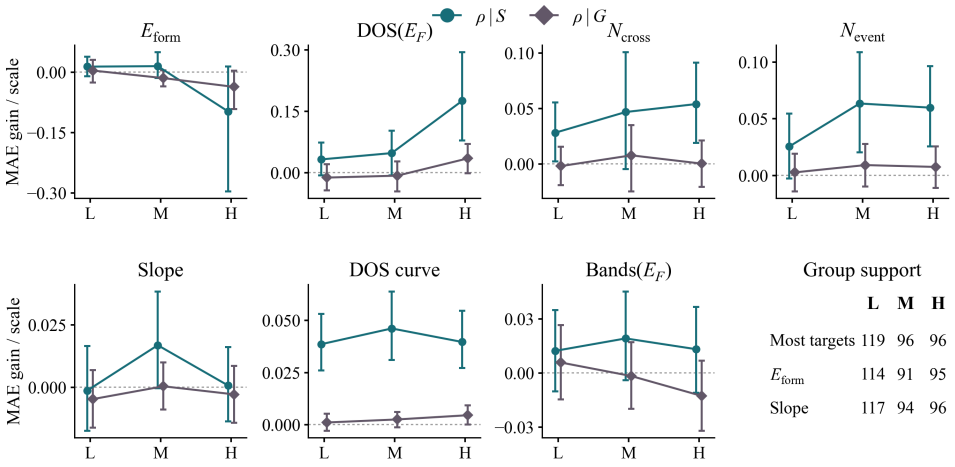


Figure 8: Environment-conditioned MAE gains: $\rho|S$ compares $C+S+\rho$ with $C+S$; $\rho|G$ compares it with $C+S+G$. Intervals are paired, pointwise 95%; gains use training-target scales and panel-specific axes. L/M/H denote low/middle/high groups; their sample counts appear at right.

The environment contrasts show no consistent cross-target trend. For d_F, the exploratory $\rho|S$ high–low contrast is 0.05584 with a pointwise interval of $[0.01372, 0.10657]$, while its Bonferroni interval $[-0.00231, 0.13583]$ crosses zero. All seven $G|S$ MAE contrasts also cross zero. Three of the 21 RMSE contrasts exclude zero pointwise, but their directions are not consistent. These results do not support a general environment-dependent fusion advantage.

Complementary scores. Figure 9 complements the point-error results with calibrated NLL and RMSE contrasts. For the DOS curve, the NLL gain for $\rho|G$ is 0.02941 with a pointwise interval of $[0.00526, 0.06167]$, consistent with the more stable improvement of charge-density fusion observed for DOS prediction. In contrast, the event-count NLL favors Gaussian fusion, -0.00673 $[-0.00874, -0.00520]$, despite lower point errors with charge-density fusion. Thus, MAE, RMSE, and calibrated NLL emphasize different aspects of prediction error and do not indicate a uniform advantage of charge density across electronic targets.

C.2 Cross-Modal Representation and Retrieval Analyses

Absolute scores and reference differences. Table 9 supplies absolute scores for the 16 learned encoding combinations and the Gaussian radial reference. The Joint columns underlie the main-text difference table: the reference scores are 77.93% R@1 and 87.48% MRR, and 13 combinations exceed both in point estimate. Differences are calculated before rounding, so subtracting the displayed absolute scores may differ by 0.01 percentage points. All learned combinations exceed the radial reference in full-gallery point estimates. These protocols use different candidate counts and query populations; cross-protocol score changes are not paired effects.

Matched candidate-size comparison. On the same 2,777 queries with nested Joint negatives, increasing from 2-way to 4-way lowers R@1 by 0.13 percentage points for EquiformerV2 with LSHCE, 14.93 for MACE with LSHCE, and 11.52 for the Gaussian radial reference (Figure 10a).

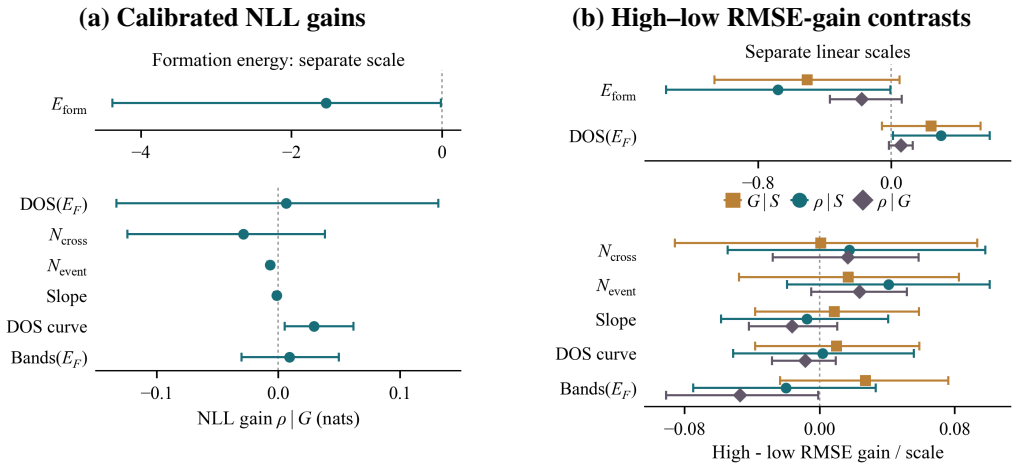


Figure 9: Complementary scores with pointwise 95% intervals. (a) Positive NLL gains favor $C + S + \rho$ over $C + S + G$. (b) Positive contrasts indicate larger gains in the high-deviation group; $\rho|S$, $\rho|G$, and $G|S$ use the comparisons defined above. RMSE contrasts use training-target scales. Separate axes retain extreme intervals.

Table 9: Absolute bidirectional R@1 and MRR (%; higher is better). 11-way Same-parent uses 3,750 queries; Joint 4-way uses 2,777; the full gallery contains 3,904 pairs. Within each protocol, all methods share queries and candidates.

		11-way: same-parent		4-way: Joint		Full gallery	
Structure	Charge density	R@1	MRR	R@1	MRR	R@1	MRR
EquiformerV2	PointVector	96.16	97.94	97.16	98.55	87.42	92.20
	X-3D	95.43	97.53	97.35	98.64	82.99	89.29
	PTv3	89.45	94.26	95.66	97.77	48.32	64.48
	LSHCE	99.47	99.72	99.75	99.87	90.82	94.61
GeoFormer	PointVector	84.07	90.73	88.73	93.94	68.83	78.93
	X-3D	77.87	86.97	84.75	91.77	54.00	68.01
	PTv3	75.29	85.80	87.85	93.60	32.56	49.71
	LSHCE	98.08	98.92	98.97	99.47	86.81	91.55
CHGNet	PointVector	75.04	85.04	83.78	91.18	61.63	74.13
	X-3D	70.57	82.34	81.65	90.01	53.07	67.58
	PTv3	68.03	80.70	81.31	89.83	51.52	66.63
	LSHCE	88.48	93.44	91.75	95.64	65.65	76.80
MACE	PointVector	66.72	79.31	78.32	87.96	50.49	65.12
	X-3D	61.19	75.61	75.24	86.22	41.05	57.33
	PTv3	60.27	74.97	74.92	86.03	41.39	57.47
	LSHCE	62.11	76.09	72.29	84.47	34.43	49.75
Gaussian radial (non-learned)		60.13	73.80	77.93	87.48	12.59	17.44

This directly tests sensitivity to additional competitors while keeping the queries and identity constraints fixed, unlike a comparison between the main 11-way and 4-way protocols.

C.3 Magnetic Transfer Analyses

Absolute performance and cell-level transfer. Table 10 summarizes site-, subgroup-, and cell-level errors across label budgets. Structure–ρ pretraining reduces site MAE and RMSE at 50, 300, and 620 labels. At 300 labels, its site-level errors are close to those of Scratch at the full 620-label budget, although this does not establish equivalence or an exact label-saving factor. Cell-level results are less consistent: ρ pretraining gives lower cell MAE at 50 and 300 labels but higher error at 620 labels.

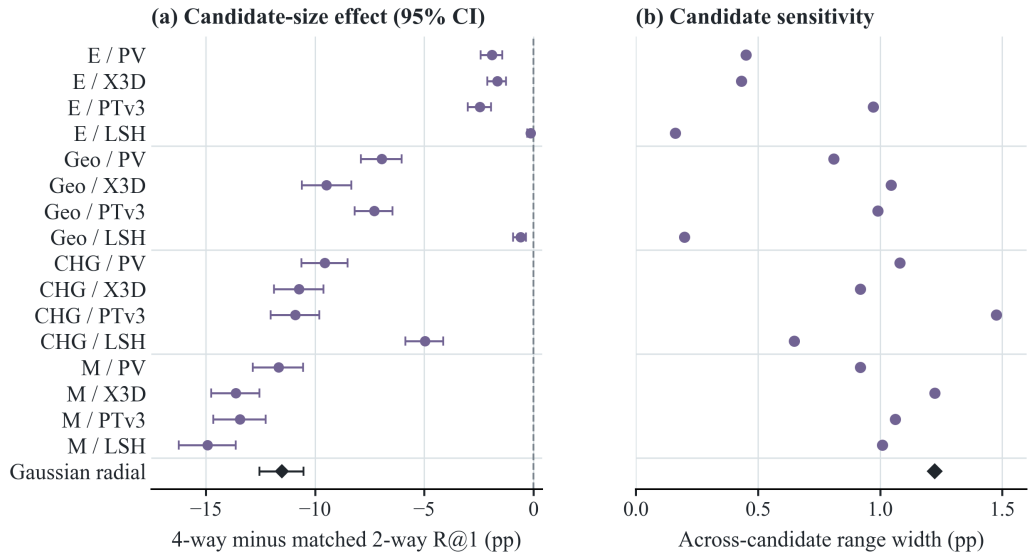


Figure 10: Joint retrieval diagnostics. (a) 4-way minus matched 2-way R@1 with paired pointwise 95% parent-bootstrap intervals. (b) R@1 range width over 21 candidate draws. E, Geo, CHG, M: structure approaches; PV, X3D, LSH: PointVector, X-3D, LSHCE. The final row is the Gaussian radial reference.

Table 10: Absolute magnetic errors. Paired entries give MAE/RMSE; site values are in μ_B and cell MAE in μ_B/cell. Dashes denote unreported metrics, not zeros.

Labels	Initialization	Site MAE/RMSE	Active MAE/RMSE	Low MAE/RMSE	Cell MAE
50	Scratch	0.2252/0.5370	0.4050/0.7583	0.0579/0.1468	7.8099
50	G pretrain	0.2253/0.4895	0.3575/0.6542	0.1021/0.2533	8.7035
50	ρ pretrain	0.2007/0.4642	0.3586/0.6543	0.0536/0.1325	7.1072
300	Scratch	0.1637/0.4172	0.2962/—	—/—	3.6222
300	ρ pretrain	0.1496/0.3957	0.2718/—	—/—	3.5028
620	Scratch	0.1502/0.4032	0.2736/—	—/—	2.6658
620	ρ pretrain	0.1411/0.3830	0.2567/—	—/—	3.0338

Figure 11 further summarizes cell-level transfer. At 50 labels, the paired cell-MAE intervals cross zero, so the site-level improvement does not establish a corresponding cell-level benefit.

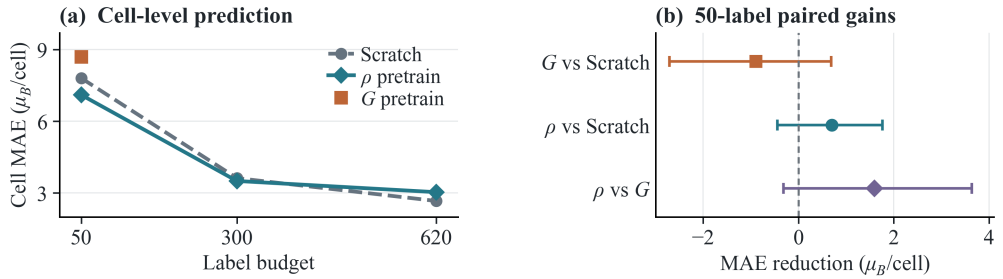


Figure 11: Cell-level transfer. (a) MAE versus label budget (configurations). (b) Paired reductions at 50 labels with pointwise 95% parent-bootstrap intervals. Positive favors the first-named condition.

Distribution of low-label benefits. Table 11 reports paired error reductions at the 50-label budget. Both G and ρ pretraining improve active-site MAE and RMSE relative to Scratch, with no resolved difference between the two pretraining fields on active sites. For low-moment sites, G increases error, whereas ρ substantially improves over G; the corresponding differences between ρ and Scratch

remain unresolved. The additional benefit of charge-density pretraining over the Gaussian control is therefore concentrated mainly in low-moment prediction.

Table 11: Paired error reductions at 50 labels with pointwise 95% parent-bootstrap intervals. Positive favors the first-named condition. Units: μ_B (sites), μ_B/cell (cell); four-decimal rounding.

Target	Metric	G vs Scratch	ρ vs Scratch	ρ vs G
All	MAE	$0.0000\ [-0.0207, +0.0171]$	$+0.0246\ [+0.0158, +0.0335]$	$+0.0246\ [+0.0102, +0.0427]$
	RMSE	$+0.0476\ [+0.0159, +0.0760]$	$+0.0728\ [+0.0508, +0.0963]$	$+0.0252\ [+0.0044, +0.0504]$
Active	MAE	$+0.0475\ [+0.0310, +0.0660]$	$+0.0463\ [+0.0289, +0.0651]$	$-0.0011\ [-0.0106, +0.0080]$
	RMSE	$+0.1041\ [+0.0751, +0.1361]$	$+0.1040\ [+0.0719, +0.1402]$	$-0.0001\ [-0.0164, +0.0161]$
Low moment	MAE	$-0.0442\ [-0.0768, -0.0177]$	$+0.0043\ [-0.0016, +0.0103]$	$+0.0485\ [+0.0228, +0.0803]$
	RMSE	$-0.1065\ [-0.1799, -0.0303]$	$+0.0143\ [-0.0030, +0.0317]$	$+0.1208\ [+0.0477, +0.1923]$
Cell	MAE	$-0.8936\ [-2.7094, +0.6864]$	$+0.7027\ [-0.4421, +1.7642]$	$+1.5963\ [-0.3168, +3.6413]$

Because overall MAE is additive across the two site groups, its gain can be decomposed as

$$\Delta_{\text{all}} = \frac{6413}{13300}\Delta_{\text{active}} + \frac{6887}{13300}\Delta_{\text{low}}. \tag{16}$$

This decomposition describes error contributions from the two site groups rather than fractions of improved sites or a physical attribution mechanism.

C.4 Robustness, Sensitivity, and Additional Diagnostics

Global prediction. For the environment-conditioned analyses, we further examine the sensitivity of the d_F $\rho|S$ signal to the reference definition and removal of its most influential configuration; its direction is retained in both checks. Adjustment for bulk-composition entropy changes some intervals, however, so these analyses do not remove compositional confounding or establish a general environment-dependent advantage. One valid formation-energy test case has a strong influence on mean gains and NLL and is retained in the primary analysis.

Cross-modal retrieval. The supplied Joint audits report zero candidate-constraint violations and zero mismatches between nested 2-way and 4-way lists. The Gaussian reference reproduces the primary candidate lists; its descriptor checks on 16 samples give rotation and permutation errors below 2×10^{-6}. These checks concern candidate construction and numerical consistency, not exclusion of every local composition, geometry, or preprocessing cue. Candidate-size and sampling diagnostics appear in Figure 10.

Magnetic transfer. The transfer comparisons evaluate different pretrained initializations rather than isolating the effects of pretraining data, optimization, or objective. Active and low-moment subgroups are defined by reference-moment magnitude, so the subgroup analyses should be interpreted as error decompositions rather than evidence for a specific magnetic mechanism.