Quality-Weighted Determinantal Point Processes for Sample Efficient DFT Surrogate Models

Shirley X. Yu [1] **Adji B. Dieng** [1]

Abstract

Density functional theory (DFT) is widely used to compute electronic structure and materials properties, but it is expensive to run. When compute is limited, practitioners must choose which candidate systems to evaluate. We study this selection problem on JARVIS-DFT, a large standardized dataset of DFT labeled crystal structures and a realistic testbed for training machine learning models to predict expensive DFT computed properties. We propose quality-weighted determinantal point process (QW-DPP), a label-free coreset selection method for training DFT surrogate models. QW-DPP selects candidates that are both diverse through an RBF kernel and representative through a density-based quality score from unlabeled composition fingerprints. On JARVIS-DFT, QW-DPP reduces test MAE compared with random sampling by up to 8.9% at a 1% labeling budget across formation energy and bulk modulus. This improvement holds across Ridge, Bayesian Ridge, and Gradient Boosting. We find that QW-DPP is most useful at very small budgets because it improves worst case coverage of feature space.

1. Introduction

Density functional theory (DFT) is one of the most prevalent first-principles methods in computational condensed matter physics, materials science, and chemistry, providing access to ground state energies, electronic structure, and mechanical response. Yet, a single DFT calculation can take hours to days of compute time (Choudhary et al., 2020), and higher-level methods such as hybrid functionals or GW calculations can cost orders of magnitude more (Choudhary et al., 2020; Burke, 2012; Hedin, 1965). To reduce this cost, researchers increasingly train machine learning (ML) models to predict DFT derived properties without running a new DFT calculation for every candidate (Schmidt et al., 2019; Butler et al., 2018). These models, often called surrogate models, can accelerate property prediction and screening across condensed matter, electronic materials, and energy applications. However, their training data remains a bottleneck because every label requires a DFT calculation.

Since every label is expensive, limited compute turns surrogate training into a candidate selection problem. This setting is especially natural in materials science, where researchers often face large unlabeled candidate pools (Jain et al., 2013; Curtarolo et al., 2012) and well defined property prediction tasks tied to clean energy, catalysis, and electronic devices. Given a candidate pool of 10^4 to 10^6 unlabeled crystal structures and a DFT budget of k labels, which structures should be selected for calculation? This problem is especially important for novel material families with no existing database, higher-level DFT methods where each calculation is prohibitively expensive, and active learning campaigns that require an initial labeled set. In each case, the practitioner must select candidates using only unlabeled features, before any new DFT calculations are run.

Existing data selection methods do not address this regime cleanly. Active learning approaches such as uncertainty sampling (Lewis & Gale, 1994; Settles, 2009) and Bayesian optimization (Shahriari et al., 2015; Lookman et al., 2019), widely used in materials science, require an initial labeled set to bootstrap the loop. Modern coreset selection methods such as gradient matching (Mirzasoleiman et al., 2020) or validation-loss-driven selection (Killamsetty et al., 2021) require either trained models or labels. Methods that can run on fingerprints alone, including random sampling, k-Means clustering (Lloyd, 1982), and farthest point sampling (FPS) (Eldar et al., 1997), do not explicitly balance diversity against representativeness in chemical space. As a result, they either select uniformly or systematically over sample outlier structures.

We propose QW-DPP, a quality-weighted determinantal point process (DPP) that selects coresets for DFT surrogate training without consuming any labels at selection time. QW-DPP factors the DPP L-ensemble as $L =$

[1]Department of Computer Science, Princeton University, Princeton, USA. Correspondence to: Shirley X. Yu <shirley.yu@princeton.edu>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

$\mathrm{diag}(q)\, K\, \mathrm{diag}(q)$, combining diversity from an RBF kernel K with representativeness from a density-based quality-weight q computed from unlabeled fingerprints. On JARVIS-DFT, QW-DPP improves test MAE over random sampling by up to 8.9% at the 1% budget on formation energy and bulk modulus, across the downstream models Ridge, Bayesian Ridge, and Gradient Boosting. We observe a sharp crossover near the 5% budget, beyond which the methods converge. While we evaluate on DFT surrogate training, the framework applies more broadly to label-free coreset selection from candidate pools where each label requires an expensive simulation or experiment.

Our contribution is twofold. First, we introduce QW-DPP, a label-free coreset selection method for allocating limited DFT compute budgets. QW-DPP instantiates a quality-weighted DPP using a density-based quality function over materials chemical space, and we evaluate it across three target properties and three downstream model classes. Second, we characterize when QW-DPP helps. Its gains come from improved worst case feature space coverage, at the cost of a small target distribution distortion on the most predictable property. Combining feature target predictability and budget size identifies the regime in which QW-DPP is most useful, using unlabeled pool diagnostics together with a small labeled seed set, before committing the full DFT budget. Beyond materials science, this addresses a broader question in data scarce scientific ML: when does diversity-aware data selection improve surrogate models, and how can practitioners decide before spending the computational budget?

2. Related Work

2.1. Determinantal point processes for subset selection

Determinantal point processes (DPPs) define probability distributions over subsets that favor diversity (Kulesza & Taskar, 2012). Given a similarity matrix over candidate points, a DPP assigns higher probability to subsets whose elements are mutually dissimilar: the determinant of the selected submatrix is large when the chosen points span a large volume in feature space. This makes DPPs a natural tool for coreset selection, where the goal is to cover a candidate pool rather than select many near duplicate examples.

For DFT coreset selection, however, diversity alone is not sufficient. A diversity-only selector can over sample isolated structures while under sampling dense regions that represent much of the candidate pool. The standard DPP L-ensemble factorization provides a way to combine pairwise diversity with per point weights with $L = \mathrm{diag}(q)\, K\, \mathrm{diag}(q)$, where K captures pairwise similarity and q assigns each point an individual weight (Kulesza & Taskar, 2012). The k-DPP variant conditions the distribution on a fixed subset size, matching the case where a practitioner has a budget of exactly k DFT labels (Kulesza & Taskar, 2011). By itself, however, a k-DPP only enforces the budget; with $q \equiv 1$, it remains a diversity-only selector and can still concentrate on outliers.

Prior work has used the weight q as a learned or hand crafted relevance score in retrieval and summarization. Related spectral approaches, such as the Vendi Score (Friedman & Dieng, 2023) and quality-weighted Vendi Scores (Nguyen & Dieng, 2024), use eigenvalues of similarity matrices to quantify diversity and quality-weighted diversity. QW-DPP differs from this line of work in the selection regime. Vendi style methods provide diversity metrics or guidance for search, while our setting requires choosing exactly k DFT calculations in a single shot before any labels or trained models are available. We therefore define q as a density-based representativeness score computed only from unlabeled fingerprints and use the resulting DPP objective directly for fixed-budget coreset selection. This preserves broad feature space coverage while avoiding the outlier heavy behavior of diversity-only selection.

2.2. Label-using data selection

Many modern data selection methods rely on labels, trained models, or training dynamics. CRAIG (Mirzasoleiman et al., 2020) selects subsets that match training gradients, GLISTER (Killamsetty et al., 2021) selects subsets that minimize validation loss through bilevel optimization, and forgetting events (Toneva et al., 2019) identify informative examples from changes in training behavior. Active learning methods such as BALD (Gal et al., 2017) and BatchBALD (Kirsch et al., 2019) iteratively query examples with high model uncertainty.

In materials science, Bayesian optimization and adaptive sampling are the standard active learning paradigm (Lookman et al., 2019; Tran & Ulissi, 2018; Ling et al., 2017; Balachandran et al., 2016). These methods are powerful once a labeled seed set exists, but they do not solve the initial allocation problem studied here. At the cold start stage, the practitioner has a candidate pool and a DFT budget, but no labels, trained surrogate, validation set, or uncertainty estimates. QW-DPP is designed for this stage: it selects the initial coreset before any DFT calculation is run.

2.3. Label-free baselines

The methods most directly comparable to QW-DPP are label-free selectors that operate only on fingerprints. Random sampling is the de facto baseline in materials ML benchmarks (Dunn et al., 2020). k-Means clustering on features (Lloyd, 1982) returns the train pool point nearest each centroid, favoring cluster centers. FPS (Eldar et al., 1997) greedily appends the point farthest from the current se-

lected set, emphasizing extreme geometric spread. Standard DPPs also operate label-free, but reduce to diversity-only selection when $q \equiv 1$ in the L-ensemble factorization.

These baselines represent the main alternatives available before any DFT labels are computed. They differ in how they trade off coverage and representativeness: random sampling is unbiased but not coverage seeking, k-Means favors dense regions but may miss boundary cases, and FPS and standard DPPs can over emphasize isolated points. QW-DPP is designed to occupy the middle ground, combining DPP diversity with a density-based weight that keeps selection tied to representative regions of the unlabeled pool. We compare against all four baselines in §5.

3. Preliminaries

3.1. Problem formulation

Given N unlabeled crystal structures with fingerprints $\{x_i \in \mathbb{R}^d\}_{i=1}^N$ and a DFT budget of $k \ll N$ labels, we select a subset $\mathcal{S} \subset [N]$ of size k that minimizes the test MAE of a downstream regressor trained on $\{(x_i, y_i)\}_{i \in \mathcal{S}}$. The selector is label-free. It sees only the fingerprints, never the targets y_i, since the targets are precisely what the DFT budget would compute.

3.2. Determinantal point processes

A DPP over a ground set $[N]$ with positive semidefinite L-ensemble $L \in \mathbb{R}^{N \times N}$ assigns each subset $\mathcal{S} \subset [N]$ probability $\mathcal{P}(\mathcal{S}) \propto \det(L_{\mathcal{S}})$, where $L_{\mathcal{S}}$ is the principal submatrix indexed by $\mathcal{S}$ (Kulesza & Taskar, 2012). The k-DPP variant (Kulesza & Taskar, 2011) conditions this distribution on $|\mathcal{S}| = k$ and samples stochastically. Following Kulesza & Taskar (2012), an L-ensemble can be factored as

$$L \;=\; \mathrm{diag}(q)\, K\, \mathrm{diag}(q), \qquad L_{ij} \;=\; q_i\, K_{ij}\, q_j, \tag{1}$$

where $K \in \mathbb{R}^{N \times N}$ is a similarity kernel and $q \in (0, 1]^N$ is a per point quality-weight. The eigenvalues of $L_{\mathcal{S}}$ encode both how diverse $\mathcal{S}$ is (through K) and how representative its members are (through q). Setting $q \equiv 1$ recovers the standard diversity-only DPP. Prior work has typically used learned or hand crafted quality scores in retrieval and summarization. We instantiate q as a label-free, density-based statistic and treat the DFT budget k as a hard constraint, deterministically maximizing $\log\det(L_{\mathcal{S}})$ through greedy MAP selection.

4. Method

QW-DPP constructs the L-ensemble in Eq. (1) from a similarity kernel K and a density based quality-weight q, both computed from the unlabeled fingerprints. We then select the coreset by greedy log-determinant maximization.

4.1. Similarity kernel

We ℓ_2-normalize the fingerprints, $\tilde{x}_i = x_i / \|x_i\|_2$, so that the RBF kernel becomes a monotone function of cosine similarity. The bandwidth γ is set by the median heuristic on a 2,000 point seeded subsample, $\gamma = 1/(2\,\mathrm{med}^2)$, with med the median nonzero pairwise distance on $\tilde{x}$. The kernel is $K_{ij} = \exp\!\big(-\gamma\, \|\tilde{x}_i - \tilde{x}_j\|^2\big)$ and is stored as float32.

4.2. Density-based quality

For each point i we compute the mean Euclidean distance $\bar{d}_i$ to its κ nearest neighbors (NNs) in $\tilde{x}$ (excluding self), with $\kappa = 10$, giving a local density $\rho_i = 1/(\bar{d}_i + \varepsilon)$ with $\varepsilon = 10^{-8}$. We rank normalize and soften:

$$q_i \;=\; \sqrt{\mathrm{rank}(\rho_i)\,/\,M} \;\in\; (0, 1], \tag{2}$$

where M is the size of the subsample on which selection is performed (defined in §4.3). Rank normalization is robust to the skewed density tails common in materials databases, and the square root prevents the dense cluster mode from dominating, keeping moderate density points eligible for selection. Crucially, q depends only on $\{\tilde{x}_i\}$, never on y_i.

4.3. Greedy selection via incremental Cholesky

Algorithm 1 QW-DPP Coreset Selection (Label-Free)

Input : Fingerprints $\{x_i\}_{i=1}^N$, budget k, neighborhood size κ, subsample cap $M_{\max}$
Output : Coreset indices $\mathcal{S} \subset [N]$ with $|\mathcal{S}| = k$

ℓ_2-normalize fingerprints $\tilde{x}_i \leftarrow x_i / \|x_i\|_2$
Draw seeded subsample $\mathcal{I} \subseteq [N]$ with $|\mathcal{I}| = M = \min(M_{\max}, N)$
Set bandwidth γ via median heuristic on a 2,000 point subsample
Compute RBF kernel $K_{ij} \leftarrow \exp(-\gamma\, \|\tilde{x}_i - \tilde{x}_j\|^2)$ for $i, j \in \mathcal{I}$
Compute density $\rho_i \leftarrow 1/(\bar{d}_i + \varepsilon)$ from mean distance $\bar{d}_i$ to κ NNs, and quality $q_i \leftarrow \sqrt{\mathrm{rank}(\rho_i)/M}$
Form L-ensemble $L_{ij} \leftarrow q_i\, K_{ij}\, q_j$
Initialize residuals $r_i \leftarrow L_{ii}$, projection buffer C, and $\mathcal{S} \leftarrow \emptyset$
for $s = 1, \ldots, k$ **do**
  $j^\star \leftarrow \arg\max_{i \in \mathcal{I} \setminus \mathcal{S}} r_i$
  if $\sqrt{r_{j^\star}} < 10^{-10}$ **then**
    break
  Update projection column $C[\,\cdot\,, s]$ via incremental Cholesky against $j^\star$
  Decrement residuals $r_i \leftarrow \max(0,\, r_i - C[i, s]^2)$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{j^\star\}$
return $\mathcal{S}$ *mapped back to original pool indices via* $\mathcal{I}$

Exact DPP MAP is NP-hard. Algorithm 1 ties the kernel, quality, and selection components into a single label-free pipeline. We adopt the incremental Cholesky greedy of Chen et al. (2018), which avoids recomputing the marginal log-determinant from scratch at each step by maintaining a running Cholesky factorization and updating the conditional variance of every remaining candidate incrementally, where M is the size of the subsample on which the kernel is built. While $\mathcal{S} \mapsto \log\det(L_{\mathcal{S}})$ is submodular, it is not monotone, so the classic $(1 - 1/e)$ guarantee for monotone submodular maximization (Nemhauser et al., 1978) does not transfer directly; the greedy selector is a standard and empirically strong heuristic for this NP-hard objective rather than an approximation algorithm with a matching bound. The procedure maintains residual conditional variances $r_i = L_{ii} - \|\mathrm{proj}_i\|^2$ and an $M \times k$ projection buffer; at step s, it picks $j^\star = \arg\max_i r_i$, updates the projection column, and decrements $r_i \leftarrow \max\big(0,\, r_i - \mathrm{proj}[i,s]^2\big)$. The procedure runs in $\mathcal{O}(M\,k^2)$ time and $\mathcal{O}(M\,k)$ extra memory beyond storing K, and stops at budget exhaustion or numerical singularity ($\sqrt{r_{j^\star}} < 10^{-10}$). Because the algorithm is incremental, coresets at smaller budgets are prefixes of the greedy order: one run produces all of them.

A dense K on the full 60,795 point pool would exceed 14 GB at float32, so we draw a single seeded subsample of $M = \min(M_{\max},\, n_{\text{train}})$ points with $M_{\max} = 35{,}000$ before kernel construction and run the greedy selector on that subsample. Coreset indices are mapped back to the original pool through a stored subsample to pool index map. At $M = 35{,}000$ the kernel occupies ≈ 4.9 GB and a full 50% budget run completes on a single CPU node.

Remark 4.1 (Quality weighting as an additive representativeness prior). Because $L_{\mathcal{S}} = \mathrm{diag}(q_{\mathcal{S}})\, K_{\mathcal{S}}\, \mathrm{diag}(q_{\mathcal{S}})$, the greedy objective decomposes exactly as

$$\log\det(L_{\mathcal{S}}) \;=\; \log\det(K_{\mathcal{S}}) \;+\; 2\sum_{i\in\mathcal{S}} \log q_i. \qquad (3)$$

The first term is the diversity (covered-volume) term of a standard DPP, while the second is a separable, additive penalty $2\log q_i$ that discourages selecting low-density (outlier) points. Quality weighting thus does not distort the diversity objective so much as place a per point representativeness prior on top of it, which is exactly the behavior we want for cold start selection: spread coverage while staying on well populated regions of chemical space. This additive structure is specific to the log-determinant objective; entropy-based selectors such as Vendi style MAP admit no analogous clean decomposition, and we show in Appendix F that pairing quality weighting with those objectives can instead hurt.

4.4. No DFT computation at selection time

The full pipeline is structures $\rightarrow$ composition $\rightarrow$ Magpie features $\rightarrow$ (ρ, q) $\rightarrow$ (K, L) $\rightarrow$ greedy $\rightarrow$ S. DFT is computed only on the resulting coreset S. No y_i value ever enters the selector. This is what makes QW-DPP a DFT budget allocator; the practitioner can run selection on a 10^5 entry candidate database and only then spend the DFT budget on the chosen k structures, rather than labeling everything and subsampling post hoc.

5. Experiments

5.1. Setup

We use the JARVIS-DFT 3D database from jarvis.db.figshare. Each structure is featurized into a 132 dimensional Magpie composition descriptor (Ward et al., 2016) from matminer's ElementProperty featurizer, computed from the pymatgen Composition. No explicit structural features are used. We use a single fixed train/test split (80/20, seed 42), saved once and shared across all methods, budgets, and downstream models. Three target properties are evaluated: formation energy per atom (eV/atom, $N_{\text{valid}} = 60{,}795$), bulk modulus (GPa, $N_{\text{valid}} = 18{,}610$), and bandgap (eV, $N_{\text{valid}} = 60{,}795$, used as a negative case because Magpie features predict it only weakly, cross validated Ridge $R^2 \approx 0.39$).

Selected coresets train three downstream regressors: (i) **Ridge regression**, with α chosen from a 20 point log spaced grid in $[10^{-3}, 10^{3}]$ through cross validation; (ii) **Bayesian Ridge** (sklearn.linear_model.BayesianRidge with max_iter=300 and tol=1e-3); and (iii) **Gradient Boosting** (GradientBoostingRegressor with n_estimators=200, max_depth=5, learning_rate=0.1, and random_state=42), our main shallow nonlinear model. We additionally evaluate **ALIGNN** (Choudhary & DeCost, 2021), a structure aware line graph GNN, as a cross architecture transfer check in §5.4; the full ALIGNN configuration is provided in Appendix A.

All baselines use the same ℓ_2-normalized fingerprints and return coresets of identical size. **Random** draws uniformly without replacement, with 10 independent seeds (0–9) at every budget. **k-Means medoids** fits sklearn.cluster.KMeans (Lloyd, 1982) with k equal to the budget size and returns the train pool point nearest each centroid. **FPS** (Eldar et al., 1997) starts from a seeded initial point and greedily appends $\arg\max_i \min_{j\in S} \|\tilde{x}_i - \tilde{x}_j\|$. **Standard DPP** runs the same greedy log-determinant selection as QW-DPP but with $q \equiv 1$ in Eq. (1), isolating the contribution of the density based quality term.

Two independent sources of randomness affect our results. First, coreset variability across selectors: Random sampling

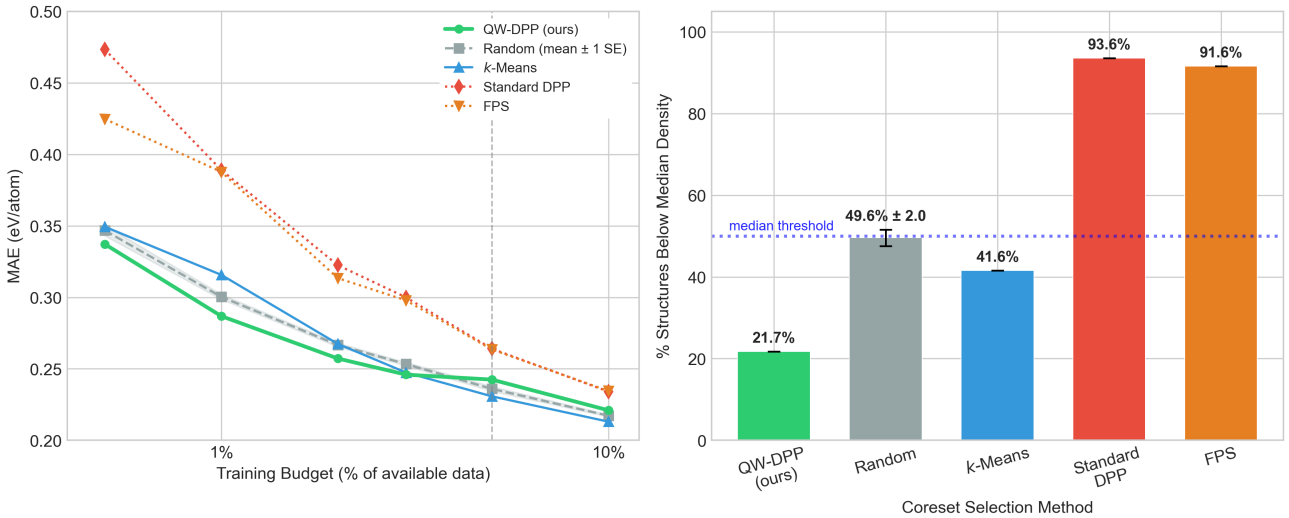

Figure 1. **QW-DPP outperforms baselines at low DFT budgets.** (left) Test MAE of Gradient Boosting on JARVIS formation energy versus training budget (% of available data). (right) Outlier selection rate at the 1% budget. Random is mean ± one standard error over 10 seeds; deterministic methods are single runs.

Table 1. **QW-DPP achieves the lowest test MAE at low budgets on JARVIS formation energy with Gradient Boosting.** Random is mean ± one standard error over 10 seeds; deterministic methods are single runs. $*$: $p < 0.05$ vs. Random (t-test). Best per column in **bold green**, second best in blue.

Method	0.5%	1%	2%	3%	5%	10%
QW-DPP (ours)	**0.337***	**0.287***	**0.257***	**0.246***	0.242*	0.221*
Random	0.347 ± 0.004	0.300 ± 0.002	0.267 ± 0.001	0.253 ± 0.001	0.236 ± 0.002	0.217 ± 0.001
k-Means	0.349	0.316*	0.267	0.247*	**0.231***	**0.213***
Standard DPP	0.473*	0.389*	0.323*	0.300*	0.264*	0.234*
FPS	0.425*	0.388*	0.313*	0.298*	0.264*	0.234*

produces a different coreset per seed (10 seeds, 0–9) and is our only stochastic selector. QW-DPP, k-Means, FPS, and Standard DPP are deterministic and produce a single coreset per (method, budget) setting. Second, training randomness: ALIGNN is stochastic with random initialization, batch order, CUDA non determinism. The other three regressors are deterministic where Ridge and Bayesian Ridge are closed form, and Gradient Boosting is deterministic given random_state=42. Because ALIGNN is a secondary cross architecture check rather than a primary regressor, and because per-coreset GPU training is expensive, we run a single ALIGNN training per coreset and do not average over training seeds.

Consequently, for Ridge, Bayesian Ridge, and Gradient Boosting, the only reported error bars come from the 10 Random seeds; the deterministic selectors are point estimates, and we test each one against the Random distribution. ALIGNN results carry both sources of variance combined into a single number per cell, which is why we treat ALIGNN as a robustness check rather than a primary statistical claim.

We report MAE on the held-out test set in natural property units; the headline statistic is percent MAE improvement over Random at the 1% budget. We test each deterministic method against the Random distribution at every budget by a one sample two sided t-test. Stars therefore indicate a difference from the random seed distribution under the fixed split, not uncertainty over dataset splits or over the deterministic selectors themselves, and we apply no multiple-comparison correction. Budgets sweep $\{0.5, 1, 2, 3, 5, 10, 25, 50\}\%$ of the train pool; the paper focuses on the $\leq 10\%$ regime where coreset choice matters.

5.2. Headline result on JARVIS formation energy

We anchor the analysis on Gradient Boosting on formation energy. Gradient Boosting is fast enough to permit 10 seed statistical testing at every setting while still capturing nonlinear structure, and formation energy is the most widely benchmarked DFT property with full coverage in JARVIS.

Figure 1 (left) and Table 1 report test MAE versus budget. QW-DPP attains the lowest MAE at every budget $< 5\%$,

beating Random by ≈4.5% at 1% and ≈2.8% at 3%. The gap is widest at the smallest budgets, where the choice of which DFT calculations to run matters most. A crossover occurs near 5%: above that budget, k-Means and Random catch up and ultimately overtake QW-DPP. Standard DPP and FPS lag at every budget, suggesting that diversity-only objectives are not just suboptimal but actively harmful in this setting.

Figure 1 (right) shows a slice at the 1% budget. Only 21.7% of QW-DPP's picks fall below the median feature density, compared with 93.6% for Standard DPP, 91.6% for FPS, and 48.5% for Random. Diversity-only methods collapse onto the manifold's outliers; the quality-weighting of QW-DPP in $L = \mathrm{diag}(q)\, K\, \mathrm{diag}(q)$ pulls selection back onto well populated regions while preserving the spread that the DPP contributes. This is the additive penalty of Eq. (3) at work: with $q \equiv 1$ (Standard DPP) the $2\sum_i \log q_i$ term vanishes and nothing offsets the diversity term's pull toward isolated points, whereas QW-DPP assigns small q_i to low-density structures and so penalizes exactly those picks.

5.3. Mechanism: worst case coverage at the cost of distribution match

Why does QW-DPP win at low budgets? Two properties of the coreset explain the gain: a feature-space coverage property that is fully label-free, and a target distribution property that needs only a small seed set.

Worst case feature space coverage. Figure 2 reports the maximum NN distance from a test point to the coreset at the 1% budget. This worst case metric measures whether any test structure is left stranded. QW-DPP achieves the smallest worst case distance on both properties, shown by the leftmost green line. The quality-weighted L-ensemble fills feature space more uniformly than any baseline by combining diversity from K with representativeness from q: the worst case test-to-coreset distance is reduced, and fewer coreset points are spent on low density regions.

Target distribution match. QW-DPP does not match the test target distribution as closely as Random. Figure 3 compares marginal target distributions on formation energy at the 1% budget: Random tracks the test marginal closely because uniform sampling is unbiased in expectation, while QW-DPP shows mild distortion from down weighting outlier structures. The gap is small in absolute terms, but it accumulates as budget grows, eventually contributing to the crossover at ∼5%.

These two findings tell a coherent story. QW-DPP wins at low budgets because it dramatically improves worst case coverage at the cost of a small amount of target distribution match. At small budgets, where coverage is the binding constraint, this trade pays off. As budget grows, coverage saturates across all methods while the target match gap remains; Random catches up and eventually overtakes QW-DPP. This is exactly the crossover seen in Figure 1.

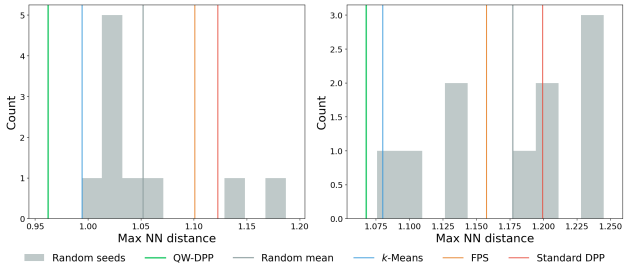

Figure 2. **QW-DPP achieves the smallest worst case feature space distance.** Maximum NN distance from any test point at the 1% budget, $\max_{x \in \text{test}} \min_{c \in \text{coreset}} \|x - c\|_2$, in ℓ_2-normalized Magpie feature space (lower is better). Histogram: 10 random seed coresets. Vertical lines: deterministic selectors. (left) Formation energy. (right) Bulk modulus.

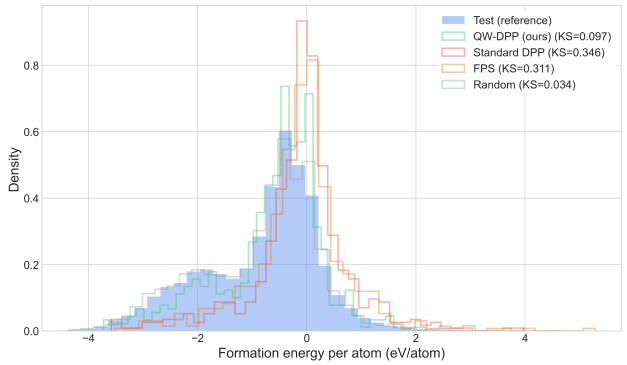

Figure 3. **QW-DPP trades distribution match for coverage.** Coreset target densities overlaid on the test reference at the 1% budget on formation energy, scored by Kolmogorov–Smirnov (lower is better). Random (KS = 0.034) closely matches the test marginal; QW-DPP (KS = 0.097) lags on distribution match but wins on coverage (Figure 2).

The crossover through the lens of Eq. (3). The decomposition also explains why the advantage is transient. At a small budget most of the kernel mass is still uncovered, so the diversity term $\log\det(K_S)$ has large marginal gains and the additive penalty $2\log q_i$ meaningfully reshapes which points are chosen, steering selection off the outliers a bare DPP would take. As the budget grows, feature space fills in and the marginal diversity gain of each new point shrinks, so the same fixed penalty changes the selection order less; we conjecture that QW-DPP and a bare DPP increasingly agree, though we do not measure this directly. The crossover near 5% marks the budget at which this front loaded coverage benefit has largely been realized while the target distribution cost of the q-penalty continues to accumulate, so an unbiased Random selector catches up.

5.4. Generalization across model, property, and architecture

The §5.2 result holds Gradient Boosting and formation energy fixed. We now vary each axis independently to test whether QW-DPP's gain depends on either choice.

Model axis. Holding formation energy fixed, we replace Gradient Boosting with Ridge. Figure 4 (left) shows the Ridge curve. QW-DPP improves over Random by $+3.0\%$ at the 1% budget in this model as well, confirming the win is not specific to gradient boosted ensembles.

Property axis. Holding Gradient Boosting fixed, we replace formation energy with bulk modulus, a mechanical property with a smaller candidate pool (18.6K vs. 60.8K) and a more clustered density distribution (CV $= 8.5$ vs. 6.3). Figure 4 (right) shows the curve; QW-DPP improves over Random by $+8.9\%$ at the 1% budget, the largest gain in the paper. This matches the prediction in §5.5: quality weighted selection helps most when the density distribution is heavy tailed, since that is where outlier down weighting matters most.

Across the four (model, property) combinations in Table 2 at the 1% budget, MAE improvement ranges from $+3.0\%$ to $+8.9\%$. The win transfers across both axes.

Negative case (bandgap). Figure 5 plots the coreset target variance for each selector against the test set variance for the three properties. On formation energy and bulk modulus, QW-DPP's coreset variance (green dot) is close to the test reference (dotted line), consistent with the test MAE gains above. Bandgap is the negative case. QW-DPP's coreset variance lies substantially above the test reference, and on bandgap QW-DPP does not improve test MAE over Random at any budget. The signature of failure is visible early, in the coreset target variance, from a small seed set and before committing the full DFT budget. This is a scope condition, not a failure. QW-DPP covers feature space; the regressor cares about label space. When features predict the target well (formation energy, $R^2 \approx 0.78$), the two coincide and QW-DPP helps. As an empirical heuristic from these three properties, QW-DPP tends to help when fingerprints predict the target reasonably well ($R^2 \gtrsim 0.4$) and is no better than Random when they do not.

The scope condition has a physical reading. Magpie fingerprints encode composition and elemental statistics, which largely determine extensive thermodynamic quantities: formation energy is set mostly by which elements are present and in what proportion, so composition coverage transfers to label coverage. Bandgap, by contrast, is an electronic-structure property governed by crystal symmetry, orbital hybridization, and band dispersion in k-space, which compounds of identical composition but different structure can realize very differently. Composition fingerprints cannot resolve the variation that drives the bandgap, the feature objective QW-DPP optimizes decouples from label coverage, and selection reduces to noise. The $R^2 \gtrsim 0.4$ rule is the empirical shadow of this distinction between composition-determined and structure-determined properties.

Architecture axis (one coreset transfers across model classes). Figure 6 reruns the headline benchmark with three downstream regressors and ALIGNN at the 1% budget on JARVIS formation energy, using a single QW-DPP coreset selected on Magpie fingerprints. The pattern from §5.2 carries across model classes: QW-DPP beats Random on Gradient Boosting, Ridge, and Bayesian Ridge ($\Delta = -0.0135$, -0.0114, -0.0059 MAE; ≥ 3.8 SE in each case). As a quick check that the Magpie-selected coreset does not hurt a structure aware GNN, we also train ALIGNN on the same coreset and find QW-DPP matches Random within seed variance ($+0.0008$ MAE, 0.18 SE).

5.5. When does QW-DPP help?

§5.2–5.4 report QW-DPP wins on formation energy and bulk modulus and a clean failure on bandgap. We now examine what distinguishes the cases where QW-DPP helps from the case where it does not, using two cheap diagnostics: a feature–target predictability score that needs only a small seed set, and a label-free measure of pool density structure.

Cross validated Ridge R^2 on the full labeled pool quantifies how well Magpie features predict each target before any selection is applied: $R^2 = 0.78$ (formation energy), 0.48 (bulk modulus), 0.39 (bandgap). Across our three properties, this ordering matches the ordering of the QW-DPP gain: largest on formation energy, intermediate on bulk modulus, no gain on bandgap. QW-DPP optimizes a feature space objective. When features predict the target well, that objective transfers to label coverage and the win materializes. When features predict the target poorly, feature coverage and label coverage decouple, and QW-DPP's selection criterion no longer matches what the regressor cares about.

A second feature side property tracks where QW-DPP's gain is largest within the predictable regime: how clumpy the pool is. The density coefficient of variation is 6.3 on the formation energy pool versus 8.5 on the bulk modulus pool. The relative QW-DPP gain at the 1% budget (Table 2) is $+4.5\%$ on formation energy / Gradient Boosting but $+8.9\%$ on bulk modulus / Gradient Boosting, in the direction predicted by the cluster structure: quality-weighted selection helps most when the density distribution is heavy tailed.

Putting these together (Table 3), QW-DPP helps when two conditions both hold: the fingerprints predict the target reasonably well ($R^2 \gtrsim 0.4$ in our experiments), and the budget is small ($< 5\%$ of the candidate pool). If either condition fails, the gain disappears and Random is competitive or better. Both quantities can be estimated from a few hun-

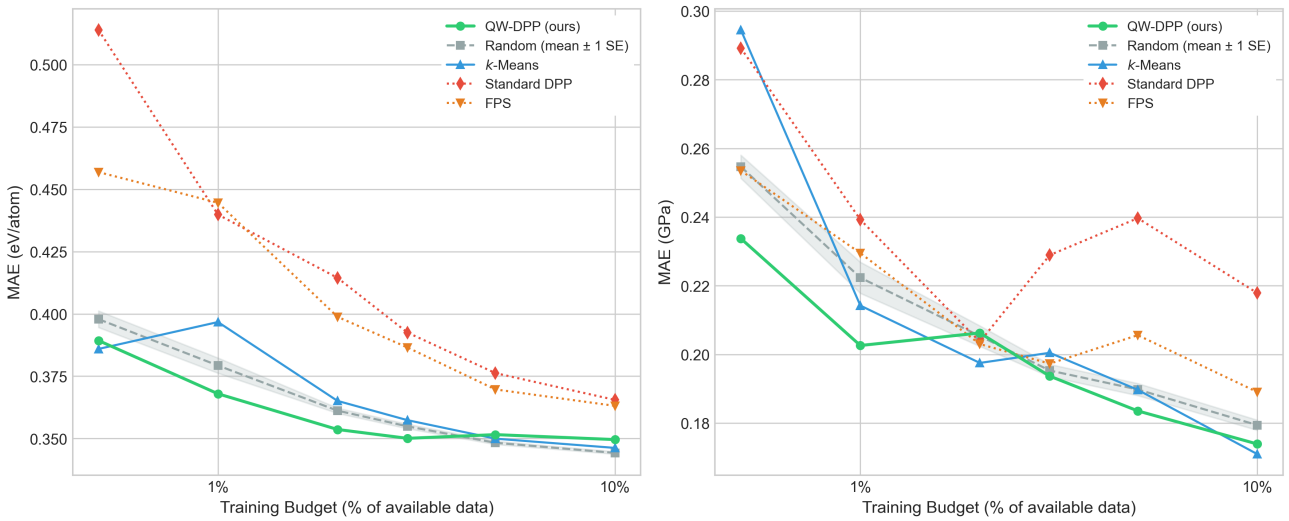

Figure 4. **Generalization across model class and target.** Test MAE versus training budget. (left) Ridge on JARVIS formation energy (model swap). (right) Gradient Boosting on bulk modulus (target swap). Random is mean ± one standard error over 10 seeds; deterministic methods are single runs.

Table 2. **QW-DPP improves over Random across targets and models at the 1% budget.** Test MAE for each (target, model) pair on JARVIS; the final column shows QW-DPP's percentage improvement over Random. Random is mean ± one standard error over 10 seeds; deterministic methods are single runs. $*$: $p < 0.05$ vs. Random (t-test). Best per row in **bold green**, second best in blue.

Target	Model	QW-DPP (ours)	Random	k-Means	Standard DPP	FPS	MAE reduction vs Random (%)
Formation Energy	Ridge	**0.368***	0.379 ± 0.003	0.397*	0.440*	0.445*	+3.0%
Formation Energy	GBR	**0.287***	0.300 ± 0.002	0.316*	0.389*	0.388*	+4.5%
Formation Energy	Bayesian Ridge	**0.371***	0.377 ± 0.002	0.392*	0.443*	0.447*	+1.6%
Bulk Modulus	Ridge	**0.220***	0.230 ± 0.004	0.231	0.250*	0.246*	+4.2%
Bulk Modulus	GBR	**0.203***	0.222 ± 0.005	0.214	0.239*	0.230	+8.9%

Table 3. **QW-DPP decision rule.** Whether to allocate the DFT budget with QW-DPP or with Random, as a function of feature–target predictability (cross-validated Ridge R^2 on a seed set) and budget size (fraction of the candidate pool). QW-DPP helps only when both conditions are favorable; otherwise Random is competitive or better. The thresholds are an empirical heuristic calibrated on three JARVIS-DFT properties, not a universal rule.

	$R^2 \gtrsim 0.4$	$R^2 < 0.4$
Budget $< 5\%$	**QW-DPP**	Random
Budget $\geq 5\%$	Random	Random

dred seed labels, so a practitioner facing a new property can run these checks before committing the full DFT budget. We caveat that these thresholds come from three properties in JARVIS-DFT; whether they transfer to other properties, fingerprints, or databases is an open question.

A concrete seed set protocol. The two diagnostics suggest a simple check before committing a large DFT campaign. First, label a small seed set of a few hundred randomly drawn structures. Second, fit a cross-validated Ridge model on the seed fingerprints and read off R^2 as a cheap proxy for feature–target predictability. Third, compute the density coefficient of variation of the unlabeled pool, which needs no labels at all. If $R^2 \gtrsim 0.4$ and the intended budget is below $\sim 5\%$ of the pool, allocate the remaining budget with QW-DPP; otherwise fall back to Random. The seed labels are not wasted, since they fold directly into the final training set, and because both quantities are read off the seed set rather than the full pool, the check costs only a small fraction of the budget it governs.

6. Conclusion

We introduced QW-DPP, a quality-weighted determinantal point process that allocates DFT budgets without consuming a label. On JARVIS-DFT, QW-DPP improves test MAE over random sampling by up to $+8.9\%$ at the 1% budget on formation energy and bulk modulus, across three downstream models, with a crossover near 5% beyond which the methods converge. The mechanism is a tradeoff: QW-DPP trades a small amount of target distribution match for substantial gains in worst case feature space coverage. Across our three properties, QW-DPP's effectiveness correlates with feature target predictability and inversely with budget size, suggest-

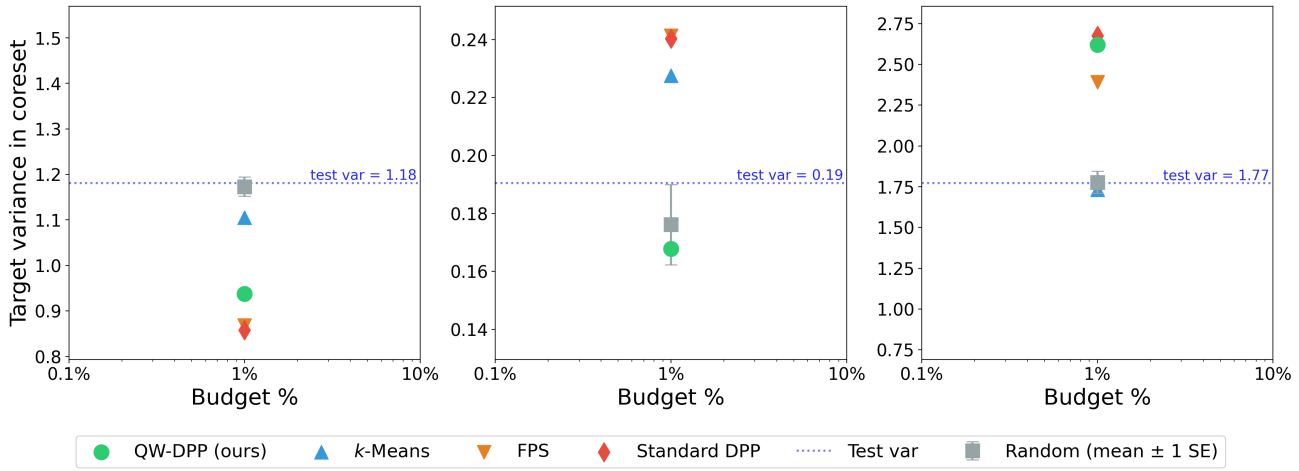

Figure 5. **QW-DPP preserves target variance on formation energy and bulk modulus but over-spreads on bandgap.** Coreset target variance for each method at the 1% budget; dotted line is the test set variance. (left) Formation energy. (middle) Bulk modulus. (right) Bandgap (negative case). Random is mean ± one standard error over 10 seeds; deterministic methods are single runs.

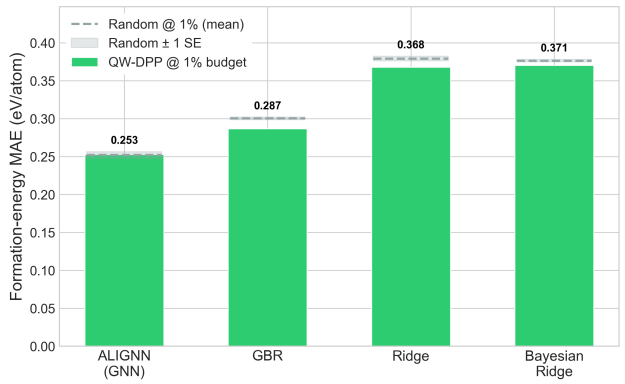

Figure 6. **One QW-DPP coreset transfers across model classes without hurting any.** A single QW-DPP coreset trains ALIGNN (GNN), Gradient Boosting, Ridge, and Bayesian Ridge at the 1% budget on JARVIS formation energy. QW-DPP beats Random on the three regressors and matches Random within seed variance on ALIGNN. Random is mean ± one standard error over 10 seeds; deterministic methods are single runs.

ing two cheap diagnostics—one needing only a small seed set, one fully label-free—that practitioners can compute before committing a large DFT campaign. By tying selection effectiveness to quantities that are cheap to estimate, QW-DPP lets practitioners check whether selection will help before spending the DFT budget.

Limitations and future work. Bandgap is a clean negative case: Magpie R^2 is only 0.39, and bandgap is dominated by electronic structure effects that composition fingerprints cannot see. Structure aware fingerprints (CGCNN/SOAP/ALIGNN embeddings) would presumably move it closer to the predictable regime. Our evaluation also uses JARVIS-DFT alone; whether the patterns transfer to other databases (Materials Project, OQMD, AFLOW), property classes, or fingerprint families remains open, as does comparison against learned, embedding-based selection strategies. Our statistical evidence further rests on a single fixed train/test split, with uncertainty quantified only over the Random selector's seeds; the deterministic selectors are point estimates tested against that distribution, and multi-split confidence intervals are left to future work. Three engineering considerations apply. (i) Composition-only fingerprints. Magpie carries no structural information, capping the regime where QW-DPP can help. (ii) Quadratic in budget greedy. The log-determinant MAP is $\mathcal{O}(M\,k^2)$, capping the subsample at $M = 35{,}000$ on a single CPU node; databases beyond $\sim 10^5$ entries would need landmark/Nyström approximations or a streaming variant. (iii) Heuristic quality. The density quality $q_i = \sqrt{\mathrm{rank}(\rho_i)/M}$ materially outperforms the $q \equiv 1$ ablation in the predictable regime, but we do not claim optimality among label-free quality functions.

Broader impact. On problems where features predict the target reasonably well ($R^2 \gtrsim 0.4$) and the budget is small ($< 5\%$ of the candidate pool), QW-DPP reduces the compute and energy cost of training DFT surrogate models, which can accelerate materials discovery in clean energy, catalysis, and electronic devices. Moreover, this label-free selection framework applies wherever expensive simulations or experiments generate training data, including molecular dynamics and computational fluid dynamics.

Acknowledgements

We thank the Department of Computer Science at Princeton University for providing the computational resources used in this work.

References

Balachandran, P. V., Xue, D., Theiler, J., Hogden, J., and Lookman, T. Adaptive strategies for materials design using uncertainties. *Scientific reports*, 6(1):19660, 2016.

Burke, K. Perspective on density functional theory. *The Journal of chemical physics*, 136(15), 2012.

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., and Walsh, A. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

Chen, L., Zhang, G., and Zhou, H. Fast greedy map inference for determinantal point process to improve recommendation diversity, 2018. URL https://arxiv.org/abs/1709.05135.

Choudhary, K. and DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1), November 2021. ISSN 2057-3960. doi: 10.1038/s41524-021-00650-1. URL http://dx.doi.org/10.1038/s41524-021-00650-1.

Choudhary, K., Garrity, K. F., Reid, A. C. E., DeCost, B., Biacchi, A. J., Hight Walker, A. R., Trautt, Z., Hattrick-Simpers, J., Kusne, A. G., Centrone, A., Davydov, A., Jiang, J., Pachter, R., Cheon, G., Reed, E., Agrawal, A., Qian, X., Sharma, V., Zhuang, H., Kalinin, S. V., Sumpter, B. G., Pilania, G., Acar, P., Mandal, S., Haule, K., Vanderbilt, D., Rabe, K., and Tavazza, F. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj Computational Materials*, 6(1), November 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00440-1. URL http://dx.doi.org/10.1038/s41524-020-00440-1.

Curtarolo, S., Setyawan, W., Hart, G. L., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., et al. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.

Dunn, A., Wang, Q., Ganose, A., Dopp, D., and Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1), 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00406-3. URL http://dx.doi.org/10.1038/s41524-020-00406-3.

Eldar, Y., Lindenbaum, M., Porat, M., and Zeevi, Y. Y. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315, 1997.

Friedman, D. and Dieng, A. B. The vendi score: A diversity evaluation metric for machine learning, 2023. URL https://arxiv.org/abs/2210.02410.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data, 2017. URL https://arxiv.org/abs/1703.02910.

Hedin, L. New method for calculating the one-particle green's function with application to the electron-gas problem. *Physical Review*, 139(3A):A796, 1965.

Jain, A., Ong, S., Hautier, G., Chen, W., Richards, W., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. The materials project: a materials genome approach to accelerating materials innovation. apl mater 1: 011002. *Available from DOI*, 10(1.4812323), 2013.

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., and Iyer, R. Glister: Generalization based data subset selection for efficient and robust learning, 2021. URL https://arxiv.org/abs/2012.10630.

Kirsch, A., van Amersfoort, J., and Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, 2019. URL https://arxiv.org/abs/1906.08158.

Kulesza, A. and Taskar, B. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1193–1200, 2011.

Kulesza, A. and Taskar, B. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, December 2012. ISSN 1935-8245. doi: 10.1561/2200000044. URL http://dx.doi.org/10.1561/2200000044.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers, 1994. URL https://arxiv.org/abs/cmp-lg/9407020.

Ling, J., Hutchinson, M., Antono, E., Paradiso, S., and Meredig, B. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integrating Materials and Manufacturing Innovation*, 6 (3):207–217, 2017. ISSN 2193-9772. doi: 10.1007/s40192-017-0098-z. URL http://dx.doi.org/10.1007/s40192-017-0098-z.

Lloyd, S. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Lookman, T., Balachandran, P. V., Xue, D., and Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):21, 2019.

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. Coresets for data-efficient training of machine learning models, 2020. URL https://arxiv.org/abs/1906.01827.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Nguyen, Q. and Dieng, A. B. Quality-weighted vendi scores and their application to diverse experimental design, 2024. URL https://arxiv.org/abs/2405.02449.

Schmidt, J., Marques, M. R., Botti, S., and Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj computational materials*, 5(1):83, 2019.

Settles, B. Active learning literature survey. *Computer Sciences Technical Report 1648, University of Wisconsin-Madison*, 2009.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2015.

Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning, 2019. URL https://arxiv.org/abs/1812.05159.

Tran, K. and Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for co2 reduction and h2 evolution. *Nature Catalysis*, 1(9):696–703, 2018.

Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1), August 2016. ISSN 2057-3960. doi: 10.1038/npjcompumats.2016.28. URL http://dx.doi.org/10.1038/npjcompumats.2016.28.

A. Implementation Details

A.1. Feature representation

All coreset selection operates on 132 dimensional Magpie composition fingerprints. Features are ℓ_2-normalised before kernel/distance computations.

A.2. QW-DPP selection

The QW-DPP L-ensemble is $L = \mathrm{diag}(q)\, K\, \mathrm{diag}(q)$ where K is the RBF kernel on ℓ_2-normalised Magpie features and q is the per point density-based quality. Concretely:

- Quality weight $q_i = \sqrt{\mathrm{rank}(\rho_i)/M}$ where ρ_i is the κ-NN density (defined below), $\mathrm{rank}(\rho_i)$ is the ascending rank of ρ_i among the M subsample points, and M is the subsample size used for selection (§4.3).
- Local density: $\rho_i = 1/(\overline{d}_i + \varepsilon)$ with $\varepsilon = 10^{-8}$, where $\overline{d}_i$ is the mean distance to the $\kappa = 10$ nearest neighbours of point i in ℓ_2-normalised feature space, computed on the same M-point candidate subsample (seed 42) used for selection.
- Selection: greedy log-determinant maximisation on L over the candidate pool until the budget is filled.
- Standard DPP baseline: same algorithm with $q \equiv 1$.

A.3. Models

Gradient Boosting. sklearn.ensemble.GradientBoostingRegressor with n_estimators=200, max_depth=5, learning_rate=0.1, random_state=42. No row or column subsampling, so training is fully deterministic given the input data.

Ridge. sklearn.linear_model.Ridge with α chosen from a 20 point log-spaced grid in $[10^{-3}, 10^{3}]$ via 5 fold cross validation on the training coreset. Closed form, deterministic given α.

Bayesian Ridge. sklearn.linear_model.BayesianRidge with max_iter=300, tol=1e-3; remaining hyperparameters left at sklearn defaults. Closed form variational fit, deterministic.

ALIGNN. Reference implementation of (Choudhary & DeCost, 2021). Training configuration: 50 epochs, batch size 32, neighbour-list cutoff 8.0 Å, hidden width 256. No early stopping. Training is stochastic (weight initialisation and mini-batch order vary across runs); per seed spread is reported in Appendix D.

A.4. Coresets

Random gets ten seeds. Deterministic methods (QW-DPP, k-Means, FPS, Standard DPP) emit a single file per budget.

B. Additional Coverage and Distribution Match Metrics

The main paper's Figure 2 reports max NN distance from test to coreset (worst case feature space coverage) at the 1% budget, and Figure 3 reports Kolmogorov–Smirnov distance between coreset and test target distributions on formation energy. This appendix adds six complementary metrics across both axes (feature space coverage and distribution match in feature and target space) on the same coresets. For all metrics, lower is better.

Metric definitions.

- **M1 mean NN**: mean nearest neighbour distance from test to coreset in ℓ_2-normalised feature space.
- **M2 p95 NN**: 95th percentile nearest neighbour distance (tail coverage).
- **M3 max NN**: maximum nearest neighbour distance (worst case coverage; reported in Figure 2).
- **M4 MMD^2 (RBF)**: squared maximum mean discrepancy between coreset and test feature samples with median heuristic bandwidth.

- **M5 energy distance**: distribution level discrepancy between coreset and test feature samples.
- **M6 stratified mean NN**: mean NN distance averaged across five target quantile bins (uniform weight on rare and common target regions).
- **M7 KS-target**: Kolmogorov–Smirnov statistic between coreset and test target distributions.
- **M8 W_1-target**: 1-Wasserstein distance between coreset and test target distributions.

Results. Table 4 reports each metric by (property, method) at the 1% budget on JARVIS.

Table 4. **QW-DPP leads on worst case coverage (M3) across all three properties and on most target distribution metrics (M7, M8) on bulk modulus and bandgap.** Eight coverage and distribution metrics at the 1% budget on JARVIS, by (property, method). Lower is better. Random is mean over 10 seeds; deterministic methods are single runs. $*$: $p < 0.05$ vs. Random (t-test). Best per column in **bold green**, second best in blue.

Property	Method	M1 mean NN	M3 max NN	M5 energy	M7 KS-y	M8 W_1-y
formation_energy	QW-DPP (ours)	0.461*	**0.962***	0.0109*	0.0973*	0.190*
	Random	0.493	1.052	**0.0038**	**0.0344**	**0.0556**
	k-Means	**0.428***	0.994*	0.0138*	0.0950*	0.189*
	Standard DPP	0.631*	1.122*	0.0693*	0.346*	0.711*
	FPS	0.609*	1.101*	0.0655*	0.311*	0.705*
bulk_modulus	QW-DPP (ours)	0.590*	**1.068***	0.0092	**0.0301***	**0.0201***
	Random	0.620	1.177	**0.0083**	0.0630	0.0442
	k-Means	**0.580***	1.080*	0.0098*	0.0511	0.0346
	Standard DPP	0.767*	1.199	0.0604*	0.141*	0.104*
	FPS	0.735*	1.157	0.0524*	0.144*	0.102*
bandgap	QW-DPP (ours)	0.461*	**0.962***	0.0109*	**0.0152***	0.0721*
	Random	0.493	1.052	**0.0038**	0.0289	**0.0582**
	k-Means	**0.428***	0.994*	0.0138*	0.0476*	0.0868*
	Standard DPP	0.631*	1.122*	0.0693*	0.125*	0.206*
	FPS	0.609*	1.101*	0.0655*	0.142*	0.244*

C. Per-Property Extended Results

This appendix reports per property results that did not fit in the main paper. All experiments use the same train/test split and selection pipeline as §5.

C.1. Bulk modulus

The main paper reports the 1% Ridge / Gradient Boosting cells in Table 2. Table 5 adds the full Gradient Boosting budget sweep. QW-DPP achieves the lowest MAE at the smallest budgets (0.5%, 1%) and at 3%–5%, where the gap over Random is largest; k-Means is competitive throughout and overtakes at 2% and at higher budgets ($\geq 10\%$).

Table 5. Bulk modulus test MAE under Gradient Boosting across budgets. Random is mean $\pm$ one standard error over 10 seeds; deterministic methods are single runs. $*$: $p < 0.05$ vs. Random (t-test). The best method per column is shown in **bold green**; the second best is shown in blue.

Method	0.5%	1%	2%	3%	5%	10%	25%	50%
QW-DPP (ours)	**0.234***	**0.203***	0.206	**0.194**	**0.184***	0.174*	0.169	0.159*
Random	0.255 ± 0.003	0.222 ± 0.005	0.205 ± 0.003	0.195 ± 0.002	0.190 ± 0.002	0.179 ± 0.001	0.168 ± 0.001	0.162 ± 0.001
k-Means	0.295*	0.214	**0.198***	0.200*	0.190	**0.171***	**0.163***	0.159*
Standard DPP	0.289*	0.239*	0.204	0.229*	0.240*	0.218*	0.176*	0.158*
FPS	0.254	0.230	0.203	0.197	0.206*	0.189*	0.169	**0.157***

D. ALIGNN Details

D.1. Training configuration

ALIGNN reference implementation (Choudhary & DeCost, 2021), trained for 50 epochs with batch size 32, neighbour-list cutoff 8.0 Å, hidden width 256, learning rate per the upstream default. No early stopping. Random rows aggregate ten coreset seeds; deterministic coreset rows report a single ALIGNN training run per (method, budget) setting.

D.2. Per budget results

Table 6 reports ALIGNN MAE at every budget across all five selectors. Deterministic coreset entries are point estimates; the Random column is the 10 seed mean ± standard error. QW-DPP remains within 0.003 eV/atom of the leading method at every budget tested, confirming that one QW-DPP coreset selected on Magpie fingerprints transfers cleanly to a structure aware GNN without hurting performance (Figure 6 in the main paper).

Table 6. **One QW-DPP coreset transfers to ALIGNN across all budgets without hurting performance.** ALIGNN test MAE on JARVIS formation energy by selector and budget. Random is mean ± one standard error over 10 seeds; deterministic methods are single runs. Best per column in **bold green**, second best in blue.

Method	0.5%	1%	2%	3%	5%	10%
QW-DPP	**0.304**	0.253	0.198	**0.154**	**0.124**	0.102
Random (10s.)	0.330 ± 0.008	**0.252 ± 0.004**	**0.187 ± 0.002**	0.165 ± 0.004	0.131 ± 0.002	0.101 ± 0.001
k-Means	0.307	0.253	0.191	0.161	0.131	**0.099**
FPS	0.344	0.281	0.225	0.165	0.131	0.103
Standard DPP	0.384	0.292	0.226	0.183	0.146	0.106

E. Reproducibility

E.1. Data

- **JARVIS-DFT.** 75,993 structures total, split 80 / 20 into train (60,795) and test (15,198) by random index assignment with a fixed seed.
- **Feature pipeline.** 132 dimensional Magpie composition descriptors via matminer 0.10.0.

E.2. Software

Python 3.13.5; scikit-learn 1.7.2; NumPy 2.2.6; SciPy 1.17.1; pandas 2.3.3; matplotlib 3.10.6; matminer 0.10.0; PyTorch 2.9.1.

E.3. Hardware

Magpie feature pipeline, kernel construction, greedy QW-DPP selection, Gradient Boosting / Ridge / Bayesian Ridge training, and all coverage metrics run on a single CPU node (∼minutes per budget sweep). ALIGNN training runs on a single GPU.

E.4. Random seeds

- **QW-DPP / k-Means / FPS / Standard DPP selection.** Deterministic; no seed required after the candidate subsample is fixed.
- **Candidate subsample for selection.** np.random.seed(42) when subsampling $M = \min(35{,}000, n_{\text{train}})$ candidates from the train pool. For formation energy and bandgap ($n_{\text{train}} = 60{,}795$), $M = 35{,}000$; for bulk modulus ($n_{\text{train}} = 14{,}888$, the 80% train split of $N_{\text{valid}} = 18,610$), $M = 14{,}888$.
- **Random coreset baseline.** Seeds 0–9.
- **Gradient Boosting.** random_state=42.

- **ALIGNN.** random_seed=42.

F. Decomposing QW-DPP: Kernel vs. Objective

QW-DPP combines two design choices that could each independently drive its gain over baselines: (i) a quality-weighted L-ensemble $L = \mathrm{diag}(q)\, K\, \mathrm{diag}(q)$ that down weights outliers, versus a bare similarity kernel K that does not; and (ii) a log-determinant MAP selector that maximizes covered volume, versus a Rényi-entropy MAP selector—a generalization from the Vendi Score (VS) family (Friedman & Dieng, 2023)—that maximizes spectral entropy. Crossing these two design choices yields four selectors, which we denote by (objective)(kernel):

- **DPP**(L): log-det MAP on the weighted kernel; this is QW-DPP.
- **DPP**(K): log-det MAP on the bare kernel; this is Standard DPP from the main paper.
- **VS**$_\alpha(L)$: greedy Rényi-α entropy MAP on the weighted kernel; this is the Vendi Score family selector (Friedman & Dieng, 2023) applied to L.
- **VS**$_\alpha(K)$: greedy Rényi-α entropy MAP on the bare kernel; this is the Vendi Score family selector applied to K.

We note that $\mathrm{VS}_\alpha(L)$ is not equivalent to the quality-weighted Vendi Score of Nguyen & Dieng (2024): quality-weighted Vendi reweights the spectrum directly, while we apply quality weighting through the kernel factorization $L = \mathrm{diag}(q)\, K\, \mathrm{diag}(q)$ and then take the entropy of the resulting spectrum. The two yield different selection criteria; we use the kernel side formulation here for direct comparability with DPP-MAP(L).

Two contrasts isolate the contribution of each design choice:

- Comparing L to K within the same selector (DPP(L) vs. DPP(K); $\mathrm{VS}_\alpha(L)$ vs. $\mathrm{VS}_\alpha(K)$) isolates the contribution of **quality weighting**.
- Comparing log-det MAP to VS-MAP on the same kernel (DPP(L) vs. $\mathrm{VS}_\alpha(L)$; DPP(K) vs. $\mathrm{VS}_\alpha(K)$) isolates the contribution of the **selection objective**.

We run all four cells on JARVIS formation energy with three downstream regressors. Random is included as a non coreset reference.

F.1. Headline 2×2 with Rényi-2 entropy

Rényi-2 (collision) entropy admits a closed-form $\mathcal{O}(s)$ incremental update at our $M = 35{,}000$ subsample, so the comparison runs at the same subsample as the main paper experiments and all four cells are directly comparable. We use $\alpha = 2$ here because it is the only Rényi order whose greedy MAP selector scales to our full subsample size; we relax this in §F.2.

Table 7. **Quality weighting and the log-determinant objective both contribute; the objective contributes more.** JARVIS formation energy MAE at the 1% budget across three downstream regressors, with all four cells of the (kernel × selector) grid plus Random. Random is mean over 5 seeds ± standard error; deterministic methods are single runs. Best per row in **bold green**.

Model	DPP(L)	DPP(K)	VS$_2(L)$	VS$_2(K)$	Random
Ridge	**0.368**	0.426	0.480	0.441	0.385 ± 0.005
Gradient Boosting	**0.287**	0.359	0.390	0.354	0.300 ± 0.003
Bayesian Ridge	**0.371**	0.429	0.484	0.442	0.380 ± 0.002

Table 7 lets us read off both contrasts directly:

Quality weighting (L vs. K, fixing the objective). Under log-det MAP, quality weighting helps on every regressor: DPP(L) beats DPP(K) by 0.058 eV/atom on Ridge, 0.072 on Gradient Boosting, and 0.058 on Bayesian Ridge. Under VS-MAP, however, quality weighting hurts: $\mathrm{VS}_2(L)$ is worse than $\mathrm{VS}_2(K)$ by 0.036–0.042 eV/atom across the three regressors. The two design choices therefore interact rather than decompose additively; quality weighting helps only when the selector is log-det MAP.

Selection objective (DPP-MAP vs. VS-MAP, fixing the kernel). On the weighted kernel L, log-det MAP beats VS_2 by 0.103–0.113 eV/atom across the three regressors. On the bare kernel K, log-det MAP also beats VS_2 but by smaller margins (0.005–0.013 eV/atom). The objective contributes regardless of which kernel it is paired with.

The full QW-DPP cell (DPP(L)) is therefore the best of the four on every regressor, with much of the gain coming from pairing quality weighting with the log-det objective.

F.2. Robustness to Rényi order ($\alpha = 1$)

The canonical Vendi Score (Friedman & Dieng, 2023) uses Shannon entropy ($\alpha = 1$). Greedy maximization of Shannon entropy of the spectrum has no closed form incremental update—each candidate evaluation requires an $\mathcal{O}(s^3)$ eigendecomposition—so we re-ran the comparison at a smaller subsample $M = 1{,}000$ where this is tractable. Both selectors saw the same $M = 1{,}000$ subsample for fairness.

We compare $\alpha = 1$ to log-det MAP on the same kernel L in Table 8.

Table 8. **The log-determinant objective beats the Shannon-entropy Vendi Score on the same weighted kernel.** JARVIS formation energy MAE at the 1% budget; $\alpha = 1$ (Shannon-entropy VS) versus log-det MAP, both on the weighted kernel L at $M = 1{,}000$. Best per row in **bold green**.

Regressor	$\mathrm{VS}_1(L)$	DPP-MAP(L)	Gap
Ridge	0.391	**0.381**	+0.010
Gradient Boosting	0.345	**0.312**	+0.034
Bayesian Ridge	0.401	**0.381**	+0.020

The objective isolation contrast survives at $\alpha = 1$: DPP-MAP(L) beats $\mathrm{VS}_1(L)$ on every regressor, by 0.010–0.034 eV/atom. Sweeping α across the standard range (Table 9) confirms that no Rényi order matches log-det MAP on average.

Table 9. **No Rényi-α Vendi Score variant matches log-det MAP on average.** Rényi-α sweep at the 1% budget on the weighted kernel L at $M = 1{,}000$. Columns are MAE per regressor; the final column averages across the three regressors. Best per column in **bold green**, second best in blue.

Selector	Ridge	GB	Bay. Ridge	Average
$\mathrm{VS}_{0.1}(L)$	**0.380**	0.329	0.387	0.365
$\mathrm{VS}_{0.5}(L)$	0.383	0.322	0.390	0.365
$\mathrm{VS}_1(L)$	0.391	0.345	0.401	0.379
$\mathrm{VS}_2(L)$	0.383	0.324	0.390	0.366
$\mathrm{VS}_\infty(L)$	0.386	0.341	0.393	0.373
DPP-MAP(L)	0.381	**0.312**	**0.381**	**0.358**

$\mathrm{VS}_{0.1}$ wins on Ridge by a single point (0.380 vs. 0.381), but DPP-MAP wins on the other two regressors and on the average, reproducing the $\alpha = 2$ ordering at smaller subsample. The ablation confirms that QW-DPP's gain is not specific to the particular Rényi order chosen for the headline experiment.