# Maximizing Efficiency of Dataset Compression for Machine Learning Potentials With Information Theory

**Benjamin Yu,[1] Vincenzo Lordi,[2] Daniel Schwalbe-Koda[1*]**

[1]Department of Materials Science and Engineering, University of California, Los Angeles, CA, United States
[2]Materials Science Division, Lawrence Livermore National Laboratory, CA, United States

**Machine learning interatomic potentials (MLIPs) balance high accuracy and lower costs compared to density functional theory calculations, but their performance often depends on the size and diversity of training datasets. Large datasets improve model accuracy and generalization but are computationally expensive to produce and train on, while smaller datasets risk discarding rare but important atomic environments and compromising MLIP accuracy/reliability. Here, we develop an information-theoretical framework to quantify the efficiency of dataset compression methods and propose an algorithm that maximizes this efficiency. By framing atomistic dataset compression as an instance of the minimum set cover (MSC) problem over atom-centered environments, our method identifies the smallest subset of structures that contains as much information as possible from the original dataset while pruning redundant information. The approach is extensively demonstrated on the GAP-20 and TM23 datasets, and validated on 64 varied datasets from the ColabFit repository. Across all cases, MSC consistently retains outliers, preserves dataset diversity, and reproduces the long-tail distributions of forces even at high compression rates, outperforming other subsampling methods. Furthermore, MLIPs trained on MSC-compressed datasets exhibit reduced error for out-of-distribution data even in low-data regimes. We explain these results using an outlier analysis and show that such quantitative conclusions could not be achieved with conventional dimensionality reduction methods. The algorithm is implemented in the open-source QUESTS package and can be used for several tasks in atomistic modeling, from data subsampling, outlier detection, and training improved MLIPs at a lower cost.**

## 1. Introduction

Machine learning (ML) interatomic potentials (IPs) show excellent ability to approximate potential energy surfaces (PES) with lower cost compared to density functional theory (DFT) calculations.[1–11] Particularly with neural network (NN) models, MLIPs reach impressively high accuracy across spaces of structures and compositions, but are often governed by power laws that relate dataset sizes to model performance.[9,12] In general, larger datasets tend to substantially improve the accuracy of models, though at the expense of higher costs for generating larger datasets and training models on them. Multiple works have proposed "foundation-like" approaches to MLIPs, where NNs are trained on very large datasets of energy minimization trajectories, well-sampled PESes, and others.[12–16] Nevertheless, training models on increasingly diverse and well-sampled PESes requires ensuring that atomistic datasets are not becoming redundant, and are instead well-representative of increasingly larger phase spaces.[17–26] Thus, there is a need for tools that analyze and subsample datasets to remove redundancies, maximizing the diversity of the data while also minimizing dataset sizes, which reduces training cost and complexity.[27–30]

*E-mail: dskoda@ucla.edu

In MLIPs, dataset subsampling methods often rely on partitioning schemes that sample structures (i.e., collections of atoms in space, either with or without boundary conditions) that are mostly distinct from each other based on a given representation.[31–33] Often, to encode structures in a size-invariant manner, representations are created by taking the average of either learned or fixed per-atom representations across environments in a molecule or unit cell. However, this averaging scheme often leads to information losses or introduction of redundancies in a dataset. For instance, average representations are unable to distinguish between a supercell and a primitive cell of a ground state structure, which carry the same amount of information, yet contain different number of atoms. A dataset built using the supercell has a larger memory footprint at training even though the additional atoms do not contribute additional information to the dataset and indeed can overweight those redundant atomic environments. On the other hand, information losses also occur when interesting outlier behavior — e.g., rare events, metastability, and more — are underrepresented in subsampled datasets due to their sparse nature. In many subsampling methods, there is no guarantee that rare events will be reliably captured in the downsampled datasets despite their importance in atomistic modeling. The redundancy-outlier tradeoff is exacerbated when outlier environments in large cells are overlooked due to the use of average, per-structure representations.

In this work, we use information theory and combinatorics to minimize information losses when subsampling atomistic datasets. Specifically, we created an algorithm that compresses datasets by solving the minimum set cover (MSC) problem over the space of atomic environments. The MSC analysis attempts to find a minimum number of structures that spans the entire distribution of environments in the original dataset, thus maximizing the retained information on a per-dataset basis. Our algorithm is demonstrated to preserve outliers, the long tail of force distributions, and the diversity of the datasets, outperforming baseline dataset compression methods, including farthest point sampling. Our approach is exemplified with case studies on the GAP-20[32] and TM23[34] datasets, where outlier analysis and model training demonstrate the efficiency of our compression algorithm compared to the baseline approaches. We also benchmark the algorithm on 64 different datasets obtained from the ColabFit repository[35] and show our approach outperforms the baseline algorithms in the vast majority of cases. This work provides an information-theoretical strategy to automate the compression of atomistic datasets while minimizing loss of outliers and information. The algorithms and implementation are publicly available in the QUESTS package at https://github.com/dskoda/quests.

## 2. Methods

### 2.1. Information entropy and QUESTS method

Central to the idea of measuring efficiency of data compression is the theory of information from Shannon.[36] Here, we employ an information theoretical perspective to atomistic simulations[25] to propose algorithms that compress atomistic datasets, as well as figures of merit to evaluate their performance. Below, we describe the representation and relevant information-theoretical quantities used in the manuscript.

**Representation:** The representation of atomic environments was computed as described in our previous work.[25] In short, an atom-centered descriptor is obtained by sorting the distances from a central atom $i$ to its $k$ nearest neighbors (considering periodic boundary conditions), forming a vector $\mathbf{X}_i^{(1)}$ of length $k$ as

$$\mathbf{X}_i^{(1)} = \left[ \frac{w(r_{i1})}{r_{i1}} \quad \dots \quad \frac{w(r_{ik})}{r_{ik}} \right]^T, \quad r_{ij} \leq r_{i(j+1)}, \tag{1}$$

where $1 \leq j \leq k$ and $w$ is a smooth cutoff function,

$$w(r) = \begin{cases} \left[ 1 - \left( \frac{r}{r_c} \right)^2 \right]^2, & 0 \leq r \leq r_c, \\ 0, & r > r_c. \end{cases} \tag{2}$$

As $\mathbf{X}_i^{(1)}$ accounts only for two-body terms, we proposed a second term $\mathbf{X}_i^{(2)}$ that accounts for three-body terms by measuring the distances between neighbors of atom $i$:

$$X_{in}^{(2)} = \left\langle \frac{\sqrt{w(r_{ij}) w(r_{il})}}{r_{jl}} \right\rangle_n, \quad j, l \in \mathcal{N}(i),\ X_{in} \geq X_{i(n+1)}, \tag{3}$$

where $j$ and $l$ are atoms in the neighborhood $\mathcal{N}(i)$ of $i$, $\langle . \rangle_n$ denotes the average over the $n$-th terms, and $1 \leq n \leq k-1$. The final descriptor $\mathbf{X}_i$ is formed by concatenating $\mathbf{X}_i^{(1)}$ and $\mathbf{X}_i^{(2)}$. We used the default hyperparameters for the representation, as published in our prior work, namely a number of $k = 32$ neighbors and a cutoff of $r_c = 5$ Å were used to represent each environment. Extensive details on the representation are provided in the original reference.[25]

**Information entropy:** the information entropy $\mathcal{H}$ of a set of descriptors $\{\mathbf{X}\}$ encodes the amount of information contained in the descriptor distribution.[25] We quantified the information entropy from a set $\{\mathbf{X}\}$ of $N$ feature vectors $\mathbf{X}_i$ using a kernel density estimate over the empirical distribution of points,

$$\mathcal{H}(\{\mathbf{X}\}) = -\frac{1}{N} \sum_{i=1}^{N} \log \left[ \frac{1}{N} \sum_{j=1}^{N} K_h(\mathbf{X}_i, \mathbf{X}_j) \right], \tag{4}$$

where our choice of the kernel $K_h$ is the Gaussian kernel,

$$K_h(\mathbf{X}_i, \mathbf{X}_j) = \exp\left( \frac{-||\mathbf{X}_i - \mathbf{X}_j||^2}{2h^2} \right) \tag{5}$$

and $h$ is the bandwidth of the kernel. For the bandwidth, we selected the default bandwidth of $h = 0.015$ Å$^{-1}$ in the QUESTS method, which has proven useful for a range of applications and datasets.[25] Throughout this work, the natural logarithm was used for computing the information entropy, scaling the units of information to nats.

**Differential entropy:** To determine whether individual data points $\mathbf{Y}$ are contained in the distribution induced by $\{\mathbf{X}\}$, we compute the relative information provided by $\mathbf{Y}$ given the current knowledge of $\{\mathbf{X}\}$, defining the so-called differential entropy $\delta\mathcal{H}$ as:[25]

$$\delta\mathcal{H}(\mathbf{Y}|\{\mathbf{X}\}) = -\log \left[ \sum_{i=1}^{N} K_h(\mathbf{Y}, \mathbf{X}_i) \right]. \tag{6}$$

With this definition, $\delta\mathcal{H} \leq 0$ only if $\mathbf{Y}$ is contained in the distribution induced by $\{\mathbf{X}\}$. If $\delta\mathcal{H} > 0$, then $\mathbf{Y}$ provides additional "surprise" to the dataset $\{\mathbf{X}\}$. Within Eq. (6), the kernel $K_h$ is chosen to be Gaussian, the bandwidth $h$ is equal to 0.015 Å$^{-1}$, and the unit of $\delta\mathcal{H}$ is nats, thus mirroring Eq. (4).

**Dataset diversity:** The diversity of a given set of descriptors $\{\mathbf{X}\}$ is defined as:[25]

$$\mathcal{D}(\{\mathbf{X}\}) = \log\left[\sum_{i=1}^{N} e^{\delta\mathcal{H}(\mathbf{X}_i|\{\mathbf{X}\})}\right], \tag{7}$$

which approximates a measure of the support of the distribution. While the dataset diversity $\mathcal{D}$ is an approximation and has not been mathematically proven to be an exact measure, it is quite useful in defining the size of the phase space. The entropy $\mathcal{H}$, on the other hand, takes into account the probability of each data point and may be low even for high-coverage phase spaces. For instance, a near-degenerate distribution (e.g., nearly all the probability is concentrated in a single point) exhibits low information entropy, even though its support may be vast. In this case, its diversity will remain high despite the low entropy.

## 2.2. Evaluation methods

Evaluating the performance of a dataset subsampling technique based on test errors obtained from models trained on these smaller datasets is the usual method to measure subsampling quality, but it relies on somewhat expensive model training runs and is prone to the variability in the training process and in the test distributions. Alternatively, our model-free analysis measure the effectiveness of the compression algorithms in preserving the original distribution of the dataset by quantifying dataset completeness and overlap between datasets from atomistic simulations.[25] In this work, we use all these metrics to compare the efficiency of dataset compression algorithms in subsampling atomistic datasets. Figure 1 illustrates three of such assessment techniques, each of which is discussed below.

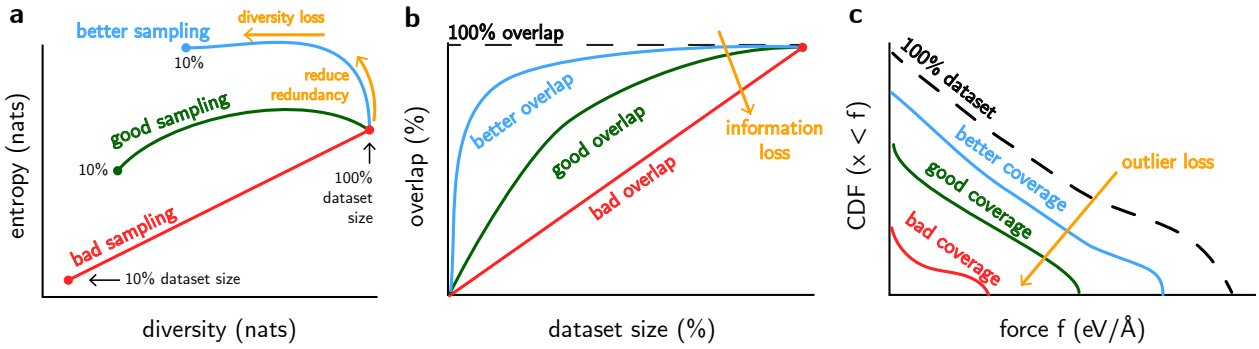


**Fig. 1**: **Figures of merit that quantify the efficiency of a dataset compression algorithm. a**, An ideal algorithm should retain as much as the original dataset diversity as possible while removing redundancies from the data. In an entropy vs. diversity plot, this is depicted by curves that increase in entropy without reducing the diversity. At higher levels of compression, however, information loss is inevitable, and dataset diversity naturally start to decrease. **b**, An ideal algorithm should maximize the overlap of a compressed dataset with the original, complete dataset. Overlaps near 100% indicate a compression with effectively no information loss. **c**, An ideal dataset compression method should retain outlier environments as much as possible. Given that high-force environments tend to be outliers in datasets sampled from molecular dynamics or optimization trajectories, preserving this "long tail" of the force distribution is evidence of an algorithm that prioritizes these outlier environments. This long tail is represented by the cumulative distribution function (CDF) of forces.

**Entropy and diversity:** First, we define subsampling performance by using the entropy and diversity metrics defined in Eqs. (4) and (7). A good subsampling algorithm should maximize the entropy

of a dataset (i.e., remove the redundancy in the data) while also preserving as much of its original diversity as possible. To quantify this, we subsample a dataset at multiple fractions of its original size and calculate the entropy and diversity of the sampled dataset for different compression algorithms (Fig. 1a). For a constant dataset size, worse subsampling methods will lead to datasets with smaller entropy and diversity, indicating information losses when subsampling the data.

**Overlap:** Given the importance of preserving the support of the data distribution, a second possible figure of merit is the overlap of a dataset $\{\mathbf{X}\}$ with respect to a reference dataset $\{\mathbf{X}_{\text{ref}}\}$. Intuitively, the overlap represents the proportion of environments from $\{\mathbf{X}\}$ which are contained in $\{\mathbf{X}_{\text{ref}}\}$. This can be quantified using the definition of $\delta\mathcal{H}$ in Eq. (6) and the notion of "contained" as $\delta\mathcal{H} \leq 0$, as

$$\text{Overlap}\left(\{\mathbf{X}\}|\{\mathbf{X}_{\mathbf{ref}}\}\right) = \frac{\#\text{ envs with } \delta\mathcal{H}(\mathbf{X}_i \in \{\mathbf{X}\}|\{\mathbf{X}_{\mathbf{ref}}\}) \leq 0}{\#\text{ envs in } \{\mathbf{X}\}}. \tag{8}$$

Given that the overlap cannot be larger than 100%, a good compression algorithm should maximize the overlap of the subsampled dataset with respect to the original dataset (Fig. 1b). On the other hand, a worse compression algorithm quickly leads to information loss as soon as the dataset size is reduced, especially for low-redundancy datasets. This loss of information is undesirable in a range of scenarios in MLIPs, given that the cost associated with generating and labeling the data may be higher than including an additional point to the training set.

**Dataset efficiency:** Similarly to the concept of "efficiency" in information theory,[36] we define the efficiency of a dataset containing $n$ environments as the information entropy of that dataset divided by the maximum entropy of the dataset ($\log n$). A dataset with 100% efficiency necessarily has a uniform distribution of environments, thus without any overlaps between environments. On the other hand, datasets with low efficiency contain a limited amount of information, but high number of environments, indicating oversampling or redundancy in the data.

**Long tail of force distributions:** Finally, beyond descriptor distributions, we evaluate the performance of compressed datasets according to the distribution of force magnitudes across all environments. Given that high-force configurations are typically rarer than low-force configurations, preserving the "long tail" of the distribution of forces is important when subsampling datasets. A more effective compression algorithm is expected to produce a force distribution with a longer tail, i.e., one that captures a wider range of force magnitudes, indicating broader coverage of the original dataset and retention of outliers. To represent this long tail, we compute a cumulative distribution function (CDF) of force magnitudes, $\text{CDF}(x < f)$, as a function of the threshold $f$ (Fig. 1c). With this definition, a reduced loss of outliers is represented by a higher value of the CDF at higher thresholds $f$.

### 2.3. Baseline dataset compression methods

To assess common baseline subsampling algorithms, we implemented three main algorithms: mean farthest-point sampling (FPS), k-means clustering, and random selection. For all these algorithms, we focus on datasets of structures with periodic boundary conditions where each cell may exhibit large variability in the number of atoms. The assumption of the compression methods in this work is that all environments of structures in the dataset will be used during the training/testing process, regardless of their (possible) internal redundancies. Many methods, on the other hand, do not require entire structures, but only selected environments for training. In these cases, all algorithms could be used in exactly the same way, but with a per-atom treatment rather than a per-structure one.

**Random Sampling:** Within random sampling, the most trivial of subsampling strategies, structures (i.e., sets of descriptors $\{\mathbf{X}\}$) are randomly selected given a target fraction of the dataset size (see Algorithm 1). In this case, representations are not needed in order to subsample the datasets. Sometimes, random sampling from uniform bins (e.g., energy levels) is used to increase the diversity of the data,[37] but in this work we consider only random samples of the dataset.

**Algorithm 1:** Random sampling

**Require:** Set of $N$ structures $\mathcal{S} = \{S_i\}$ with environments $S_i = \{\mathbf{X}_1^{(i)}, \ldots, \mathbf{X}_n^{(i)}\}$, target size $K \leq N$

1: Randomly select $K$ structures $S_i^{(c)}$ from $\mathcal{S}$ without replacement
2: **return** $\mathcal{C} = \{S_1^{(c)}, \ldots, S_K^{(c)}\}$

***k*-means sampling:** Within $k$-means sampling, we represent each structure of the dataset by taking the mean of all per-atom representations in the structure. In particular, the representation we adopted is the one within the QUESTS method (see Eqs. (1) and (3)), as it is later used to compute entropy or diversity values. Then, we cluster these per-structure vectors into $k$ groups using a typical $k$-means clustering method and randomly sample one structure from each cluster. The sampled structures are used to create a compressed dataset (see Algorithm 2). We used the $k$-means clustering method implemented in `scikit-learn` (v. 1.7.0). This approach has been used multiple times in the development of datasets for MLIPs.[38,39] Variations of this method employ hierarchical clustering rather than $k$-means to stratify the space of features in clusters which contain similar information.[33,40]

**Algorithm 2:** $k$-means sampling

**Require:** Set of $N$ structures $\mathcal{S} = \{S_i\}$ with environments $S_i = \{\mathbf{X}_1^{(i)}, \ldots, \mathbf{X}_n^{(i)}\}$, target size $K \leq N$

1: Calculate a per-structure representation $\mathbf{S}_i$ from per-atom representations of $S_i$, $\mathbf{S}_i = \langle \mathbf{X}_j^{(i)} \rangle$
2: Cluster the set of structure vectors $\{\mathbf{S}_i\}$ in $K$ clusters using a $k$-means clustering method
3: **for** $n = 1$ **to** $K$ **do**
4: **1.** Select a random element from cluster $n$
5: **2.** Add the selected structure $S_n^{(c)}$ to the compressed set $\mathcal{C}$
6: **end for**
7: **return** $\mathcal{C} = \{S_1^{(c)}, \ldots, S_K^{(c)}\}$

**Mean Farthest Point Sampling:** In mean farthest point sampling (FPS), we begin with per-structure representations obtained by averaging the per-atom QUESTS representations across each structure, as used for the $k$-means sampling method. The algorithm starts by randomly selecting one structure vector as the initial seed for the compressed dataset. Next, we iteratively compute the Euclidean distances between all vectors in the compressed dataset and those remaining in the full dataset. At each iteration, the structure whose mean vector has the largest average distance to all vectors in the compressed dataset is selected, added to the compressed dataset, and removed from the pool of remaining candidates. This process continues until the compressed dataset reaches the size specified by the user. (see Algorithm 3). Variations of FPS have been extensively employed in subsampling datasets,[41] including datasets for MLIPs.[28,31,32,42–47] Related algorithms based on matrix factorization methods such as CUR[48] have also been used to subsample atomistic datasets by maximizing the distance between representations.[28,49–52]

**Algorithm 3:** Mean farthest point sampling

**Require:** Set of $N$ structures $\mathcal{S} = \{S_i\}$ with environments $S_i = \{\mathbf{X}_1^{(i)}, \ldots, \mathbf{X}_n^{(i)}\}$, target size $K \leq N$

1: Calculate a per-structure representation $\mathbf{S}_i$ from per-atom representations of $S_i$, $\mathbf{S}_i = \langle \mathbf{X}_j^{(i)} \rangle$
2: Randomly select a structure $S_i^{(c)}$ from $\mathcal{S}$ and add it to the initialized compressed set $\mathcal{C}_1$
3: **for** $n = 1$ **to** $K - 1$ **do**
4: **1.** Calculate the distances $d\left(S_i, S_j^{(c)}\right) = \left\| \mathbf{S}_i - \mathbf{S}_j^{(c)} \right\|_2$ between $S_i \in \mathcal{S} \setminus \mathcal{C}_n$ and $S_j^{(c)} \in \mathcal{C}_n$
5: **2.** Find $S_{n+1}^{(c)} = \arg\max_i \sum_{j=1}^{n} d\left(S_i, S_j^{(c)}\right)$
6: **3.** Define $\mathcal{C}_{n+1} = \mathcal{C}_n \cup \{S_{n+1}^{(c)}\}$
7: **end for**
8: **return** $\mathcal{C} = \{S_1^{(c)}, \ldots, S_K^{(c)}\}$

## 2.4. Compressing atomistic datasets with a minimum set cover algorithm

One of the main problems of baseline methods is the reliance on per-structure representations, as most modern MLIPs require entire structures (i.e., collections of atoms) to be added to a training set rather than individual environments. In this context, one can easily propose counterexamples for which data subsampling is less efficient due to average, per-structure representations rather than sets of per-atom vectors (see Supplementary Text). To address this problem, we propose a new atomistic dataset subsampling algorithm as an instance of the minimum set cover (MSC) problem. The set cover problem is a classical mathematical formulation defined as follows: given a domain $\mathcal{U}$ and a family of subsets $\mathcal{S}$ of $\mathcal{U}$, a set cover is a subfamily $\mathcal{C} \subseteq \mathcal{S}$ such that the union of all sets in $\mathcal{C}$ equals $\mathcal{U}$.[53] The minimum set coverage corresponds to the smallest subfamily $\mathcal{C}$ that fully covers the domain $\mathcal{U}$. From an atomistic perspective, the domain $\mathcal{U}$ represents the set of all distinct atomic environments $\mathbf{X}_j$ contained in the original dataset, and thus differs slightly from the original MSC problem due to its continuous space. Nevertheless, these environments are grouped into different sets of environments (i.e., structures) $S_i = \{\mathbf{X}\}_j^{(i)}$, and $\mathcal{S} = \{S_1, \ldots, S_N\}$ denotes the initial dataset containing $N$ structures. The sets $S_i$ may have overlapping environments and typically contain different numbers of atoms. Thus, the goal of atomistic dataset compression is analogous to the set cover problem, aiming to identify a subset of structures $\mathcal{C} = \{S_1^{(c)}, \ldots, S_K^{(c)}\} \subseteq \mathcal{S}$, with $K \leq N$, such that the union of all environments in $\mathcal{C}$ span the original universe $\mathcal{U}$.

Though the optimization of the set cover problem is NP-hard,[54] we used a greedy algorithm to approximate the MSC of atomistic datasets in polynomial time.[55] Within this approach, the main rule is to iteratively select sets with the largest number of uncovered elements until the desired dataset size is achieved, thus maximizing the coverage of environments with the minimum number of sets. Using the information-theoretical quantities defined in the Methods, the atomistic version of this algorithm is initialized by first selecting the element with the highest per-structure information entropy, i.e., the structure that exhibits a large number of distinct environments and has the minimum possible redundancy. Then, at each step $n$, we iteratively compute the values of $\delta\mathcal{H}(\mathcal{S} \setminus \mathcal{C}_n \mid \mathcal{C}_n)$, where $\mathcal{C}_n$ is the set of selected structures at iteration $n$. To maximize the coverage, the new structure $S_{n+1}$ is selected according to

$$S_{n+1} = \arg\max_i \left[ \max_{\mathbf{X}_j^{(i)} \in S_i} \delta\mathcal{H}(\mathbf{X}_j^{(i)} | \mathcal{C}_n) + \mathcal{H}(S_i) \right], \tag{9}$$

where $\delta\mathcal{H}$ is computed for all environments $\mathbf{X}_j^{(i)}$ of each structure $S_i \in \mathcal{S} \setminus \mathcal{C}_n$. Finally, $\mathcal{C}_{n+1} = \mathcal{C}_n \cup \{S_{n+1}\}$. The selection rule from Eq. (9) balances two objectives: (1) it prioritizes new structures that contain a large number of uncovered environments according to the environments already present in the compressed dataset ($\max \delta\mathcal{H}$); and (2) it selects structures that are diverse rather than structures exhibiting redundant environments ($\mathcal{H}$). The pseudocode for this proposed method is shown in Algorithm 4.

**Algorithm 4:** Minimum set coverage

**Require:** Set of $N$ structures $\mathcal{S} = \{S_i\}$ with environments $S_i = \{\mathbf{X}_1^{(i)}, \ldots, \mathbf{X}_n^{(i)}\}$, target size $K \leq N$

Calculate the values of per-structure information entropy $\mathcal{H}(S_i)$
Initialize the compressed set $\mathcal{C}_1 = \{\arg\max_i \mathcal{H}(S_i)\}$
**for** $n = 1$ **to** $K - 1$ **do**
    **1.** For each structure $S_i = \{\mathbf{X}_j^{(i)}\}$, compute $\delta\mathcal{H}(\mathbf{X}_j^{(i)} | \{\mathbf{X}\}_n^{(c)})$, where $\{\mathbf{X}\}_n^{(c)}$ are all envs. in $\mathcal{C}_n$
    **2.** Choose $S_{n+1}^{(c)}$ according to Eq. (9)
    **3.** Define $\mathcal{C}_{n+1} = \mathcal{C}_n \cup \{S_{n+1}^{(c)}\}$
**end for**
**return** $\mathcal{C} = \{S_1^{(c)}, \ldots, S_K^{(c)}\}$

## 2.5. Machine learning interatomic potential

**Model training approach:** To evaluate the effect of sampling method on MLIP performance, we trained models using subsampled datasets containing 10%, 25%, 50%, 75%, and 100% of the original data. Each model was trained for a fixed total number of structure-epochs to compare performances across dataset sizes. Accordingly, models trained on 10%, 25%, 50%, 75%, and 100% of the data were run for 10,000, 4,000, 2,000, 1,333, and 1,000 epochs, respectively. Across all the dataset sizes, a batch size of 40 was used, except for the 10% dataset size, where a batch size of 4 was used. We randomly split each subsampled dataset at ratios of 80:10:10 for train:validation:test, where a single test set was shared among all experiments. The best performing model at each training run was determined as the one minimizing the validation loss.

**SevenNet architecture:** The MLIP used in this work was the SevenNet architecture from Park *et al.*,[11] which is based on the NequIP architecture.[9] We used the SevenNet code available at https://github.com/MDIL-SNU/SevenNet (v. 0.2.0). The model was created with three equivariant blocks with $L = 2$ and node features with irreducible representations of 32 scalars ($\ell = 0$), 32 vectors ($\ell = 1$), and 32 tensors ($\ell = 2$). The convolutional filter used a cutoff radius $r_c$ of 5 Å and a tensor product of learnable radial functions with 8 radial Bessel functions and spherical harmonics up to $\ell = 2$.

**SevenNet training:** The SevenNet models were trained with the Adam optimizer,[56,57] starting with a learning rate of 0.005, which decayed by a factor of 0.99 per epoch. We used the Huber loss function[58] as defined in the SevenNet paper.[11] The weight of the loss due to errors in forces was selected as 0.1, whereas the weight due to stresses was set to $10^{-6}$. The value of $\delta$ in the Huber loss that controls

the transition between the mean squared error loss to a mean absolute error[58] was set to 0.01. The dataset was shifted according to the mean per-atom energy and rescaled according to the root mean square of the forces, statistics that were obtained from each training dataset.

# 3. Results

## 3.1. Comparing compression algorithms with the GAP-20 carbon dataset

Using the figures of merit and methods defined above, we compared the performance of our information-theoretical algorithm from Sec. 2.4 against the baseline ones by compressing three subsets of the GAP-20 carbon dataset from Rowe *et al.*[32] In our previous work, we showed that the subsets of this dataset exhibit varying degrees of redundancy, leading to distinct behaviors upon random sampling.[25] Here, we investigated whether improved sampling efficiency could be achieved when different algorithms are used to subsample the data. Figures **2**a–c compare the performance of all four algorithms in compressing the Graphene, Nanotubes, and Fullerenes subsets of GAP-20. In all metrics and datasets, our MSC method outperforms all other baseline algorithms in compression efficiency. Figure **2**a shows that MSC preserves the diversity of the dataset as much as possible upon compression, while other baseline decrease data diversity at lower compression rates than MSC. The figure also shows that MSC leads to the highest increase in entropy of the subsampled dataset across all algorithms, demonstrating that the method efficiently removes redundancies in the data prior to information loss. Figures **2**b,c corroborate these results, showing that MSC-compressed datasets have the highest overlap with the original dataset and the longest tail of the distribution of forces (see also Fig. S1). The only exception to this result is the Fullerenes subset of GAP-20 compressed to 10% of its original size, where all algorithms exhibit large information loss, including MSC (see analysis in Sec. 3.2). This is partially a consequence of the overall higher dataset efficiency of Fullerenes compared to Nanotubes or Graphene (Fig. **2**d), for which higher data compression implies information loss.

The results above show that it is possible to measure information loss in subsampled datasets without training MLIPs. This model-free approach is useful, as it avoids expensive model training runs and provides a algorithms independent of model architecture. Furthermore, these metrics provide a way to define an optimal compression level instead of relying on error scans. Nevertheless, it is still relevant to measure whether MLIPs trained on subsampled datasets exhibit a reasonable range of test errors despite being trained on a small data regime. Using the SevenNet architecture[11] as described in the Methods, we trained models on compressed datasets from each algorithm and size, and evaluated them on a common test set to measure the behavior of force errors when data compression is achieved. Figure **2**e compares these errors for all models and data sizes by subtracting the minimum error from each dataset type and size across algorithms. In this case, points closer to zero would indicate that the algorithm that created the subsampled dataset leads to models with lower errors. On the other hand, higher errors indicate that the algorithm often leads to training sets that degrade model performance on the test set. As shown on Fig. **2**e, no algorithm systematically outperforms the others when it comes to the lowest possible error, which can be partly explained by the variability in the model training process and the specific test datasets used to quantify error. However, MLIPs trained on datasets compressed with MSC are more likely to exhibit lower errors compared to MLIPs trained on datasets compressed with other algorithms, as seen by smaller ranges of the distribution of errors under MSC. This result is particularly pronounced at the very low data regimes (10%), where MSC achieves the lowest possible errors among the three subsets of the GAP-20 dataset. This demonstrates that dataset compression methods can have a large impact in model

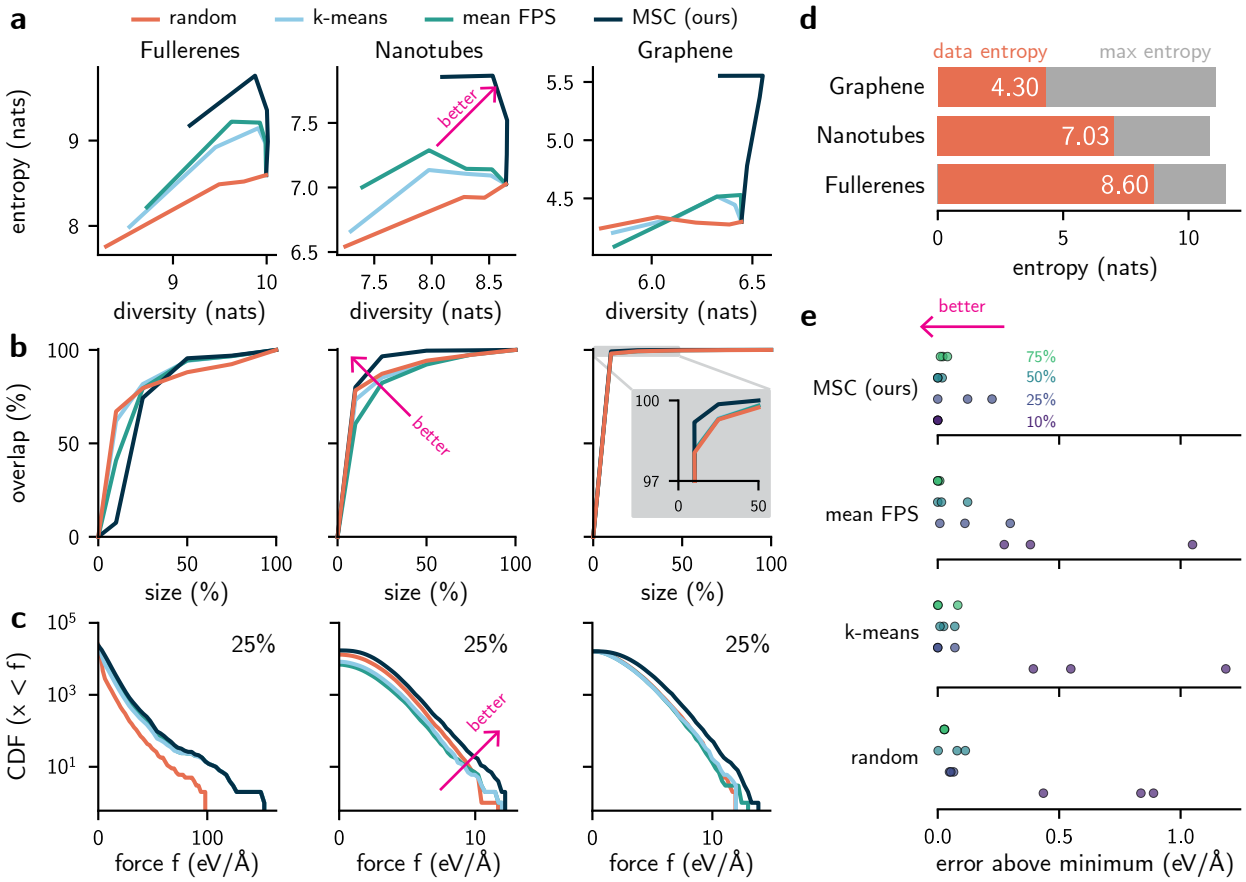


**Fig. 2**: Performance of compression algorithms in subsampling three subsets of the GAP-20 dataset (Fullerenes, Nanotubes, Graphene). Each dataset was subsampled at four different sizes: 75, 50, 25, and 10%, and compared against the full, original dataset. **a**, our MSC method exhibits the best behavior in terms of simultaneous entropy increases and diversity retention for the selected datasets across dataset sizes. On the other hand, random sampling quickly leads to information loss, as shown by immediate diversity decreases upon subsampling. **b**, datasets compressed with MSC exhibit the higher overlap with the original dataset, with a single exception of Fullerenes sampled at 10% of its original size. **c**, MSC preserves the long tail of the distribution of forces of the environments, as shown by a higher cumulative distribution function (CDF) given a force threshold for datasets compressed at 25% of their original size (see also Fig. S1). **d**, Information entropy (orange) of the three subsets of GAP-20 and the maximum value of entropy that would be possible in a dataset with that size (gray). The numerical values of entropy (in nats) are shown with white numbers. The differences between the gray and orange bars show that the subsets exhibit different levels of redundancy, explaining the results in **a**–**c**. **e**, Average force error (eV/Å) above the minimum value of error given a subset (i.e., Graphene, Nanotubes, or Fullerenes) and a dataset size (i.e., 10, 25, 50, or 75%) across algorithms. While no algorithm consistently outperforms the others, models trained on datasets compressed using our MSC method exhibit a smaller range of errors above the minimum across sampling sizes, with errors more tightly grouped around optimal performance compared to other methods.

generalization when datasets are undersampled (low-data regime), while preserving the original distribution of points and errors when the data redundancy is high. All numerical results supporting these conclusions are shown in the Supplementary Information (Tables S1-S4).

### 3.2. Explaining information loss upon dataset compression

To explain the results from the previous section, we investigated how the algorithms lead to information loss when compressing datasets. However, instead of relying on the aggregate dataset metrics from Figs. 2a–c, we analyzed individual environments in each dataset and tested their impact in the model errors shown in Fig. 2e. Figure 3a illustrates the distribution of $\delta\mathcal{H}(\mathbf{X}_i|\mathcal{C})$ as a function of compression level and algorithm, where $\mathbf{X}_i$ are all environments in the original, complete dataset. Given that $\delta\mathcal{H}$ measures how close each environment $\mathbf{X}_i$ is to the distribution of environments in the compressed dataset $\mathcal{C}$, an ideal algorithm would lead to as many negative $\delta\mathcal{H}$ values as possible. In all but one case in Fig. 3a, datasets compressed with MSC exhibit $\delta\mathcal{H}$ distributions with fewer positive values, indicating that the compressed datasets indeed cover as much as possible of the original dataset. On the other hand, the baseline algorithms almost always lead to loss of information, as seen by distributions with high values of $\delta\mathcal{H}$. The only exception to this scenario is the Fullerenes subset of GAP-20 compressed to 10% of its original size. To explain this discrepancy, we note that the MSC algorithm attempted to minimize the number of environments very far away from the original distribution, as illustrated by the number of data points with $\delta\mathcal{H} > 10$ nats in the plot. In the process, the subsampled structures end up reasonably close to the original dataset (see range of $0 < \delta\mathcal{H} < 10$ nats for 10% Fullerenes), creating a dataset with low overlap with the original, but fewer true outliers as well. These results refine the conclusions from the previous section. When the errors from Fig. 2e are broken down by environment rather than treated as average results, it becomes clear that MSC reduces the number of data points treated as outliers and, as a consequence, constrain generalization errors from the MLIPs. Figures S2–S4 further show that, by reducing the number of environments from the original dataset that become outliers (i.e., exhibit large $\delta\mathcal{H}$ values) in the compressed data, models are less likely to operate under "out-of-distribution" scenarios ($\delta\mathcal{H} > 0$).

As an alternative to our $\delta\mathcal{H}$ analysis, dataset subsampling is often examined in the literature using dimensionality reduction techniques. In such approaches, a 2D projection of the subsampled dataset is overlaid on top of that of the original dataset, and the overlap between the two distributions is qualitatively analyzed. To evaluate whether this visualization method could provide insights into the information loss caused by subsampling similarly to $\delta\mathcal{H}$, we used a visualization created with UMAP[59,60] to compare datasets subsampled with different algorithms. The UMAP transformation was fitted to the distribution of QUESTS descriptors from the original, full subsets of the GAP-20 dataset and then applied consistently to all subsampled datasets. Figure 3b illustrates how the 2D distribution of points for the Nanotubes subset changes as we reduce the dataset size from 75% to 10% (see all results in Figs. S5–S7). In contrast with the the metrics in Fig. 2 or the outlier analysis in Fig. 3a, this 2D visualization is unable to quantify information losses in the data. It can be misleading to interpret the "blank areas" within the main region of the 2D UMAP plots at 10% size as a sign of higher "information losses". If that assumption were true, it could suggest that random sampling is the best algorithm, as it seemingly covers more of the two-dimensional space represented with the UMAP visualization. As shown before, the actual coverage is minimized with the random sampling algorithm, which exhibits the worst information loss when subsampling atomistic datasets. Similar results can be found when comparing all 2D visualizations for the Graphene and Fullerene datasets (Figs. S5,S7), where the 2D distribution of data points is less representative of the true dataset coverage than an analysis in higher-dimensional spaces. This provides a cautionary tale on the use of low-dimensional representations when compressing datasets, supporting a more quantitative measure of distributions and coverage of local environments using information theory.

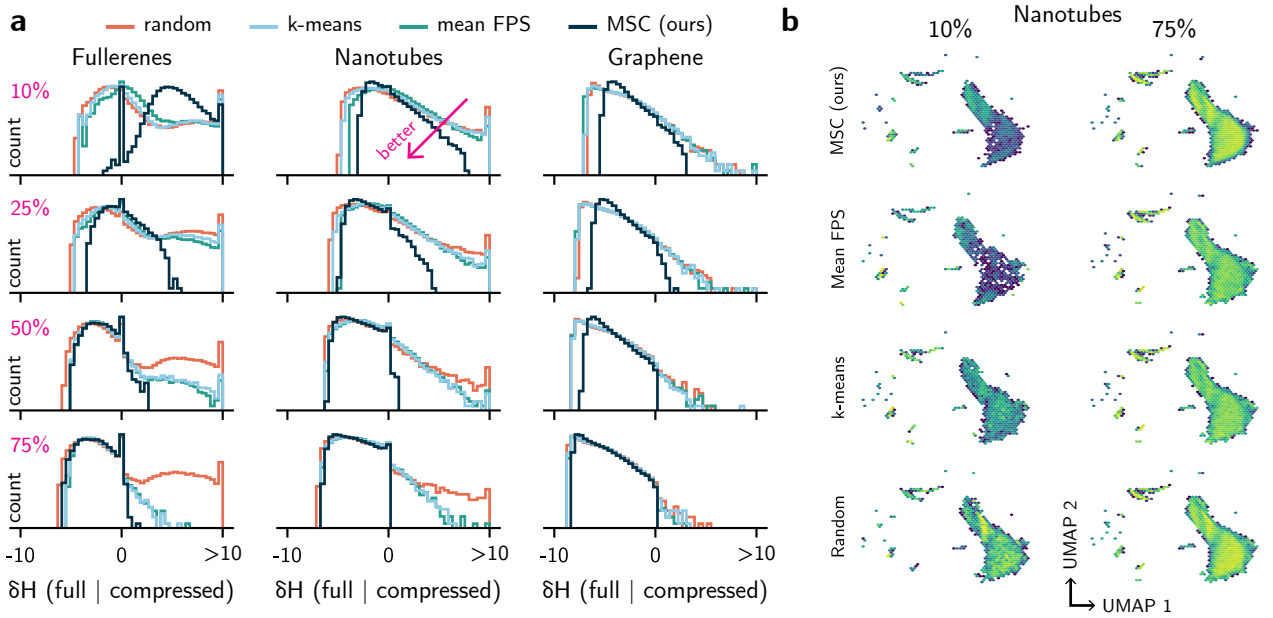


**Fig. 3**: Analysis of outlier loss across algorithms and dataset sizes for the three subsets of GAP-20 under study (Graphene, Nanotubes, and Fullerenes). **a**, the distribution of $\delta\mathcal{H}$ of environments in the full dataset w.r.t the compressed dataset across different methods indicates that MSC minimizes outlier loss, as shown by distributions of $\delta\mathcal{H}$ shifted towards less positive values. The only exception to this is the case of Fullerenes compressed at 10% of its original size, where the algorithm prioritized reducing the number of extreme outliers (i.e., those with $\delta\mathcal{H} > 10$ nats) at the expense of a reduction of true overlap ($\delta\mathcal{H} \leq 0$) with the dataset. **b**, a low-dimensional representation of the datasets, obtained with a UMAP visualization, illustrates how outlier loss is not trivially detected with these methods. When comparing the distribution of environments within the 10% and 75% datasets of Nanotubes, the information loss is at best qualitative, and at worst misleading. On all UMAP plots, the axes are the arbitrary axes of UMAP and brighter colors indicate a larger number of points within each region.

### 3.3. Comparing compression algorithms with the TM23 dataset

As a second case study, we analyzed the compression of the TM23 dataset from Owen *et al.*[34] This dataset exhibits several PESes, but here we focus on seven elements whose errors are representative of the original TM23 analysis: Ag, Au, Cd, Co, Ir, Pd, and Ti, all within the "warm" subset of the dataset (i.e., sampled at 75% of the melting point of each metal). Specifically, results for four elements are shown in the main text: Ag (very low test errors of MLIPs), Ir (low errors), Co (medium errors), and Ti (high errors), with the remaining data shown in the Supplementary Information. Figures 4a–c show that MSC outperforms all other algorithms in preserving dataset diversity, increasing entropy, maintaining higher overlap with the original distribution, and retaining the long tail behavior of the distribution of forces (see also Figs. S8–S10, Tables S5–S11). Interestingly, the datasets exhibited different behavior under compression. Ag and Ir have large redundancies in the data (see also Fig. 4d), as demonstrated by overlaps close to 100% across dataset sizes. Even dataset sizes as small as 10% retain most of the original information in the data, with high data diversity and overlaps around 99% for Ag and Ir across algorithms (see discussion on the diversity increases in the Supplementary Text). On the other hand, Ti and Co rapidly lose information upon compression, reaching 90% and 20% overlaps with the original dataset, respectively, when compressed to 10% of their original sizes.

The smaller compressibility of Co is explained by the small redundancy in the original data (Fig. 4d). Whereas a slight entropy increase can be seen when compressing Ti (Fig. 4a), any subsampling in Co leads to information loss.

To evaluate the performance of MLIPs on the compressed datasets, we trained SevenNet models to each of the compressed subsets of TM23 and evaluated their error on shared, held-out test sets. As seen with the GAP-20 dataset, errors of models trained on MSC-compressed datasets are often constrained compared to the other algorithms, especially in low-data regimes (Tables S12–S18). Similarly to Fig. 3a, we observed that our MSC method leads to less outliers compared to the other algorithms even for the Co dataset (Figs. S11–S17), highlighting the importance of minimizing information loss and extreme outliers for MLIPs. Though variability in model training and errors can make it challenging to compare errors across data sizes and subsampled datasets, our information-theoretical algorithm provides a robust method to compress atomistic datasets. On the other hand, 2D visualizations created with UMAP are unable to provide quantitative insights on the dataset compression (Figs. S18–S24).

### 3.4. Compression of 64 single-element datasets from the ColabFit repository

Beyond the examples above, we evaluated all algorithms in this work on a larger and more diverse collection of datasets used for training MLIPs. Specifically, we selected 64 single-element datasets from the ColabFit repository[35] spanning a range of materials, structures, and phases (see Table S19 for full dataset names). We compressed each of the datasets by the same ratios adopted for the other examples in this work (10, 25, 50, and 75%) and compared the resulting performance metrics. Figure 5 contrasts the diversity (Fig. 5a) and overlap (Fig. 5b) of datasets compressed with MSC against the other algorithms. The negative shifts in distributions show that MSC consistently outperforms all other algorithms across nearly all 64 datasets. With a few exceptions, the diversity of MSC-compressed datasets is substantially higher than that obtained from other algorithms (see full results in Figs. S25–S28 and Tables S20–S83), and overlaps can be up to 30% higher (and 60% higher in one case) compared to other methods. Similar to what was observed in the Fullerenes subset of GAP-20 (Fig. 3a), variability in overlap is observed when the datasets are compressed to 10% of their original size even though data diversity is preserved (Figs. S29–S32). Nevertheless, these results demonstrate that our MSC method is a robust dataset compression technique and often outperforms other algorithms in reducing dataset sizes while also minimizing information loss.

## 4. Discussion

By interpreting the problem of subsampling atomistic datasets as a set cover problem, we demonstrated that a greedy algorithm can be used to obtain a minimum set cover and reduce loss of information of atomistic datasets upon compression. These results rely on both the information-theoretical formalism used to identify high-novelty data points as well as the avoidance of per-structure representations used in typical baseline algorithms. In addition to the algorithm, we proposed model-free figures of merit to quantify the efficiency of the compression algorithms across datasets and showed that our method outperforms baseline algorithms in the majority of cases.

Multiple results in this work invite additional discussion. First, when MLIPs were trained on compressed datasets, their performance was variable, but errors mostly increased compared to training MLIPs to the full dataset, as expected. Nevertheless, it is possible that subsampling algorithms could lead to different "learning curves" for several datasets, decreasing, on average, errors at the

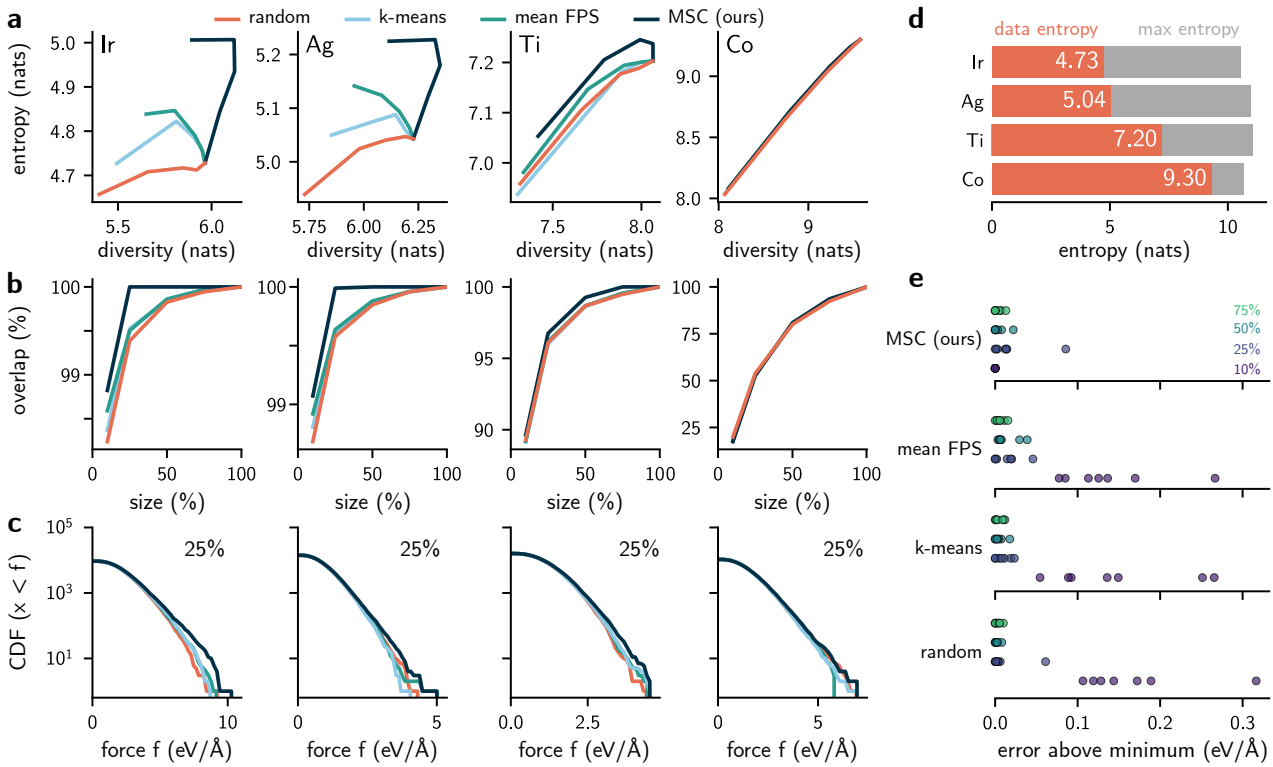


**Fig. 4**: Performance of subsampling algorithms across a subset of the TM23 dataset (elements Ir, Ag, Ti, Co). The datasets are subsampled 75, 50, 25, and 10% of their original sizes, and compared against the original datasets. **a**, our MSC method exhibits the best behavior in terms of simultaneous entropy increases and diversity retention for the subsets of TM23 across dataset sizes. In cases where there is lower data redundancy, such as the case of Ti or Co, information loss is inevitable. **b**, datasets compressed with MSC often exhibit the higher overlap with the original dataset compared to those obtained with other algorithms, except when redundancy in the original data is low, in which case all overlaps are similar. **c**, MSC preserves the long tail of the distribution of forces of the environments, as shown by a higher cumulative density function (CDF) given a force threshold $f$. **d**, Information entropy (orange) of the four subsets of TM23 and the maximum value of entropy that would be possible in a dataset with that size (gray). The numerical values of entropy are shown with white numbers, where the entropy is expressed in units of nats. The differences between the gray and orange bars show that the subsets contain different redundancy levels, explaining trends across elements in **a**–**c**. **e**, Average force error (eV/Å) above the minimum value of error given a subset (i.e., Ir, Ag, Ti, or Co) and a dataset size (i.e., 10, 25, 50, or 75%). Similarly to the GAP-20 example, while no algorithm consistently outperforms the other in terms of force errors, models trained on datasets compressed using our MSC method exhibit a smaller range of errors above the minimum across sampling sizes, with errors more tightly grouped around optimal performance in the low-data regime.

low-data regime. This result, however, may depend on biases on test sets, which are often randomly sampled from the full dataset. In some cases, specific test samples may contain more or less outliers, increasing the variance in test errors without accounting for the "fairness" of the test distribution. For instance, recent work has showed that variability in dataset construction and model training can lead to robust models even with random sampling of datasets.[61] While a full analysis of test distributions has been the scope of other studies,[62] this work provides a tool that bypasses the use of test errors as sole figure-of-merit to evaluate data compression. Moreover, eliminating data redundancies

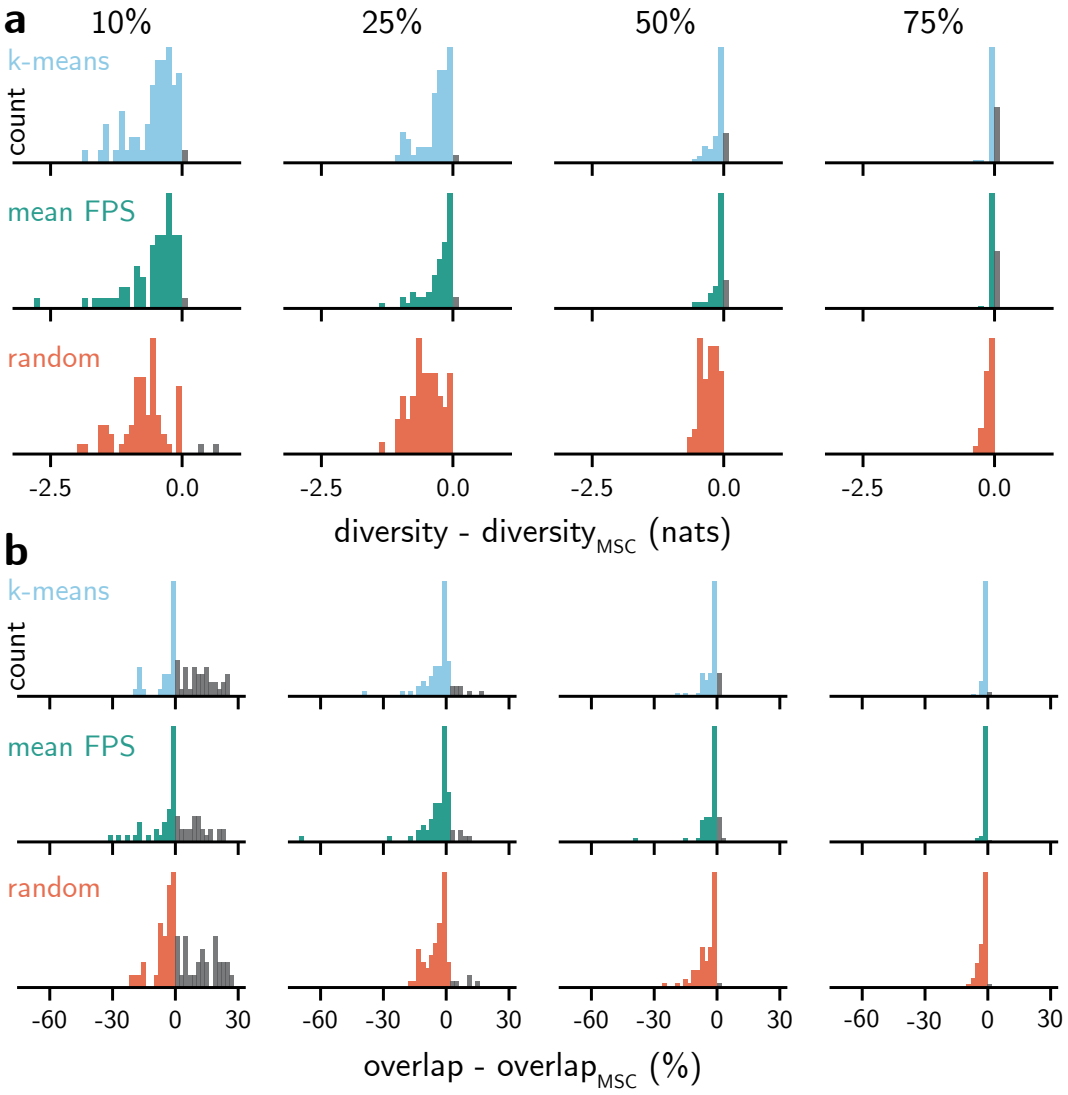


**Fig. 5**: Performance of baseline algorithms against MSC across 64 different atomistic datasets from the ColabFit repository compressed to 10, 25, 50, and 75% of their original sizes. Gray bars represent datasets for which the performance of MSC is worse than the baseline algorithm according to each metric. **a**, The difference in diversity between each method and MSC is mostly negative across all sizes and algorithms, demonstrating that MSC is a robust algorithm to preserve the original data distribution. **b**, Similarly, baseline algorithms lead to datasets with lower overlap against the original data compared to datasets subsampled with MSC. In the low-data regime, some algorithms showcase a higher overlap even though their diversity is lower. Similar to the Fullerenes example in Fig. 3, overlap does not fully account for how close outliers are to the distribution of points, but only for which points are entirely contained by a distribution.

is not only useful to reduce MLIP training costs, but also to augment training sets with new data, as is performed in active learning loops. For instance, instead of subsampling molecular dynamics trajectories at constant time steps, which may overlook outlier frames, our MSC algorithm can be used to robustly select frames that are most diverse from each other, maximizing the amount of information that can be used to build a new generation of training data points.

At the large-data regime, despite the impressive ability of universal NNIPs to simulate materials across the periodic table, recent works have shown that their degree of generalization (as measured by the $\delta\mathcal{H}$ of each environment with respect to the training data) is correlated with test errors.[25,63] Given this strong correlation between $\delta\mathcal{H}$ and errors, ensuring that MLIPs avoid extreme cases of out-of-distribution inference requires minimizing the "surprise" of outliers given known training datasets. We showed that our algorithm reduces the loss of extreme outliers even in cases where information loss is inevitable, such as in the TM23 dataset. While this work focuses on single-component systems, similar approaches can be performed for multi-component materials, especially if learned representations can be used for each environment. The definitions of entropy, diversity, and $\delta\mathcal{H}$ are agnostic to the choice of representation, and $\delta\mathcal{H}$ was demonstrated to work efficiently with multi-component systems even with the composition-agnostic QUESTS descriptor.[25]

Interestingly, we showed that dimensionality reduction techniques can be insufficient to demonstrate data compression. While qualitative behavior on dataset coverage can be sometimes extracted from these graphics, we provided examples where outlier loss is far from obvious in a UMAP plot. On the other hand, our information-theoretical metrics can effectively assess information loss, outlier behavior, and dataset compressibility, allowing a quantitative treatment of dataset subsampling while also being model-free.

One drawback of our MSC algorithm is the increased time complexity to run the algorithm. While most of the datasets in this work were compressed using a personal laptop in a matter of minutes, datasets with $\mathcal{O}(10^5) - \mathcal{O}(10^6)$ environments may require a few CPU-hours to be compressed using our method. The software implementation in the QUESTS package provides GPU acceleration, but an information-theoretical approach to data compression will always be more costly than randomly sampling structures. Some dedicated production applications may strongly benefit from our MSC approach, especially for limiting outlier loss. Nevertheless, baseline algorithms such as $k$-means sampling are scalable, simple to implement, and remain an improvement over random sampling.

## 5. Conclusions

In summary, we proposed a strategy to subsample datasets for machine learning interatomic potentials with minimal information loss using information theory. We started with the minimum set cover (MSC) problem definition in combinatorics, which tries to find the minimum number of sets that covers the entire universe of elements belonging to any of these sets, and applied their solutions to atomistic data. Then, we suggested that selecting a minimum number of structures that maximally cover the space of atom-centered environments is analogous to solving the MSC problem. Within this formalism, we created an algorithm that avoids information losses when subsampling an atomistic dataset compared to other baseline algorithms. These results were quantified by multiple figures of merit, including dataset entropy-diversity trade-offs, overlap with the original distribution of data points, and long tail of the force distributions. When compressed datasets were used to train MLIPs, models trained on MSC-compressed datasets were less prone to "out-of-distribution" predictions compared to models trained on datasets compressed with other algorithms. These results were extensively discussed in the case of GAP-20 and TM23, and remained consistent when a larger analysis was performed for 64 other atomistic datasets, demonstrating the robustness of our MSC method. All algorithms are implemented in and available as part of the QUESTS package at `https://github.com/dskoda/quests` and can be used in a wide range of applications, such as subsampling molecular dynamics trajectories, compressing atomistic datasets for MLIP training,

performing active learning loops, and more.

## Data Availability

The datasets used for training/testing ML potentials were obtained from the original sources at:

- GAP-20: https://doi.org/10.17863/CAM.54529
- TM23: https://doi.org/10.24435/materialscloud:6c-b3

The 64 datasets used for the large-scale analysis of the compression algorithms were obtained from the ColabFit repository. The references and DOIs of the datasets are provided in Table S19 in the Supplementary Information.

The raw data and analysis/plotting code necessary to reproduce all figures and results of this work are available on GitHub at https://github.com/digital-synthesis-lab/2025-compression-data, with persistent data storage on Zenodo under the DOIs: 10.5281/zenodo.17536234 and 10.5281/zenodo.17602768. Most of the raw data from this work is also available as tables in the the Supplementary Information.

## Code Availability

The code for QUESTS is available on GitHub at the link https://github.com/dskoda/quests. The version of the code used in this work (v.2025.09.29) is deposited on Zenodo for persistent storage under the DOI 10.5281/zenodo.17229448.

## Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award Number DE-SC0025642. V.L.'s contributions were performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, funded by the Laboratory Directed Research and Development Program at LLNL under project tracking code 23-SI-006. Manuscript released under the tracking code LLNL-JRNL-2013284-DRAFT. The authors acknowledge helpful discussions with Joshua Vita.

## Conflicts of Interest

The authors have no conflicts to disclose.

## Author Contributions

**Benjamin Yu:** Formal Analysis; Investigation; Software; Validation; Visualization; Writing - Original Draft; Writing - Review & Editing. **Vincenzo Lordi:** Conceptualization; Writing - Review & Editing; Funding Acquisition. **Daniel Schwalbe-Koda:** Conceptualization; Formal Analysis; Investigation; Methodology; Project Administration; Software; Validation; Visualization; Writing - Original Draft; Writing - Review & Editing; Funding Acquisition; Supervision.

## References


[1] Behler, J. & Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters* **98**, 146401 (2007).

[2] Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters* **104**, 136403 (2010).

[3] Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics* **285**, 316–330 (2015).

[4] Chmiela, S. *et al.* Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **3**, e1603015 (2017).

[5] Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical Review Letters* **120**, 143001 (2018).

[6] Mueller, T., Hernandez, A. & Wang, C. Machine learning for interatomic potential models. *The Journal of Chemical Physics* **152**, 050902 (2020).

[7] Manzhos, S. & Carrington, T. Neural Network Potential Energy Surfaces for Small Molecules and Reactions. *Chemical Reviews* **121**, 10187–10217 (2021).

[8] Unke, O. T. *et al.* Machine Learning Force Fields. *Chemical Reviews* **121**, 10142–10186 (2021).

[9] Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications* **13**, 2453 (2022).

[10] Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. In Koyejo, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 35, 11423–11436 (Curran Associates, Inc., 2022).

[11] Park, Y., Kim, J., Hwang, S. & Han, S. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation* **20**, 4857–4868 (2024).

[12] Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).

[13] Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science* **2**, 718–728 (2022).

[14] Batatia, I. *et al.* A foundation model for atomistic materials chemistry. *arXiv:2401.00096* (2023).

[15] Deng, B. *et al.* CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence* **5**, 1031–1041 (2023).

[16] Yang, H. *et al.* MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures. *arXiv:2405.04967* (2024).

[17] Karabin, M. & Perez, D. An entropy-maximization approach to automated training set generation for interatomic potentials. *The Journal of Chemical Physics* **153**, 094110 (2020).

[18] Smith, J. S. *et al.* Automated discovery of a robust interatomic potential for aluminum. *Nature Communications* **12**, 1257 (2021).

[19] Schwalbe-Koda, D., Tan, A. R. & Gómez-Bombarelli, R. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nature Communications* **12**, 5104 (2021).

[20] de Oca Zapiain, D. M. *et al.* Training data selection for accuracy and transferability of interatomic potentials. *npj Computational Materials* **8**, 189 (2022).

[21] Allen, C. & Bartók, A. P. Optimal data generation for machine learned interatomic potentials. *Machine Learning: Science and Technology* **3**, 045031 (2022).

[22] Kulichenko, M. *et al.* Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature Computational Science* **3**, 230–239 (2023).

[23] Barroso-Luque, L. *et al.* Open Materials 2024 (OMat24) inorganic materials dataset and models. *arXiv:2410.12771* (2024).

[24] Kaplan, A. D. *et al.* A foundational potential energy surface dataset for materials. *arXiv:2503.04070* (2025).

[25] Schwalbe-Koda, D., Hamel, S., Sadigh, B., Zhou, F. & Lordi, V. Model-free estimation of completeness, uncertainties, and outliers in atomistic machine learning using information theory. *Nature Communications* **16**, 4014 (2025).

[26] Tan, A. R., Dietschreit, J. C. & Gómez-Bombarelli, R. Enhanced sampling of robust molecular datasets with uncertainty-based collective variables. *The Journal of Chemical Physics* **162**, 034114 (2025).

[27] Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling and Simulation* **14**, 1153–1173 (2016).

[28] Imbalzano, G. *et al.* Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *The Journal of Chemical Physics* **148**, 241730 (2018).

[29] Kulichenko, M. *et al.* Data generation for machine learning interatomic potentials and beyond. *Chemical Reviews* **124**, 13681–13714 (2024).

[30] Sun, H., Vita, J. A., Samanta, A. & Lordi, V. Unsupervised atomic data mining via multi-kernel graph autoencoders for machine learning force fields. *arXiv:2509.12358* (2025).

[31] Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Science Advances* **3**, e1701816 (2017).

[32] Rowe, P., Deringer, V. L., Gasparotto, P., Csányi, G. & Michaelides, A. An accurate and transferable machine learning potential for carbon. *The Journal of Chemical Physics* **153**, 034702 (2020).

[33] Qi, J., Ko, T. W., Wood, B. C., Pham, T. A. & Ong, S. P. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. *npj Computational Materials* **10**, 43 (2024).

[34] Owen, C. J. *et al.* Complexity of many-body interactions in transition metals via machine-learned force fields from the TM23 data set. *npj Computational Materials* **10**, 92 (2024).

[35] Vita, J. A. *et al.* ColabFit Exchange: Open-access datasets for data-driven interatomic potentials. *The Journal of Chemical Physics* **159**, 154802 (2023).

[36] Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).

[37] Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *The Journal of Physical Chemistry C* **121**, 511–522 (2017).

[38] Williams, L., Sargsyan, K., Rohskopf, A. & Najm, H. N. Active learning for snap interatomic potentials via bayesian predictive uncertainty. *Computational Materials Science* **242**, 113074 (2024).

[39] Lebeda, M., Drahokoupil, J., Lobel, L. & Vlčák, P. k-means clustering in fingerprint-based configuration selection for fitting interatomic potentials. *Journal of Chemical Theory and Computation* **20**, 10676–10683 (2024).

[40] Sivaraman, G. *et al.* Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Computational Materials* **6**, 104 (2020).

[41] Eldar, Y., Lindenbaum, M., Porat, M. & Zeevi, Y. Y. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing* **6**, 1305–1315 (1997).

[42] Wengert, S., Csányi, G., Reuter, K. & Margraf, J. T. A hybrid machine learning approach for structure stability prediction in molecular co-crystal screenings. *Journal of Chemical Theory and Computation* **18**, 4586–4593 (2022).

[43] Boulangeot, N., Brix, F., Sur, F. & Gaudry, É. Hydrogen, Oxygen, and Lead Adsorbates on Al13Co4 (100): Accurate Potential Energy Surfaces at Low Computational Cost by Machine Learning and DFT-Based Data. *Journal of Chemical Theory and Computation* **20**, 7287–7299 (2024).

[44] Li, R. *et al.* Local-environment-guided selection of atomic structures for the development of machine-learning potentials. *The Journal of Chemical Physics* **160**, 074109 (2024).

[45] Zhang, L. *et al.* Exploring the energy landscape of aluminas through machine learning interatomic potential. *Physical Review Materials* **9**, 023801 (2025).

[46] Gurlek, B., Sharma, S., Lazzaroni, P., Rubio, A. & Rossi, M. Accurate machine learning interatomic potentials for polyacene molecular crystals: application to single molecule host-guest systems. *npj Computational Materials* **11**, 318 (2025).

[47] Trestman, M., Gugler, S., Faber, F. A. & von Lilienfeld, O. A. Gradient-Guided Furthest Point Sampling for Robust Training Set Selection. *arXiv:2510.08906* (2025).

[48] Mahoney, M. W. & Drineas, P. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences* **106**, 697–702 (2009).

[49] Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Physical Review B* **95**, 094203 (2017).

[50] Dragoni, D., Daff, T. D., Csányi, G. & Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Physical Review Materials* **2**, 013808 (2018).

[51] Bernstein, N., Csányi, G. & Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Computational Materials* **5**, 99 (2019).

[52] Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Physical Review X* **8**, 041048 (2018).

[53] Korte, B. & Vygen, J. *Combinatorial Optimization: Theory and Algorithms*, vol. 21 of *Algorithms and Combinatorics* (Springer, Berlin, Heidelberg, 2018).

[54] Lund, C. & Yannakakis, M. On the hardness of approximating minimization problems. *J. ACM* **41**, 960–981 (1994).

[55] Chvatal, V. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research* **4**, 233–235 (1979).

[56] Kingma, D. & Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)* (San Diego, CA, USA, 2015).

[57] Reddi, S. J., Kale, S. & Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations* (2018). URL https://openreview.net/forum?id=ryQu7f-RZ.

[58] Huber, P. J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* **35**, 73–101 (1964).

[59] McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).

[60] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426* (2020).

[61] Stolte, N., Daru, J., Forbert, H., Marx, D. & Behler, J. Random Sampling Versus Active Learning Algorithms for Machine Learning Potentials of Quantum Liquid Water. *Journal of Chemical Theory and Computation* **21**, 886–899 (2025).

[62] Vita, J. A. & Schwalbe-Koda, D. Data efficiency and extrapolation trends in neural network interatomic potentials. *Machine Learning: Science and Technology* **4**, 035031 (2023).

[63] Chiang, Y. *et al.* MLIP Arena: Advancing Fairness and Transparency in Machine Learning Interatomic Potentials via an Open, Accessible Benchmark Platform. *arXiv:2509.20630* (2025).

# Supplementary Information for: Maximizing Efficiency of Dataset Compression for Machine Learning Potentials With Information Theory

**Benjamin Yu,**[1] **Vincenzo Lordi,**[2] **Daniel Schwalbe-Koda**[1*]

[1]Department of Materials Science and Engineering, University of California, Los Angeles, CA, United States
[2]Materials Science Division, Lawrence Livermore National Laboratory, CA, United States

## S1. Supplementary Text

### S1.1. Example of suboptimal dataset compression with per-structure representations

Assume that a dataset to be compressed contains two structures representing the same relaxed, bulk crystal with a vacancy, but one of the structures has 255 atoms and the other one has 511 atoms. These otherwise identical structures contain mostly similar atomic environments, including the few atoms around the vacancy that deviate from bulk-like environments, even though they should exhibit small differences in electronic structures due to the concentration of defects. As a result, the two per-structure representations — obtained from the average representations of each cell — will necessarily be similar because of the large number of atoms of both structures. According to their per-structure representation, therefore, these two structures are nearly indistinguishable, and may be selected almost interchangeably if subsampled from a diverse dataset. However, the 511-atom supercell has a significantly higher memory overhead compared to the 255-atom supercell even though both carry the same amount of information. From a dataset compression perspective, therefore, it is desirable to take into account smaller supercells that cover the same space of atomic environments.

An alternative example regarding the problem above would be to consider a manual selection rule, where the structure containing the smallest number of environments is selected for each neighborhood of the high-dimensional representation space. In the example above, this rule would lead to the selection of the 255-atom supercell rather than the 511-atom supercell. However, given this rule, one can also propose a counterexample that leads to outlier loss. For instance, take a small cell representing the idealized bulk structure rather than a defected cell. Because of the large size of the supercell and the fact that the defect may be a small fraction of the cell, the per-structure representation of the small cell is likely close to the per-structure representation of the defected cell. According to the manual selection rule, the information of the defect may be lost simply because it is averaged out in a per-structure representation.

### S1.2. Diversity increases upon dataset compression

Figure 4a of the main text showcases examples where the diversity of a dataset increases upon compression. However, within a strict definition, the diversity of a dataset should not decrease as the dataset size increases. The definition of diversity in Eq. (7) was proposed in our previous work to recover two limiting cases: (1) when the dataset is a degenerate case where all environments are

*E-mail: dskoda@ucla.edu

identical, the diversity should be 0 nats; and (2) when the dataset is a degenerate case where all environments are distinct, the diversity should be equal to $\log n$, where $n$ is the number of environments. Moreover, the current definition of diversity approximates the non-decreasing behavior of a true diversity metric, but does not guarantee it. Especially in near-degenerate cases where there is large redundancy in the data, as is the case of the Ag and Ir subsets of TM23, an increase in data redundancy can slightly decrease the dataset diversity, as the support of the data distribution cannot be easily measured in the continuous space of atomic environments. Therefore, the diversity in Eq. (7) is an approximation of dataset diversity rather than an actual measure of the support of the distribution. Future work can focus on proposing more rigorous measures of dataset diversity, but the analysis in this and previous work shows that the current definition is sufficient to demonstrate the results of dataset compression.

## S2. Supplementary Figures

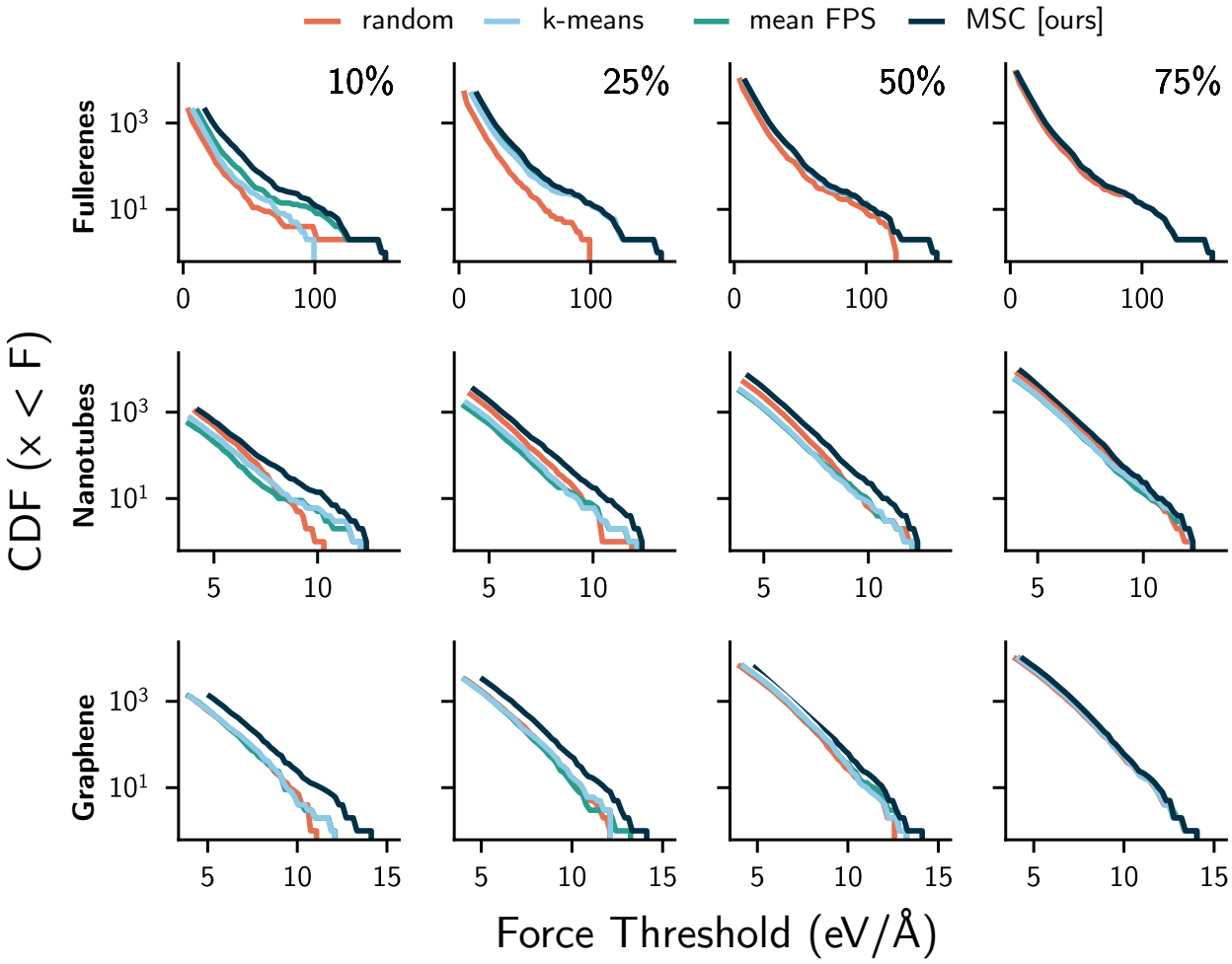


**Fig.** S1: Distribution of forces in selected GAP-20 datasets (Fullerenes, Nanotubes, Graphene) subsampled at various compression levels (75%, 50%, 25%, and 10%) using different sampling methods. All distributions are plotted starting from the 80th percentile of the corresponding full GAP-20 dataset. The consistently higher coverage of large-magnitude forces achieved by MSC across all compression levels highlights the algorithm's ability to better preserve the long tail of the force distribution compared to other methods.

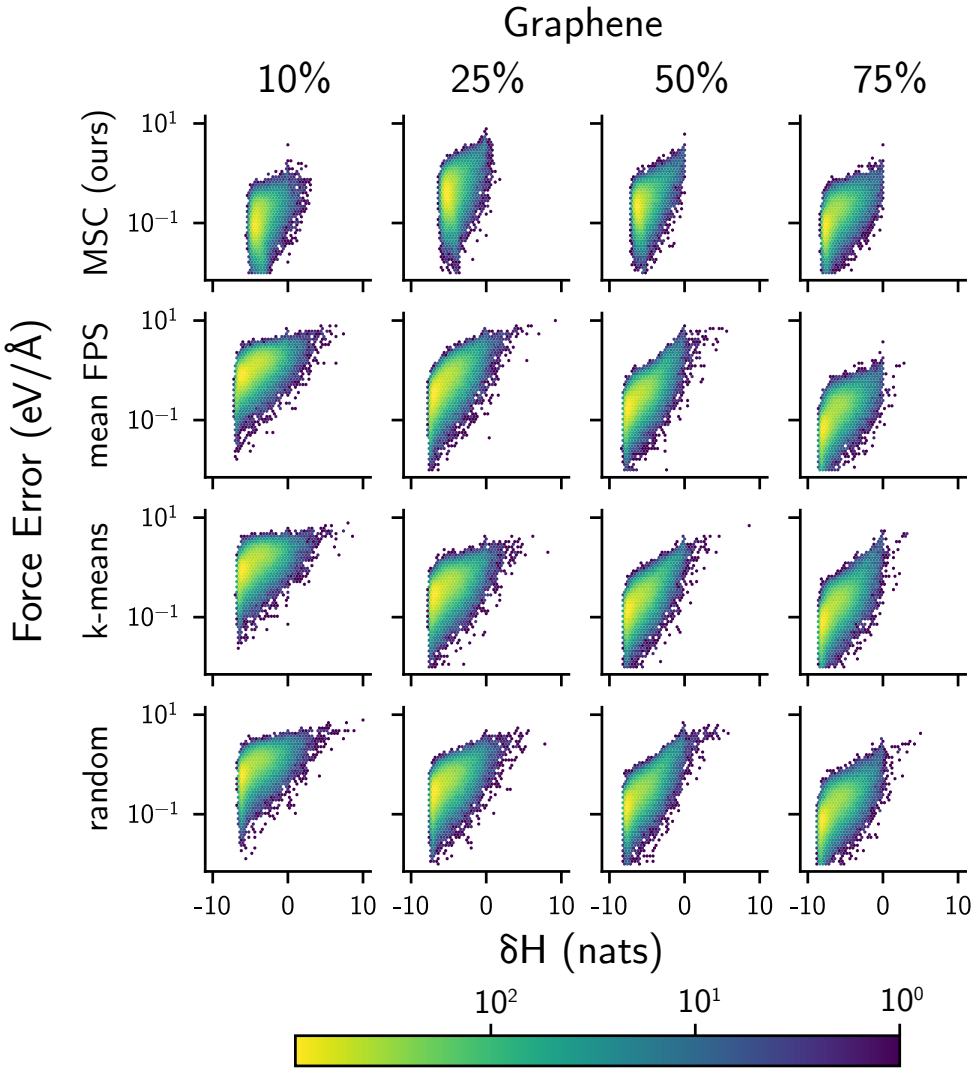


**Fig.** S2: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Graphene in GAP-20. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Across all subsampling rates, MSC results in fewer very positive $\delta\mathcal{H}$, demonstrating a smaller loss of outliers without compromising accuracy. Brighter colors represent a higher number of environments in each region of the space.

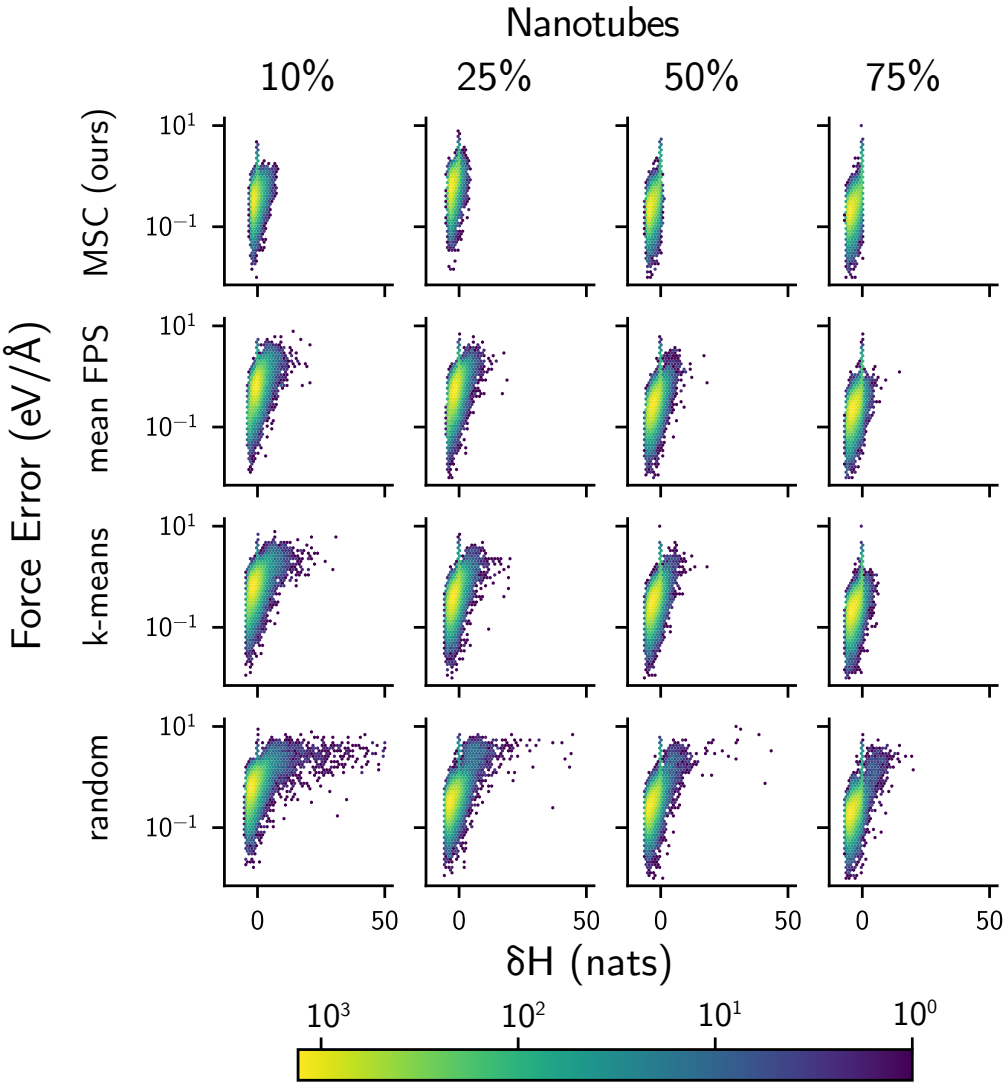


**Fig.** S3: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Nanotubes in GAP-20. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Across all subsampling rates, MSC results in fewer very positive $\delta\mathcal{H}$, demonstrating a smaller loss of outliers without compromising accuracy. Brighter colors represent a higher number of environments in each region of the space.

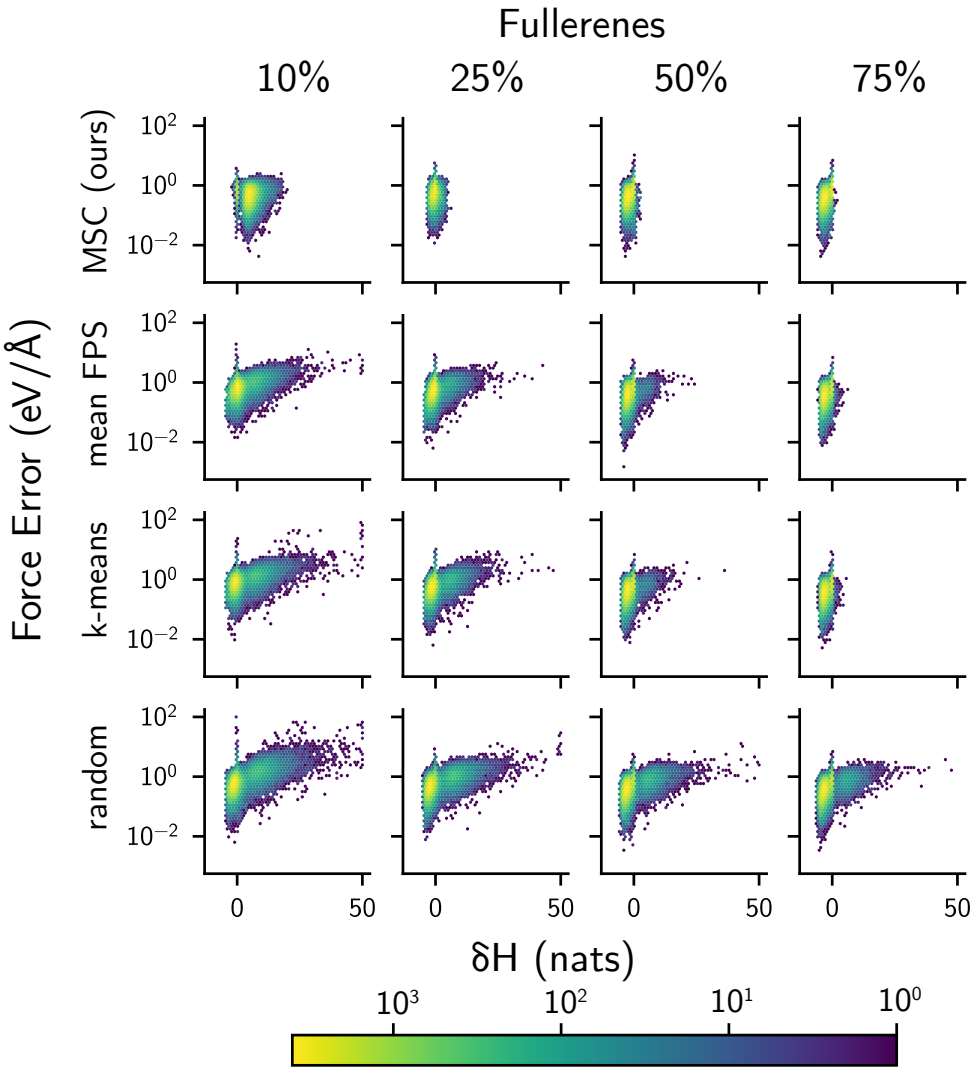


**Fig.** S4: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Fullerenes in GAP-20. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Across all subsampling rates, MSC results in fewer very positive $\delta\mathcal{H}$, demonstrating a smaller loss of outliers without compromising accuracy. Brighter colors represent a higher number of environments in each region of the space.

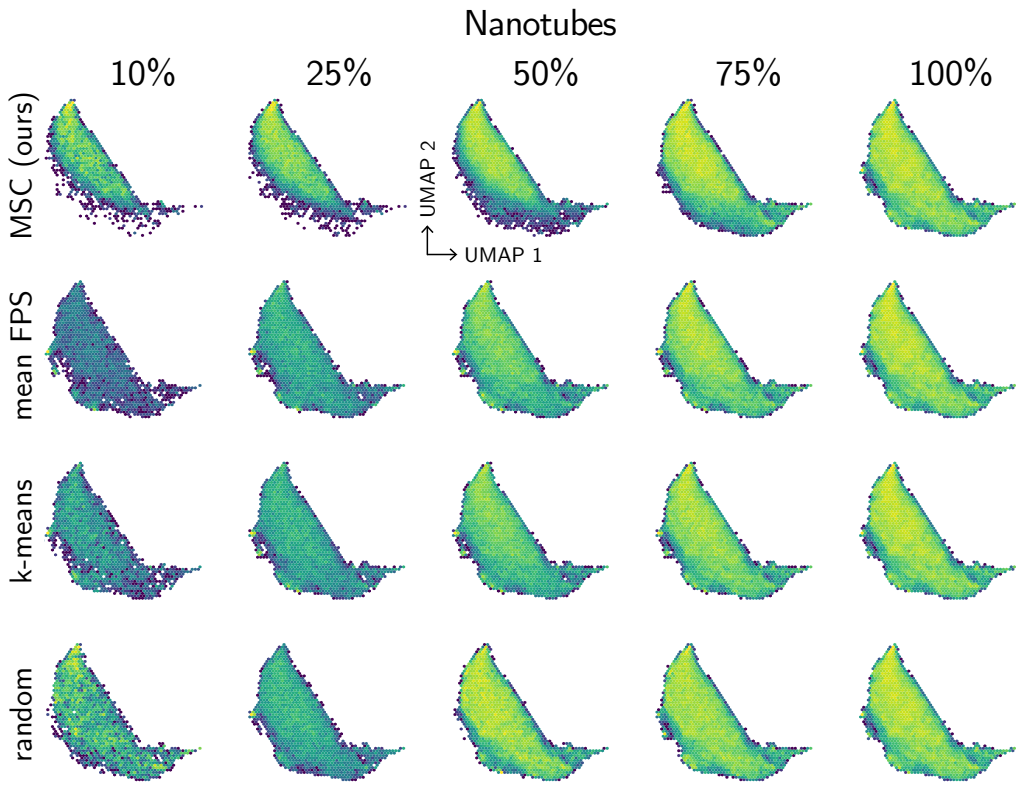


**Fig.** S5: Two-dimensional projection of per-atom descriptors from compressed datasets of the Graphene subset of GAP-20. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space. As discussed in the main text, the representations provide little information on the ability of each algorithm to compress the dataset.

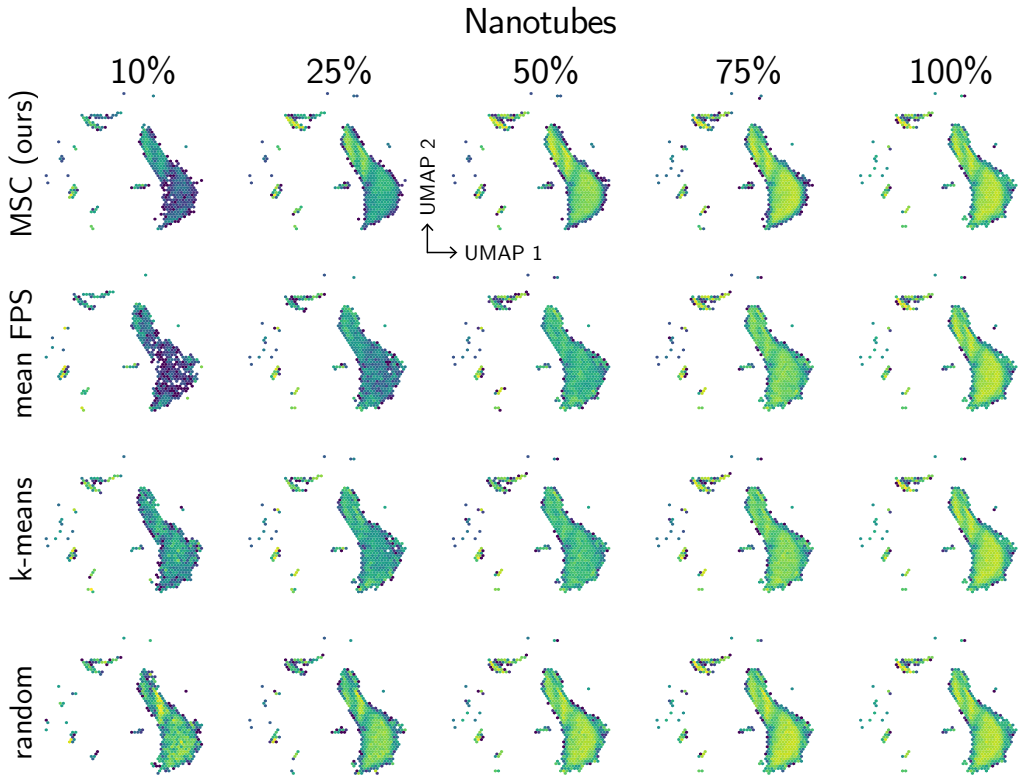


**Fig.** S6: Two-dimensional projection of per-atom descriptors from compressed datasets of the Nanotubes subset of GAP-20. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space. As discussed in the main text, the representations provide little information on the ability of each algorithm to compress the dataset.

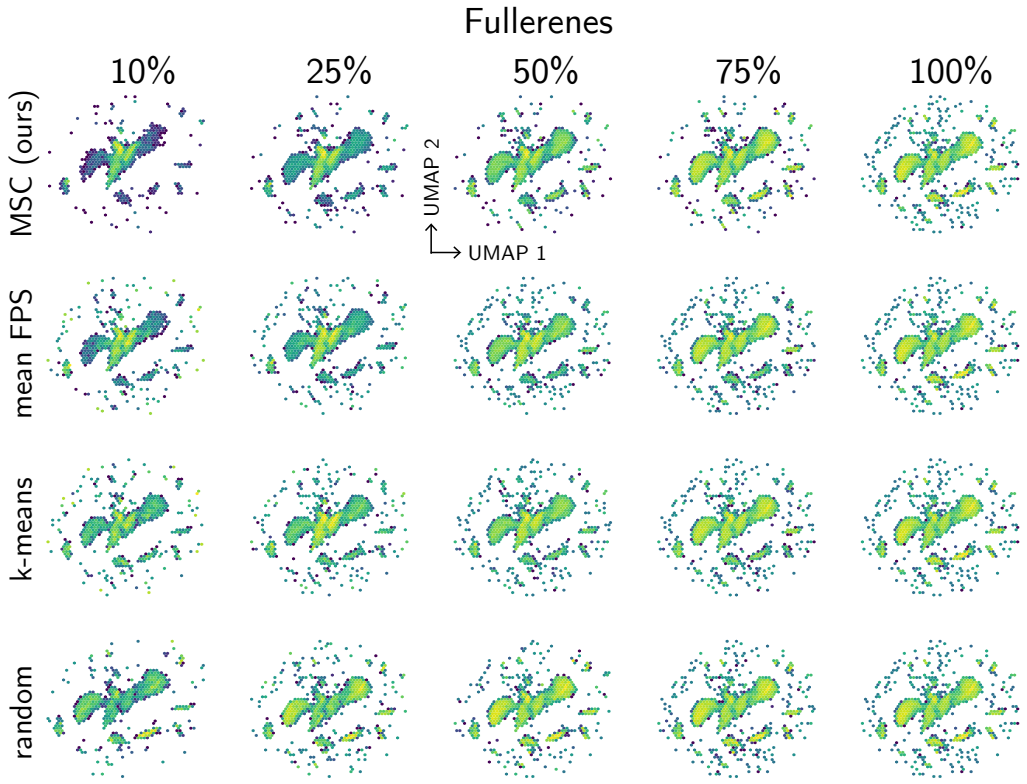


**Fig.** S7: Two-dimensional projection of per-atom descriptors from compressed datasets of the Fullerenes subset of GAP-20. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space. As discussed in the main text, the representations provide little information on the ability of each algorithm to compress the dataset.

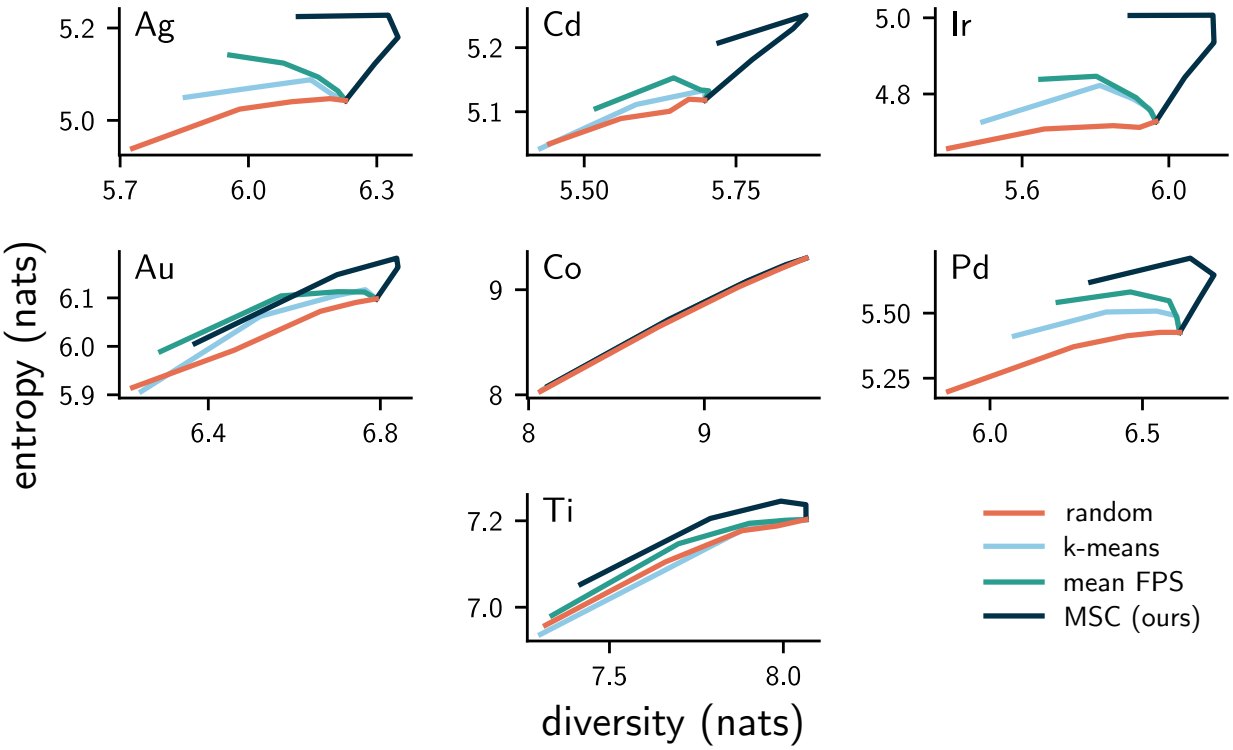


**Fig.** S8: Performance of compression algorithms in subsampling seven subsets of the TM23 dataset according to their overlap with the original dataset. Each dataset was subsampled at four different sizes: 75, 50, 25, and 10%, and compared against the full, original dataset. In nearly all cases, our MSC method exhibits the best behavior in terms of simultaneous entropy ($H$) increases and diversity ($D$) retention for the selected datasets across dataset sizes. On the other hand, random sampling quickly leads to information loss, as shown by immediate diversity decreases upon subsampling.

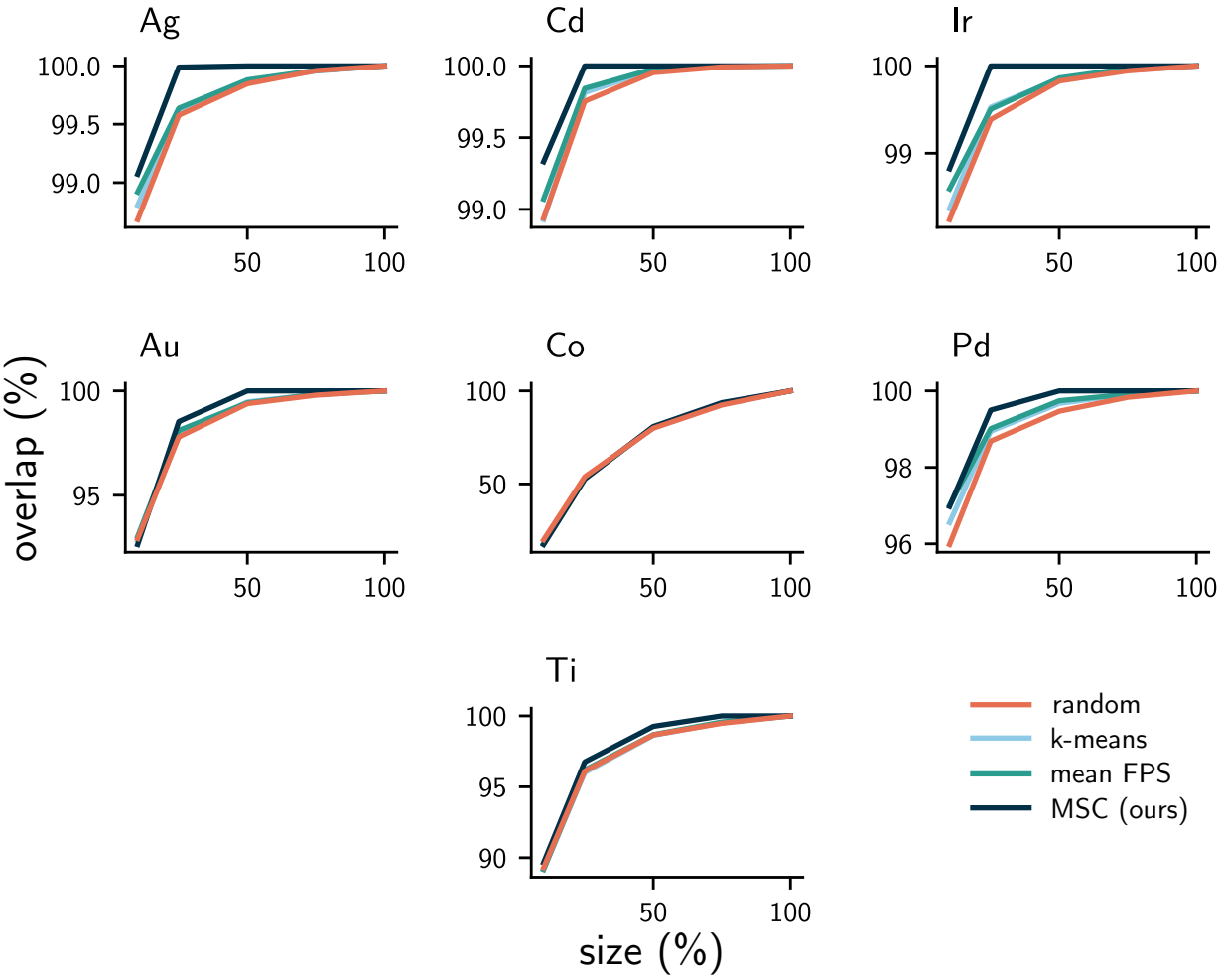


**Fig.** S9: Performance of compression algorithms in subsampling seven subsets of the TM23 dataset according to their overlap with the original dataset. Each dataset was subsampled at four different sizes: 75, 50, 25, and 10%, and compared against the full, original dataset. In all cases, our MSC method exhibits the best behavior in terms of retaining a high overlap with the original datasets upon compression.

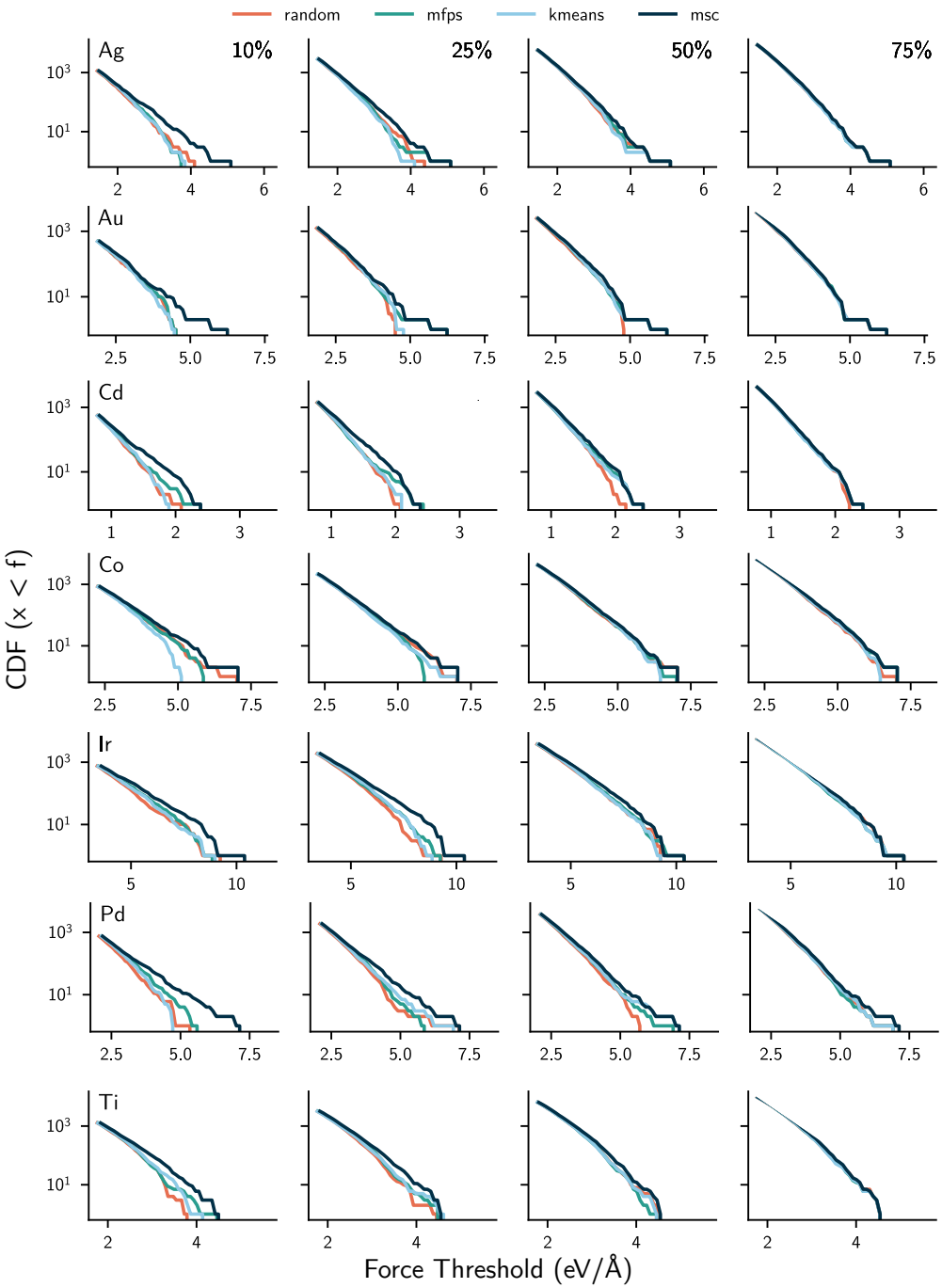


**Fig.** S10: Performance of compression algorithms in subsampling seven subsets of the TM23 dataset according to their ability to retain the long tail of forces. Each dataset was subsampled at four different sizes: 75, 50, 25, and 10%. In all cases, our MSC method exhibits the highest ability to preserve higher (and more rare) forces upon compression, as evidenced by the higher values of the cumulative density function (CDF).

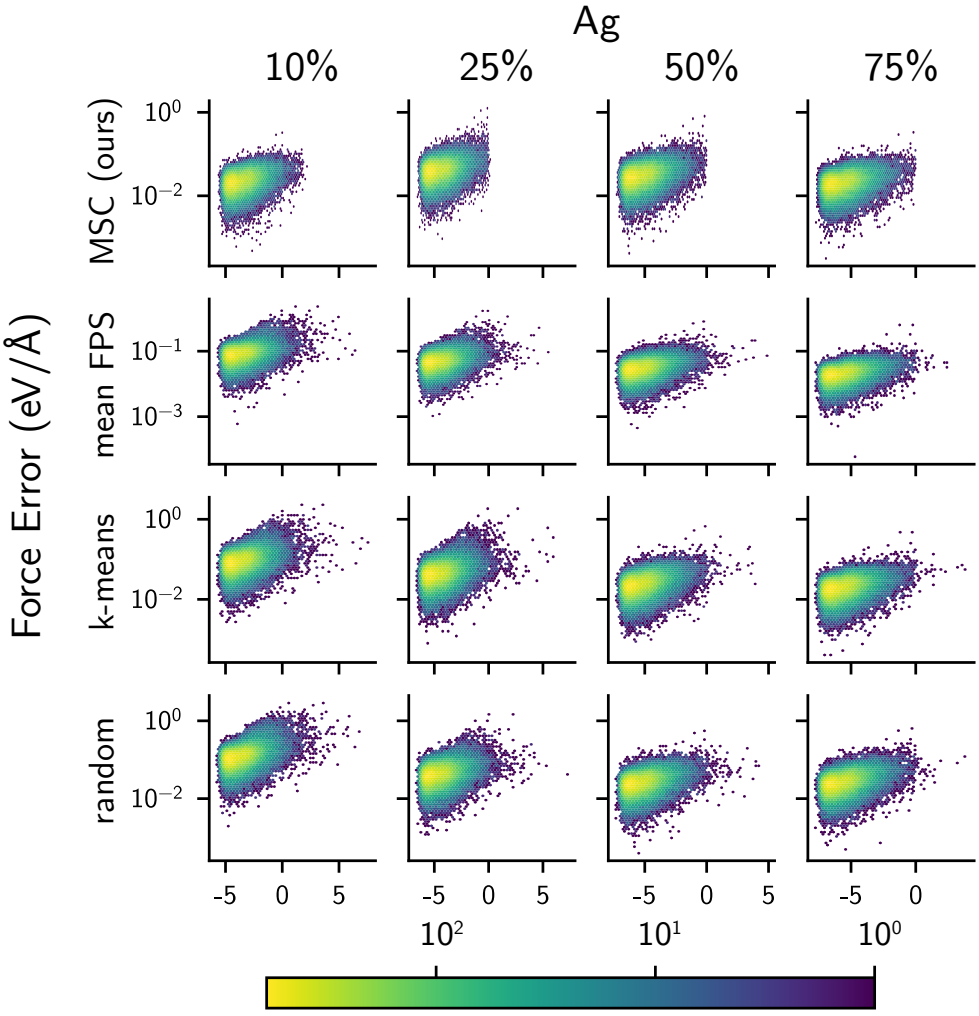


**Fig.** S11: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Ag (warm) in TM23. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Brighter colors represent a higher number of environments in each region of the space.

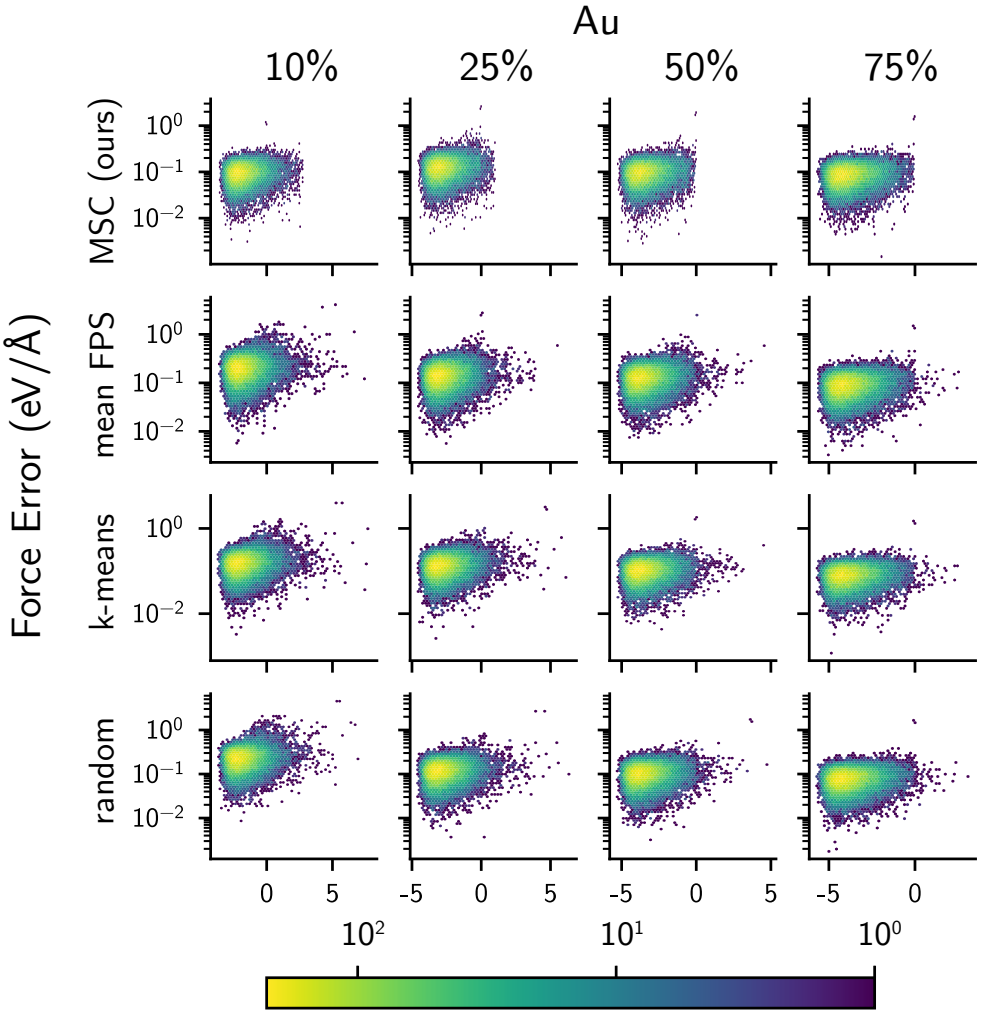


**Fig.** S12: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Au (warm) in TM23. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Brighter colors represent a higher number of environments in each region of the space.

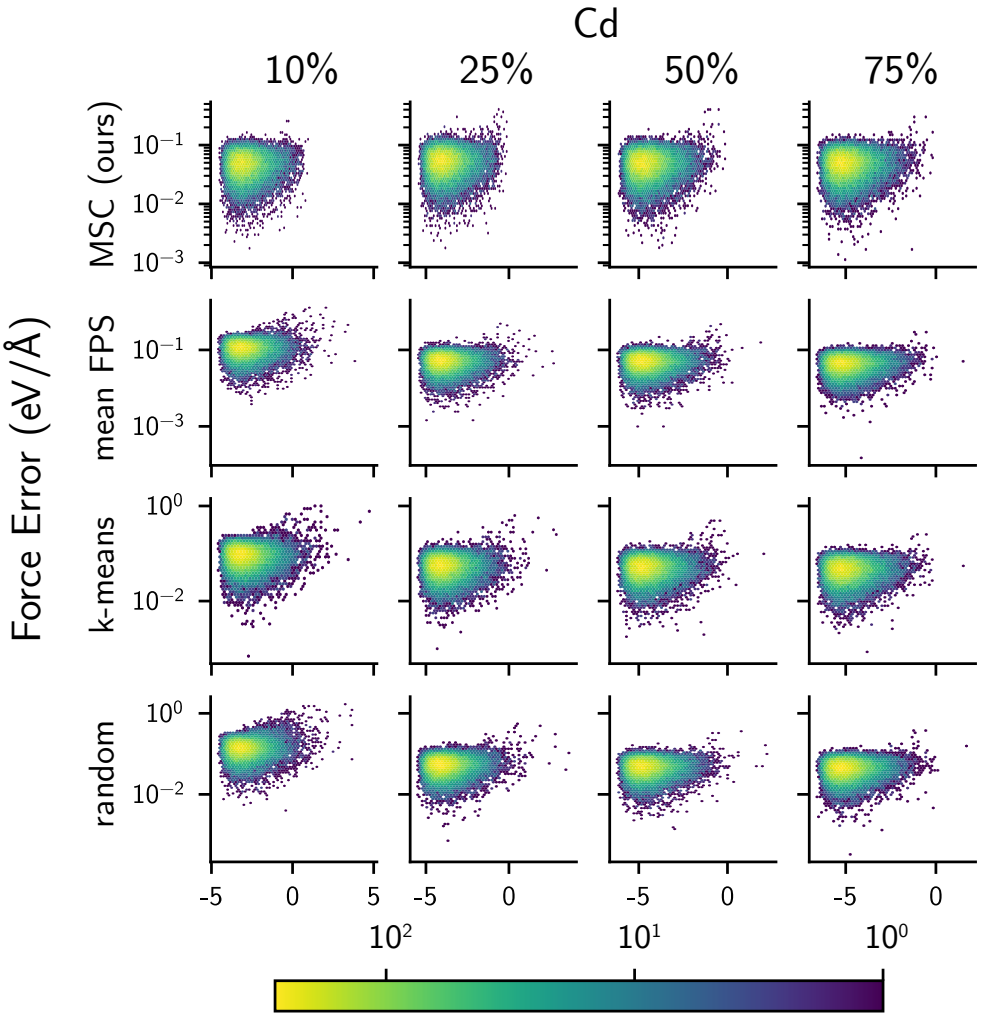


**Fig.** S13: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Cd (warm) in TM23. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Brighter colors represent a higher number of environments in each region of the space.

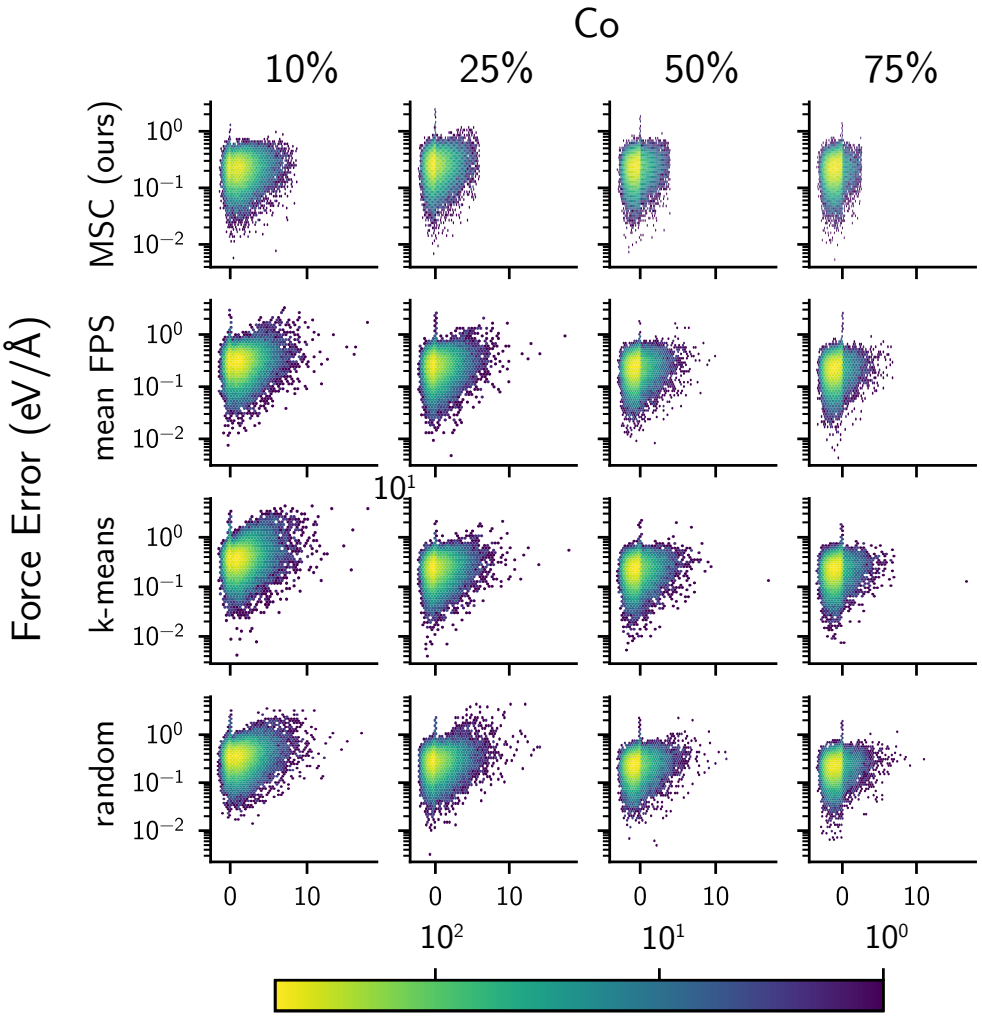


**Fig.** S14: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Co (warm) in TM23. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Brighter colors represent a higher number of environments in each region of the space.

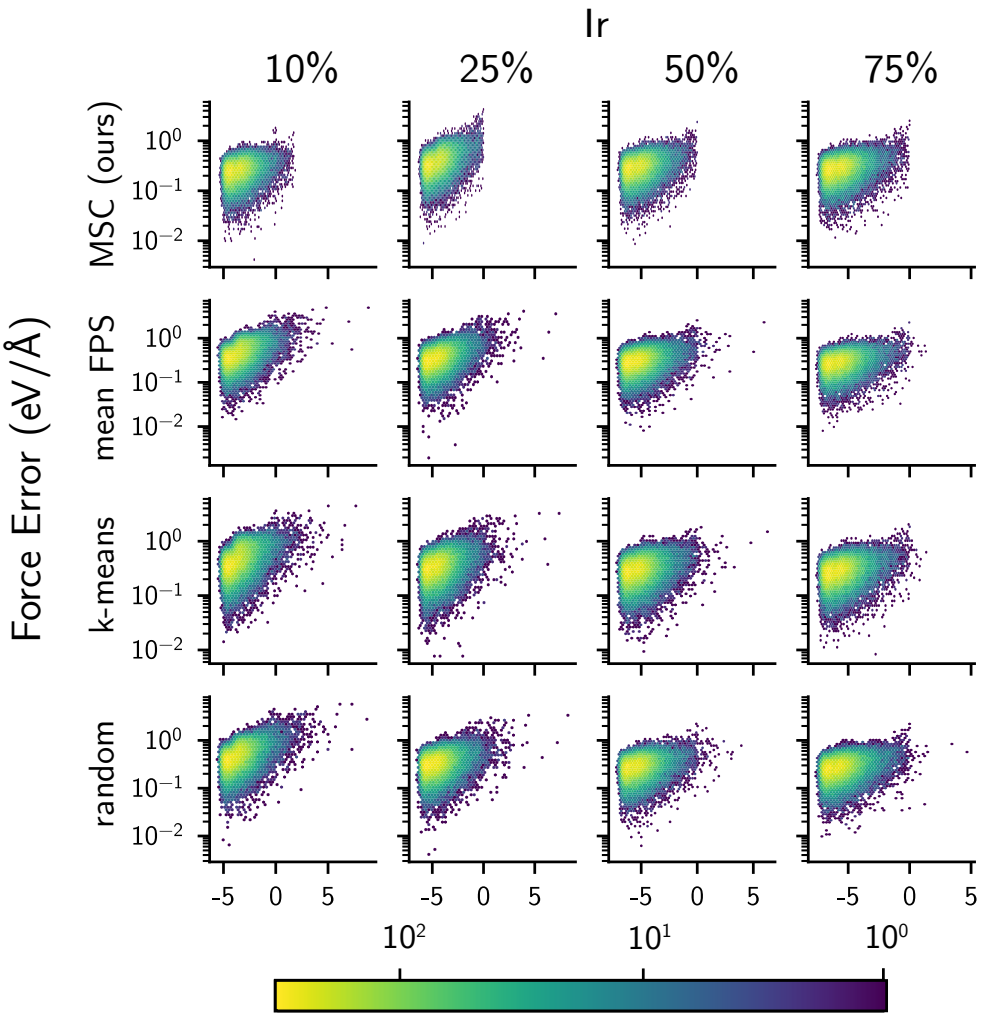


**Fig.** S15: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Ir (warm) in TM23. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Brighter colors represent a higher number of environments in each region of the space.

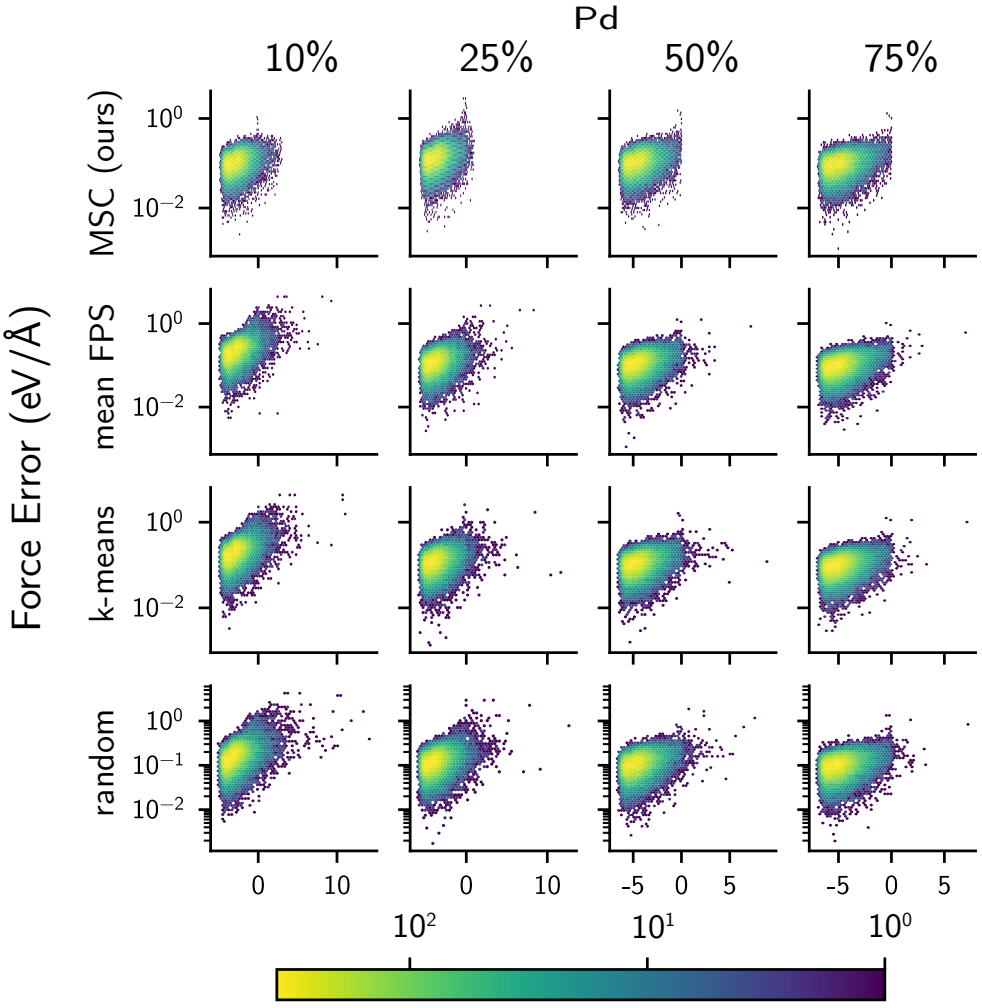


**Fig.** S16: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Pd (warm) in TM23. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Brighter colors represent a higher number of environments in each region of the space.

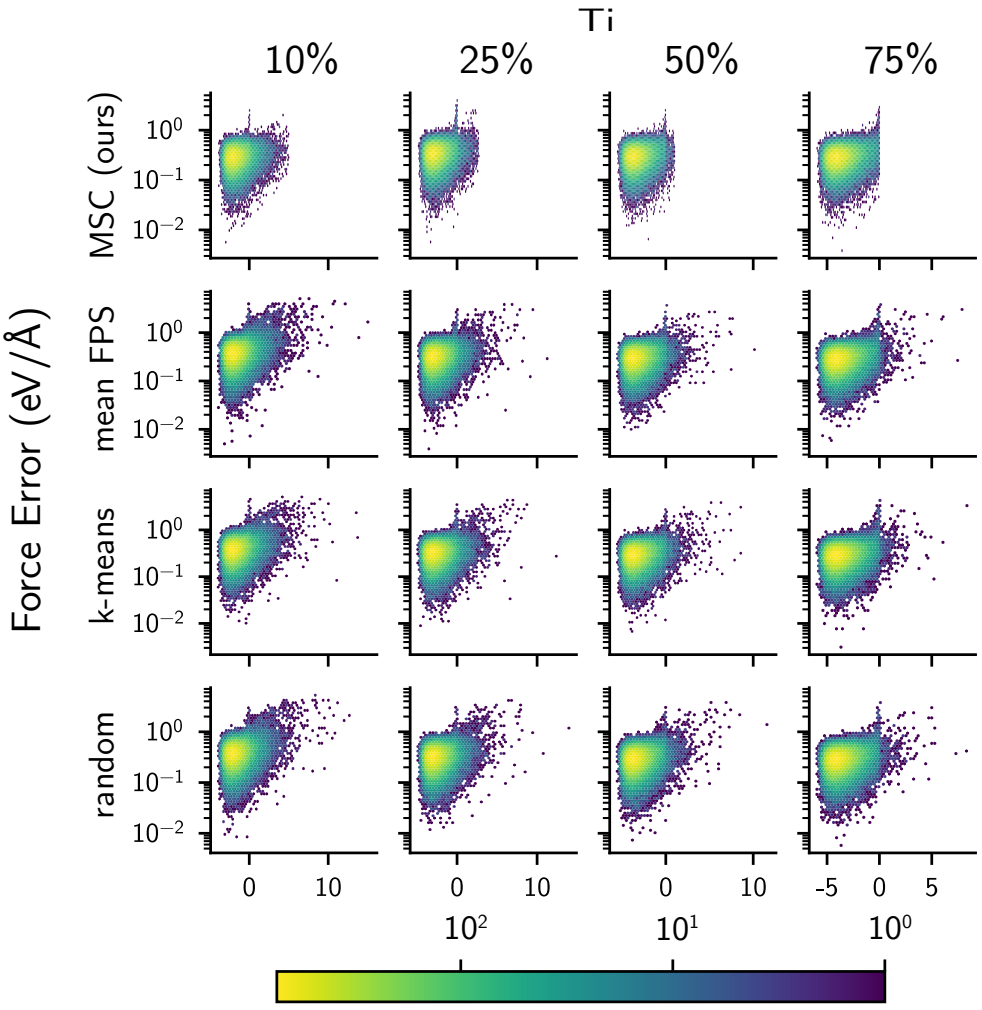


**Fig.** S17: Distribution of forces error for a SevenNet model trained on compressed datasets and tested on the original, full dataset of Ti (warm) in TM23. The distribution is correlated with the $\delta\mathcal{H}(\text{full}|\text{compressed})$, with very positive values of $\delta\mathcal{H}$ representing environments in the original dataset that are outside of the distribution of environments in the compressed dataset. Brighter colors represent a higher number of environments in each region of the space.

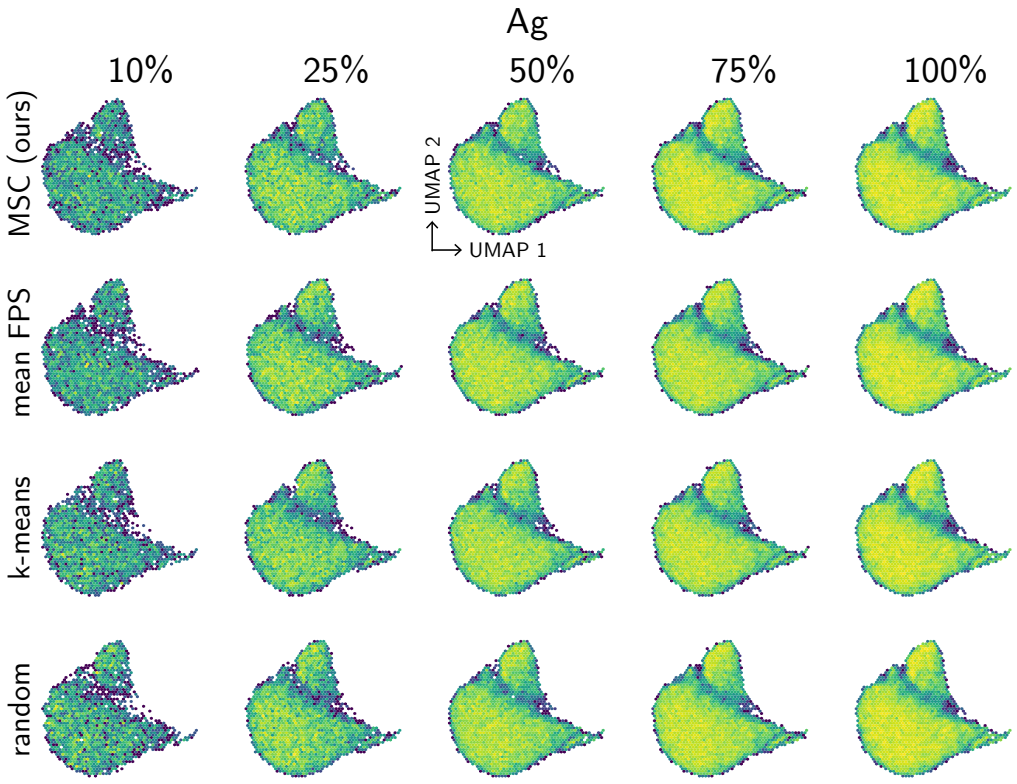


**Fig.** S18: Two-dimensional projection of per-atom descriptors from compressed datasets of the Ag (warm) subset of TM23. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space.

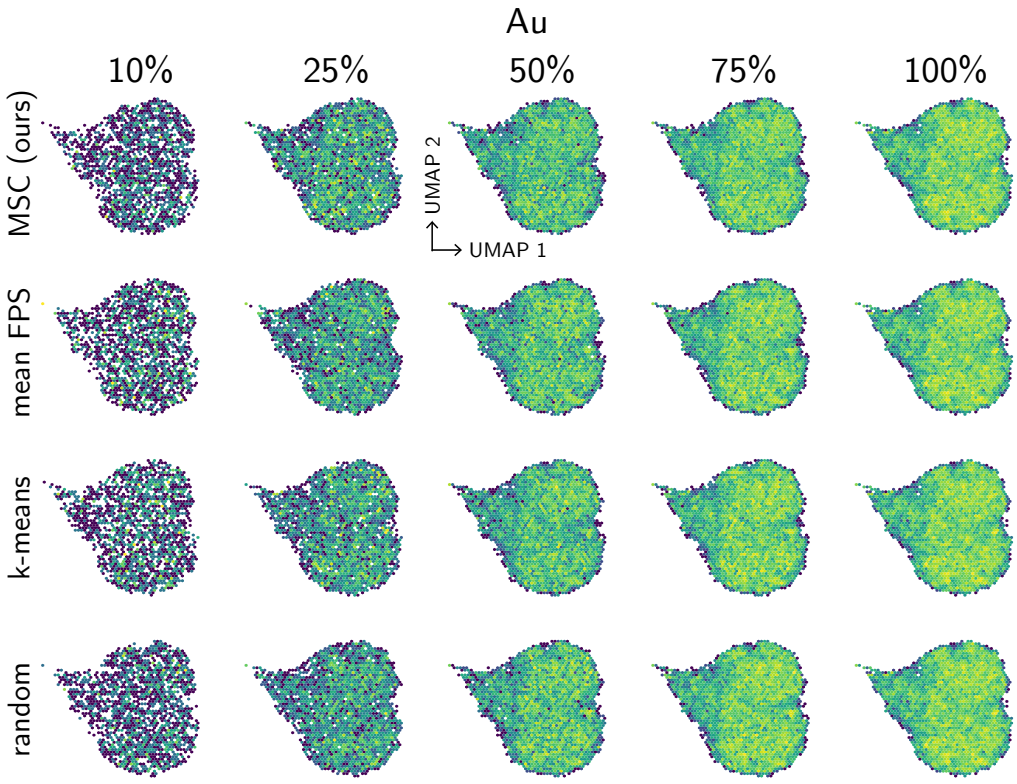


**Fig.** S19: Two-dimensional projection of per-atom descriptors from compressed datasets of the Au (warm) subset of TM23. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space.

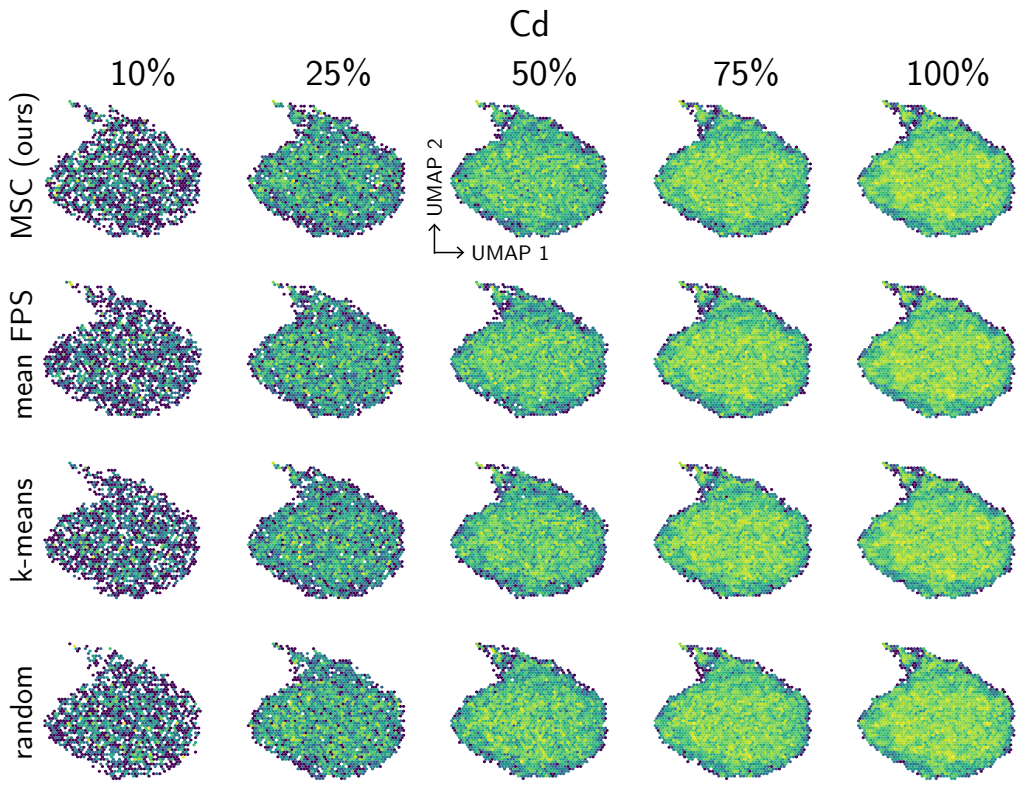


**Fig.** S20: Two-dimensional projection of per-atom descriptors from compressed datasets of the Cd (warm) subset of TM23. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space.

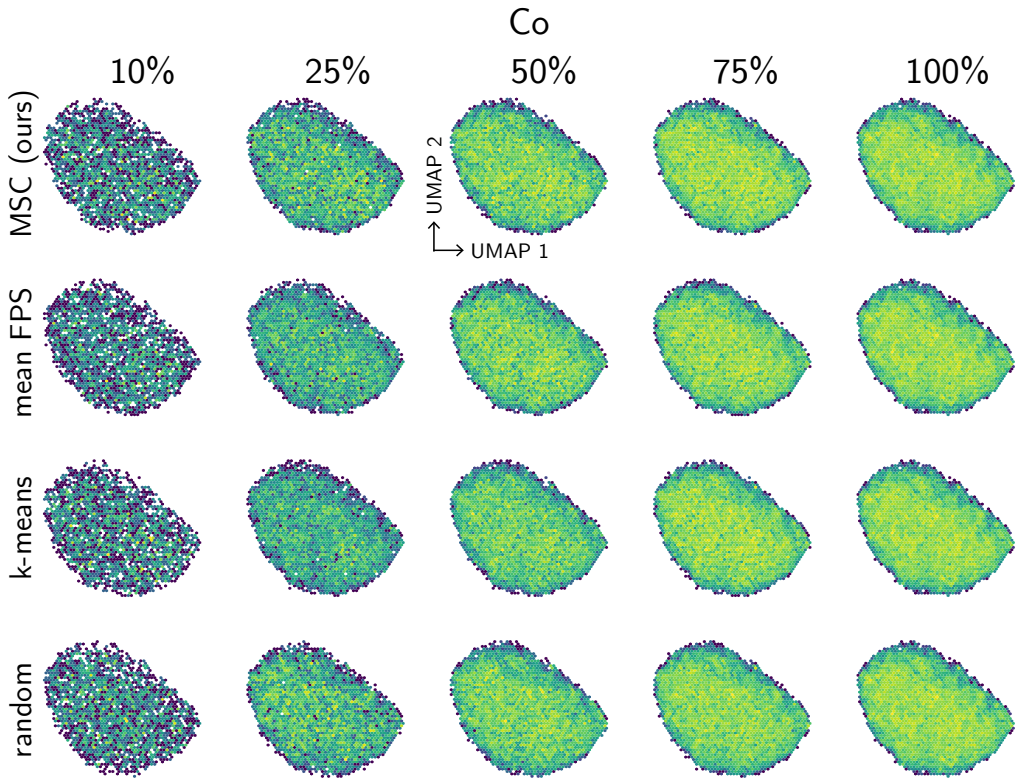


**Fig.** S21: Two-dimensional projection of per-atom descriptors from compressed datasets of the Co (warm) subset of TM23. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space.

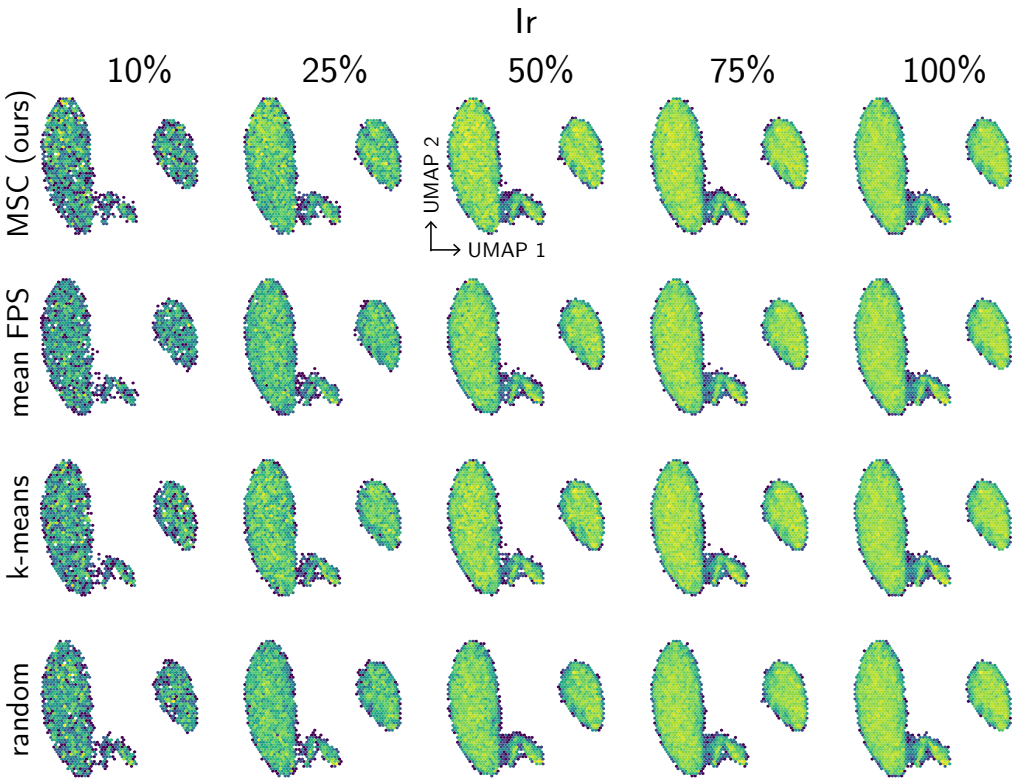


**Fig.** S22: Two-dimensional projection of per-atom descriptors from compressed datasets of the Ir (warm) subset of TM23. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space.

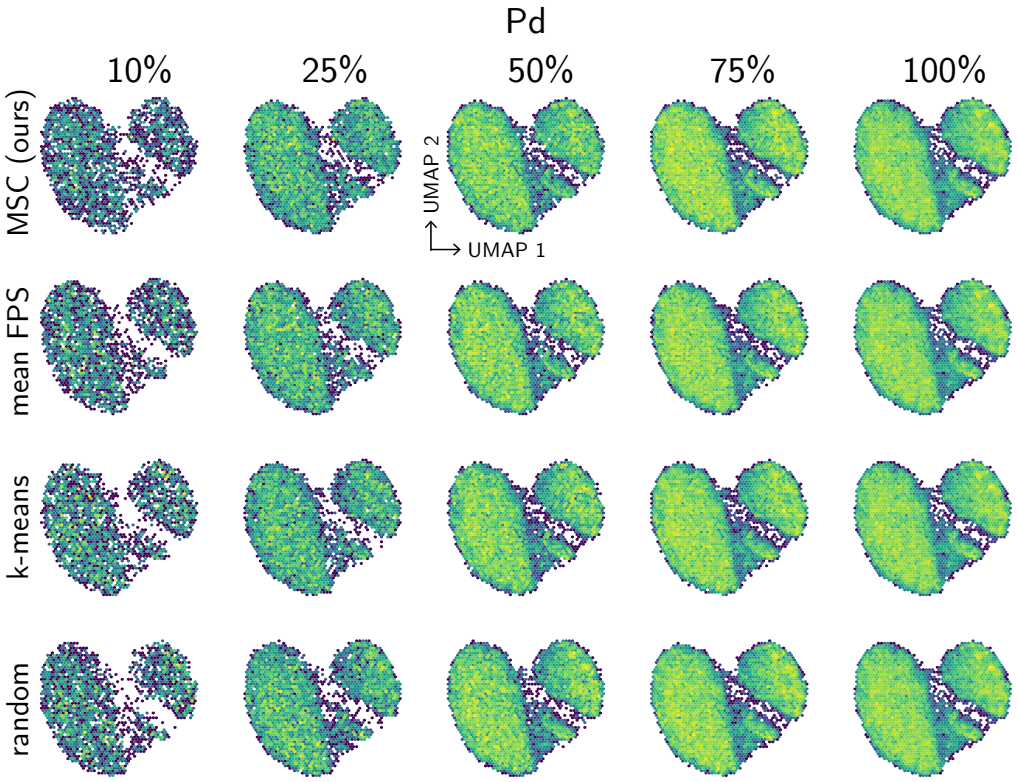


**Fig.** S23: Two-dimensional projection of per-atom descriptors from compressed datasets of the Pd (warm) subset of TM23. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space.

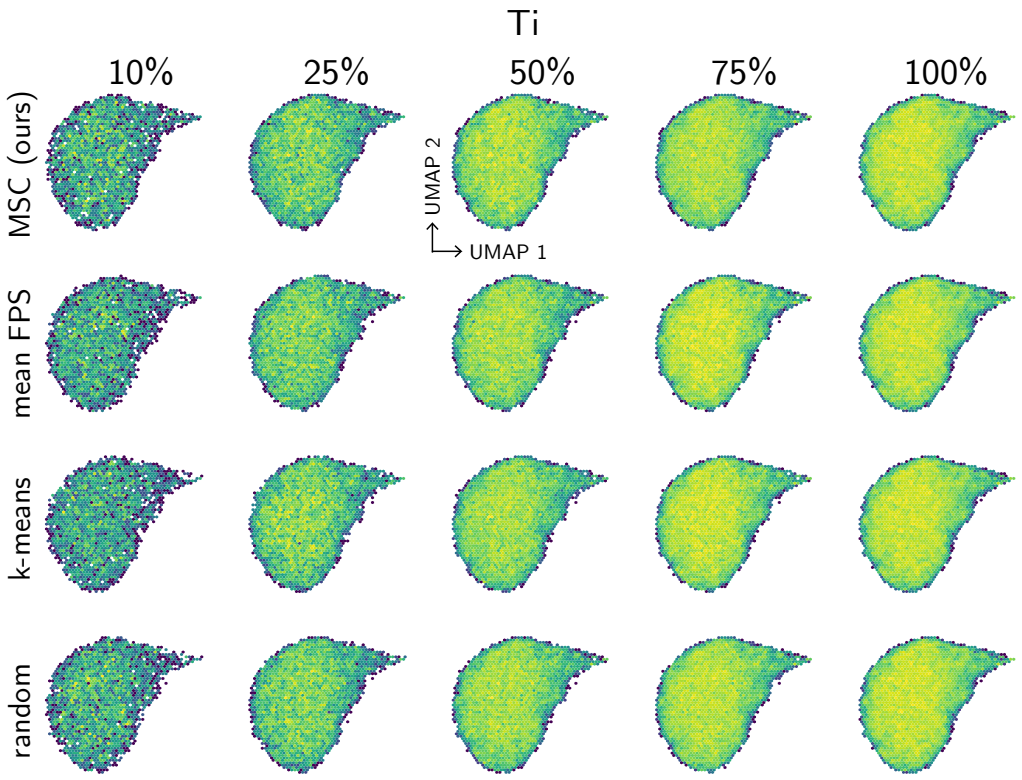


**Fig.** S24: Two-dimensional projection of per-atom descriptors from compressed datasets of the Ti (warm) subset of TM23. The projection was performed using UMAP, and the two axes correspond to the two components of UMAP. Brighter colors represent a higher number of environments in each region of the 2D space.

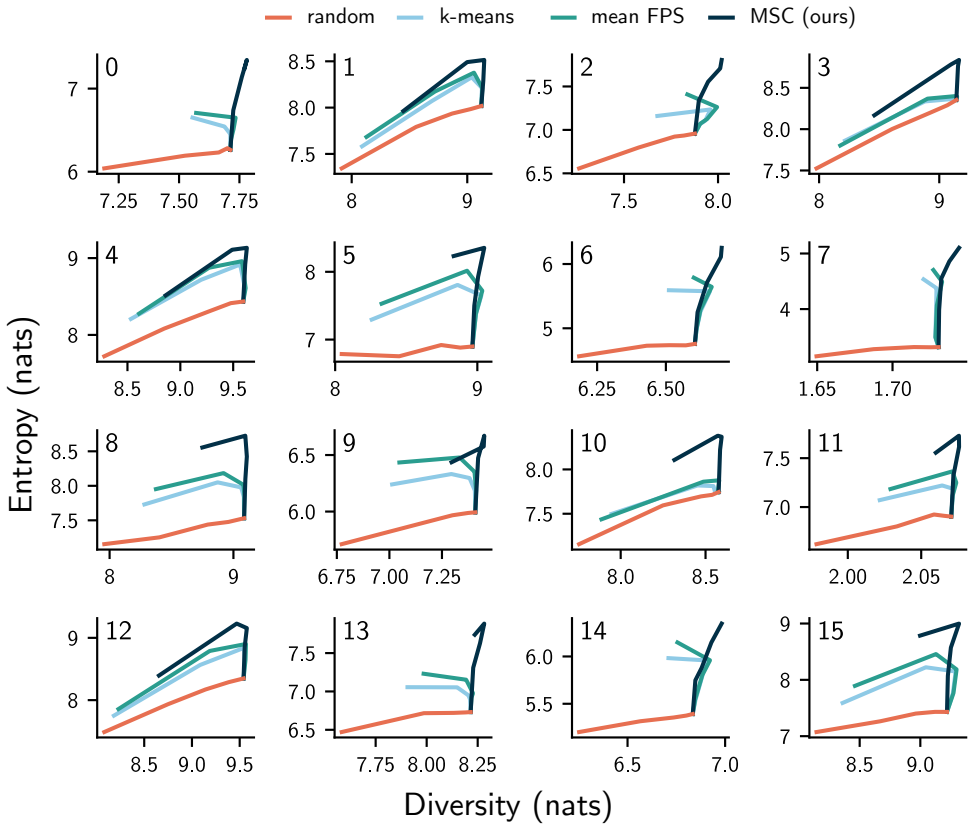


**Fig.** S25: Relationship between entropy and diversity of miscellaneous datasets obtained from the ColabFit repository sampled at different sizes (100%, 75%, 50%, 25%, 10%) and with different methods. Outside of a few outliers, datasets compressed with MSC resulted in higher diversity and entropy relative to datasets compressed with other methods, demonstrating that our compression algorithm is the most efficient in removing the redundancy of the data. The dataset numbers, shown on the top left of each plot, correspond to the ones shown in Table S19.

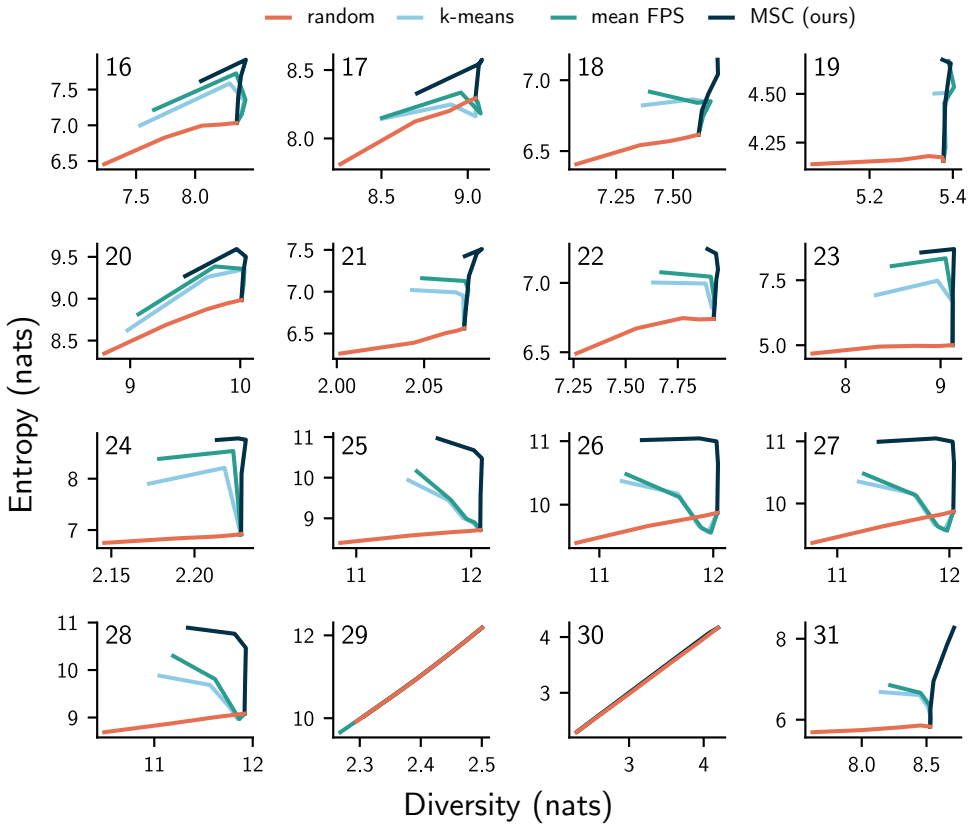


**Fig.** S26: Relationship between entropy and diversity of miscellaneous datasets obtained from the ColabFit repository sampled at different sizes (100%, 75%, 50%, 25%, 10%) and with different methods. Outside of a few outliers, datasets compressed with MSC resulted in higher diversity and entropy relative to datasets compressed with other methods, demonstrating that our compression algorithm is the most efficient in removing the redundancy of the data. The dataset numbers, shown on the top left of each plot, correspond to the ones shown in Table S19.

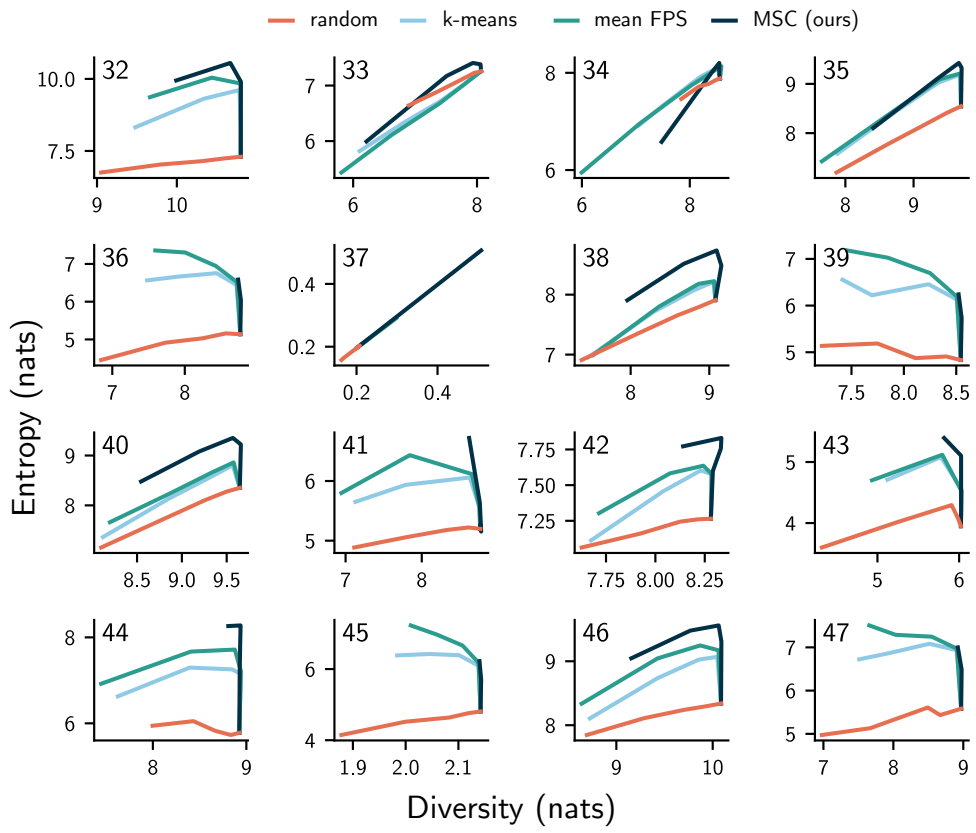


**Fig.** S27: Relationship between entropy and diversity of miscellaneous datasets obtained from the ColabFit repository sampled at different sizes (100%, 75%, 50%, 25%, 10%) and with different methods. Outside of a few outliers, datasets compressed with MSC resulted in higher diversity and entropy relative to datasets compressed with other methods, demonstrating that our compression algorithm is the most efficient in removing the redundancy of the data. The dataset numbers, shown on the top left of each plot, correspond to the ones shown in Table S19.

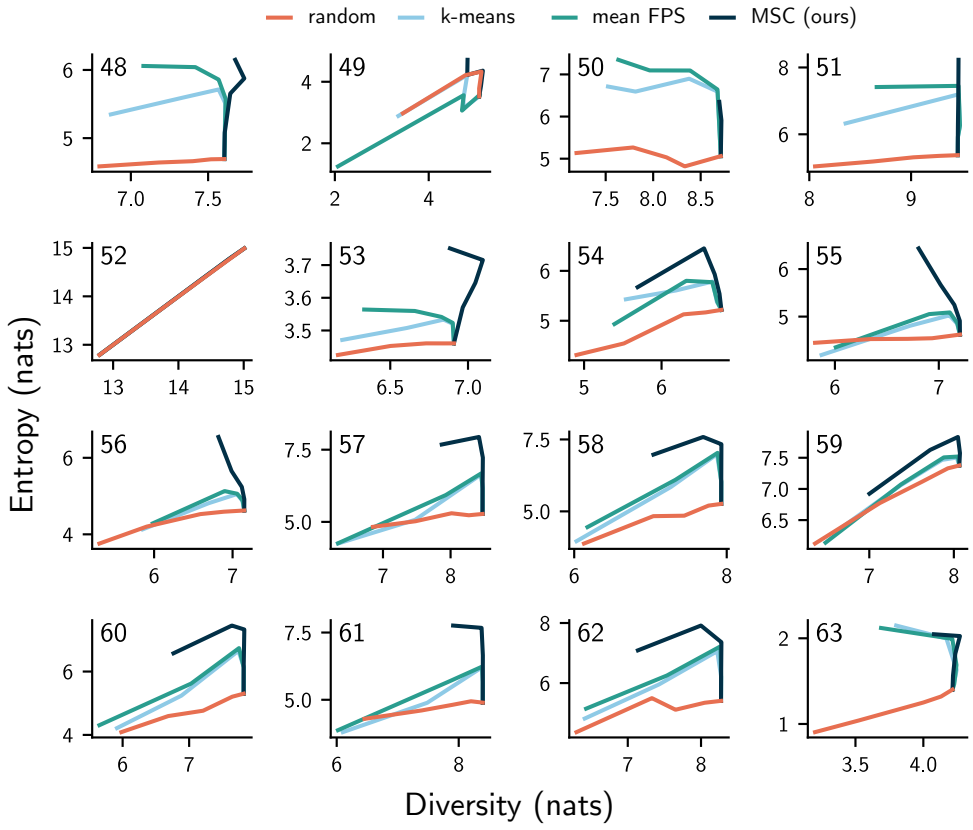


**Fig.** S28: Relationship between entropy and diversity of miscellaneous datasets obtained from the ColabFit repository sampled at different sizes (100%, 75%, 50%, 25%, 10%) and with different methods. Outside of a few outliers, datasets compressed with MSC resulted in higher diversity and entropy relative to datasets compressed with other methods, demonstrating that our compression algorithm is the most efficient in removing the redundancy of the data. The dataset numbers, shown on the top left of each plot, correspond to the ones shown in Table S19.

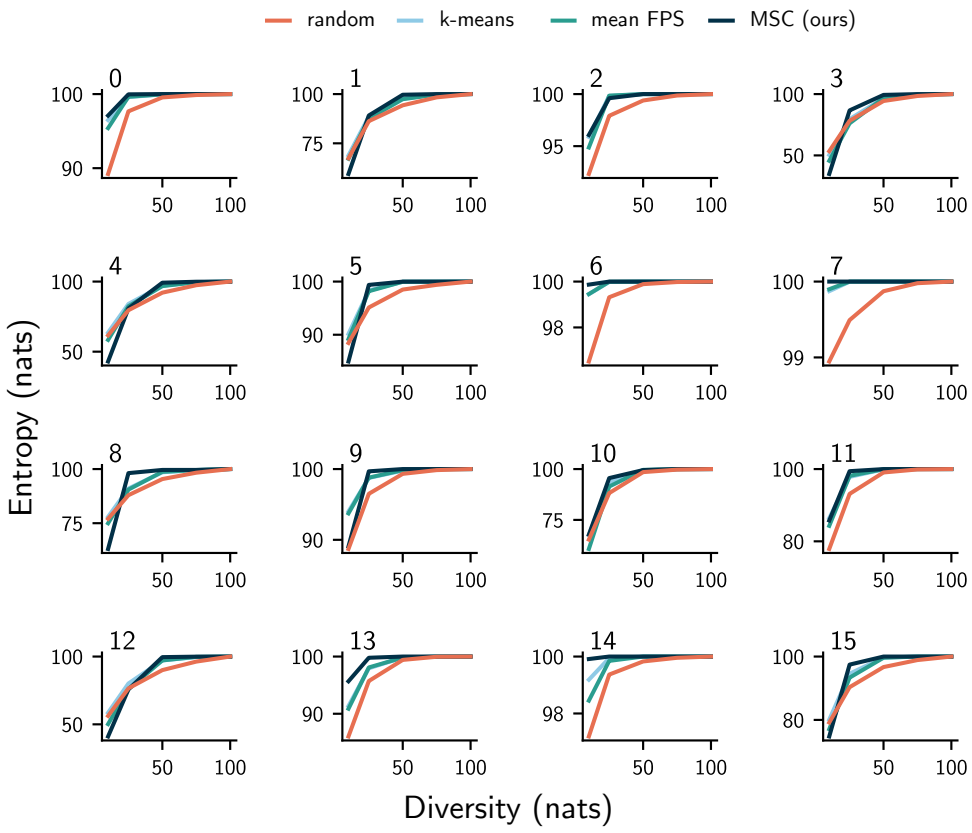


**Fig.** S29: Relationship between overlap of several compressed datasets obtained from the ColabFit repository sampled at different sizes (100%, 75%, 50%, 25%, 10%) with respect to their own, respective full datasets. Aside from a few outliers, datasets compressed with MSC consistently have higher overlap with the full dataset compared to other algorithms, demonstrating that our algorithm more efficiently preserves the distribution of the dataset. The dataset numbers, shown on the top left of each plot, correspond to the ones shown in Table S19.

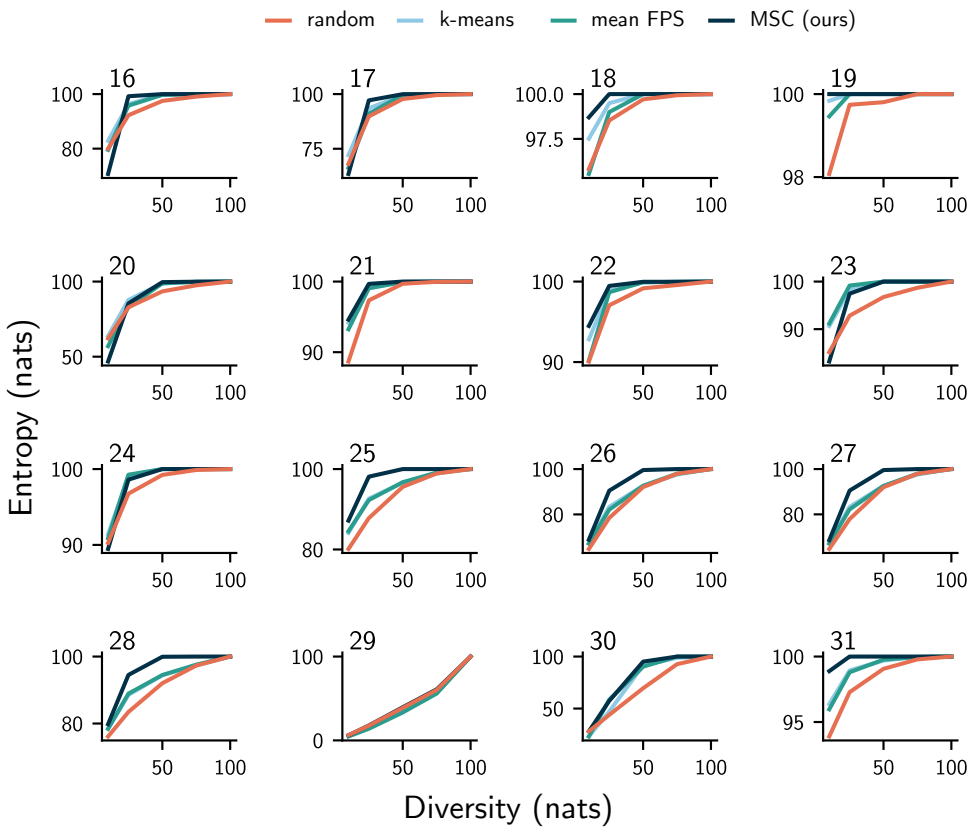


**Fig.** S30: Relationship between overlap of several compressed datasets obtained from the ColabFit repository sampled at different sizes (100%, 75%, 50%, 25%, 10%) with respect to their own, respective full datasets. Aside from a few outliers, datasets compressed with MSC consistently have higher overlap with the full dataset compared to other algorithms, demonstrating that our algorithm more efficiently preserves the distribution of the dataset. The dataset numbers, shown on the top left of each plot, correspond to the ones shown in Table S19.

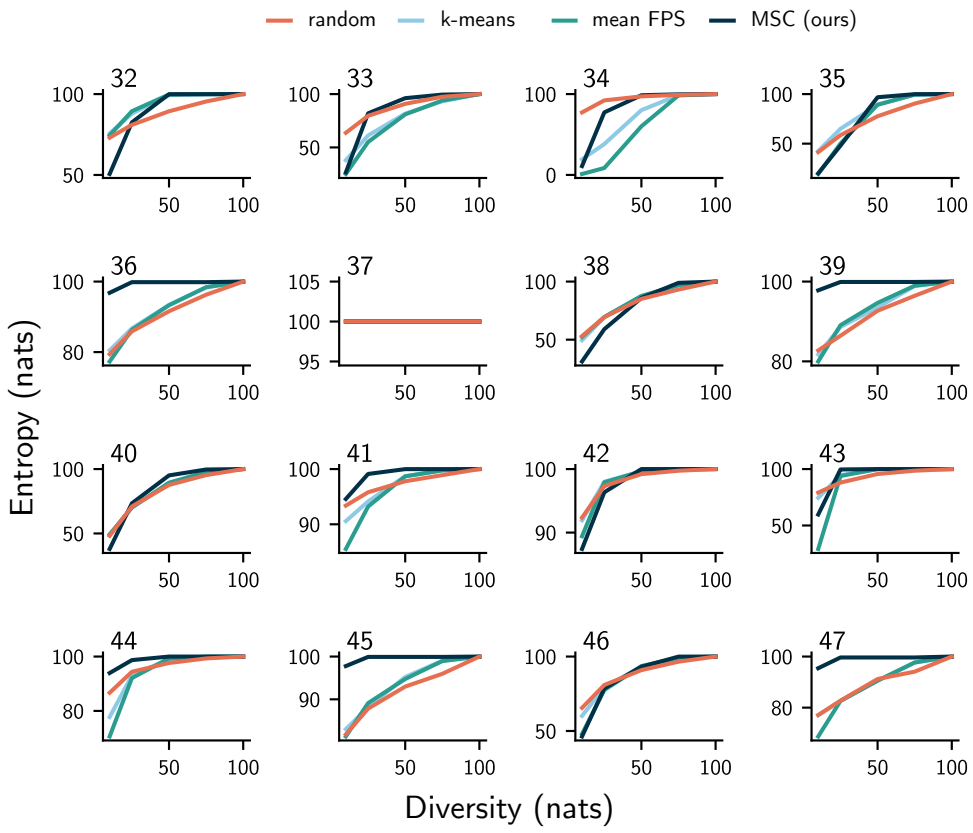


**Fig.** S31: Relationship between overlap of several compressed datasets obtained from the ColabFit repository sampled at different sizes (100%, 75%, 50%, 25%, 10%) with respect to their own, respective full datasets. Aside from a few outliers, datasets compressed with MSC consistently have higher overlap with the full dataset compared to other algorithms, demonstrating that our algorithm more efficiently preserves the distribution of the dataset. The dataset numbers, shown on the top left of each plot, correspond to the ones shown in Table S19.

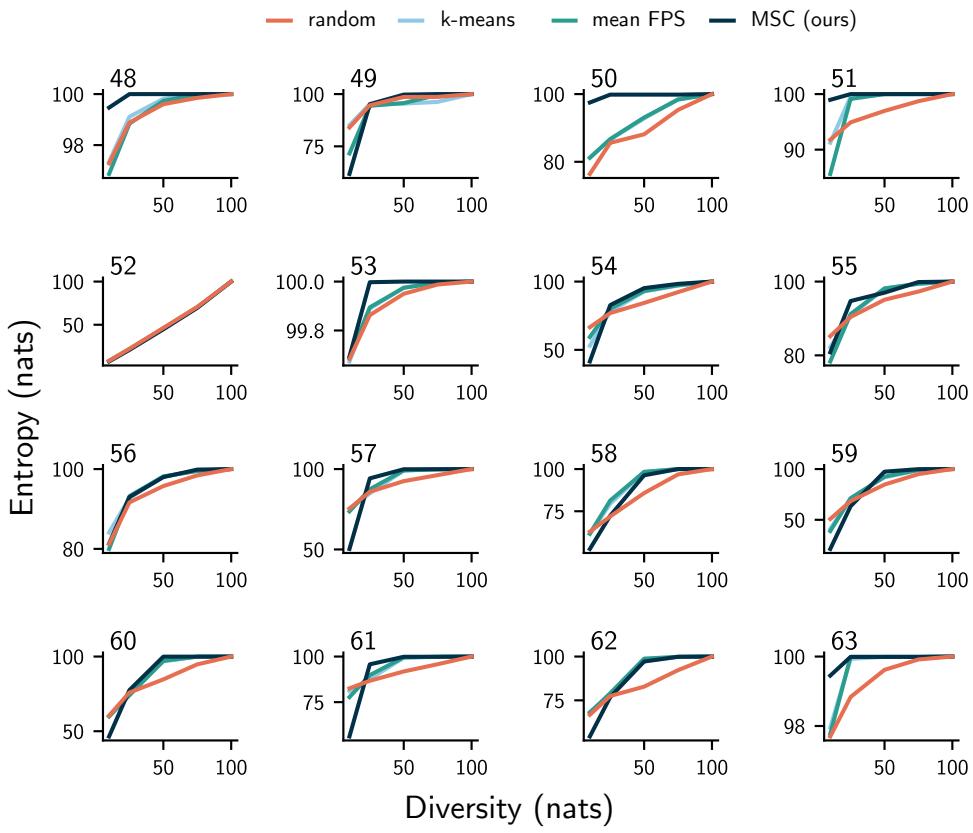


**Fig.** S32: Relationship between overlap of several compressed datasets obtained from the ColabFit repository sampled at different sizes (100%, 75%, 50%, 25%, 10%) with respect to their own, respective full datasets. Aside from a few outliers, datasets compressed with MSC consistently have higher overlap with the full dataset compared to other algorithms, demonstrating that our algorithm more efficiently preserves the distribution of the dataset. The dataset numbers, shown on the top left of each plot, correspond to the ones shown in Table S19.

# S3. Supplementary Tables

## S3.1. Results of dataset compression for GAP-20

**Table** S1: Entropy, Diversity, and Overlap results for the Graphene subset of GAP-20

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 4.24 | 4.20 | 4.09 | 5.55 | 5.75 | 5.81 | 5.82 | 6.34 | 0.98 | 0.98 | 0.98 | 0.99 |
| 25 | 4.34 | 4.33 | 4.31 | 5.55 | 6.03 | 6.11 | 6.08 | 6.55 | 0.99 | 0.99 | 0.99 | 1.00 |
| 50 | 4.29 | 4.52 | 4.52 | 5.37 | 6.22 | 6.32 | 6.32 | 6.54 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75 | 4.28 | 4.44 | 4.53 | 4.78 | 6.39 | 6.41 | 6.44 | 6.47 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100 | 4.30 | 4.30 | 4.30 | 4.30 | 6.45 | 6.45 | 6.45 | 6.45 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S2: Entropy, Diversity, and Overlap results for the Nanotubes subset of GAP-20

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 6.54 | 6.66 | 7.00 | 7.86 | 7.25 | 7.30 | 7.40 | 8.09 | 0.78 | 0.73 | 0.61 | 0.80 |
| 25 | 6.77 | 7.14 | 7.29 | 7.87 | 7.85 | 7.97 | 7.98 | 8.53 | 0.87 | 0.85 | 0.82 | 0.97 |
| 50 | 6.93 | 7.11 | 7.14 | 7.52 | 8.28 | 8.30 | 8.30 | 8.66 | 0.94 | 0.93 | 0.92 | 1.00 |
| 75 | 6.92 | 7.09 | 7.14 | 7.26 | 8.46 | 8.52 | 8.52 | 8.65 | 0.97 | 0.97 | 0.97 | 1.00 |
| 100 | 7.03 | 7.03 | 7.03 | 7.03 | 8.65 | 8.65 | 8.65 | 8.65 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S3: Entropy, Diversity, and Overlap results for the Fullerenes subset of GAP-20

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 7.76 | 7.99 | 8.22 | 9.18 | 8.29 | 8.54 | 8.72 | 9.18 | 0.67 | 0.62 | 0.41 | 0.08 |
| 25 | 8.12 | 8.92 | 9.22 | 9.76 | 8.88 | 9.45 | 9.62 | 9.88 | 0.80 | 0.82 | 0.79 | 0.74 |
| 50 | 8.49 | 9.15 | 9.21 | 9.36 | 9.49 | 9.91 | 9.93 | 10.00 | 0.88 | 0.94 | 0.94 | 0.96 |
| 75 | 8.52 | 8.96 | 8.98 | 8.99 | 9.76 | 9.99 | 9.99 | 10.01 | 0.92 | 0.97 | 0.97 | 0.97 |
| 100 | 8.60 | 8.60 | 8.60 | 8.60 | 9.99 | 9.99 | 9.99 | 9.99 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S4: Force errors for a SevenNet model trained on subsampled datasets of the Graphene, Nanotubes, and Fullerenes subsets of GAP-20. The models were tested on the original, full datasets.

| | Graphene Error (eV/Å) | | | | Fullerenes Error (eV/Å) | | | | Nanotubes Error (eV/Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 0.47 | 0.49 | 0.45 | 0.24 | 3.34 | 6.22 | 6.33 | 2.73 | 0.86 | 1.07 | 1.07 | 0.71 |
| 25 | 0.23 | 0.26 | 0.29 | 0.27 | 2.99 | 2.96 | 3.08 | 3.13 | 0.70 | 0.72 | 0.80 | 0.70 |
| 50 | 0.18 | 0.17 | 0.20 | 0.17 | 2.88 | 2.58 | 3.09 | 2.49 | 0.71 | 0.55 | 0.70 | 0.53 |
| 75 | 0.13 | 0.14 | 0.14 | 0.14 | 2.61 | 3.45 | 2.76 | 2.55 | 0.61 | 0.63 | 0.59 | 0.56 |
| 100 | 0.11 | 0.11 | 0.11 | 0.11 | 2.90 | 2.90 | 2.90 | 2.90 | 0.69 | 0.69 | 0.69 | 0.69 |

## S3.2. Results of dataset compression for TM23

**Table** S5: Entropy, Diversity, and Overlap results for the Ag (warm) subset of TM23

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 3.78 | 3.87 | 3.93 | 4.02 | 4.53 | 4.58 | 4.66 | 4.97 | 0.99 | 0.99 | 0.99 | 1.00 |
| 25 | 3.83 | 3.87 | 3.90 | 4.00 | 4.65 | 4.69 | 4.75 | 5.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 50 | 3.83 | 3.87 | 3.88 | 3.95 | 4.73 | 4.77 | 4.77 | 4.94 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75 | 3.84 | 3.85 | 3.85 | 3.90 | 4.76 | 4.77 | 4.78 | 4.86 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100 | 3.84 | 3.84 | 3.84 | 3.84 | 4.79 | 4.79 | 4.79 | 4.79 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S6: Entropy, Diversity, and Overlap results for the Au (warm) subset of TM23

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 4.91 | 4.90 | 4.99 | 5.14 | 5.21 | 5.24 | 5.31 | 5.53 | 0.96 | 0.96 | 0.96 | 0.96 |
| 25 | 4.93 | 5.02 | 5.05 | 5.18 | 5.33 | 5.43 | 5.48 | 5.66 | 0.99 | 0.99 | 0.99 | 0.99 |
| 50 | 4.99 | 5.03 | 5.03 | 5.14 | 5.46 | 5.51 | 5.52 | 5.66 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75 | 5.00 | 5.02 | 5.02 | 5.08 | 5.50 | 5.53 | 5.53 | 5.61 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100 | 5.00 | 5.00 | 5.00 | 5.00 | 5.53 | 5.53 | 5.53 | 5.53 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S7: Entropy, Diversity, and Overlap results for the Cd (warm) subset of TM23

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 4.67 | 4.67 | 4.73 | 4.83 | 4.99 | 5.00 | 5.09 | 5.28 | 0.99 | 0.99 | 1.00 | 1.00 |
| 25 | 4.71 | 4.73 | 4.76 | 4.85 | 5.08 | 5.12 | 5.17 | 5.35 | 1.00 | 1.00 | 1.00 | 1.00 |
| 50 | 4.72 | 4.74 | 4.75 | 4.82 | 5.13 | 5.18 | 5.18 | 5.31 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75 | 4.73 | 4.74 | 4.74 | 4.78 | 5.16 | 5.18 | 5.18 | 5.24 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100 | 4.73 | 4.73 | 4.73 | 4.73 | 5.17 | 5.17 | 5.17 | 5.17 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S8: Entropy, Diversity, and Overlap results for the Co (warm) subset of TM23

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 7.76 | 7.81 | 7.85 | 7.88 | 7.84 | 7.88 | 7.91 | 7.94 | 0.41 | 0.39 | 0.38 | 0.37 |
| 25 | 8.27 | 8.32 | 8.34 | 8.37 | 8.43 | 8.47 | 8.49 | 8.52 | 0.70 | 0.70 | 0.70 | 0.70 |
| 50 | 8.53 | 8.58 | 8.58 | 8.61 | 8.79 | 8.83 | 8.83 | 8.87 | 0.88 | 0.88 | 0.89 | 0.89 |
| 75 | 8.65 | 8.67 | 8.68 | 8.70 | 8.97 | 8.99 | 8.99 | 9.02 | 0.96 | 0.96 | 0.96 | 0.97 |
| 100 | 8.72 | 8.72 | 8.72 | 8.72 | 9.09 | 9.09 | 9.09 | 9.09 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S9: Entropy, Diversity, and Overlap results for the Ir (warm) subset of TM23

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 3.63 | 3.72 | 3.82 | 3.97 | 4.32 | 4.43 | 4.57 | 4.83 | 0.99 | 0.99 | 0.99 | 0.99 |
| 25 | 3.64 | 3.75 | 3.78 | 3.92 | 4.43 | 4.59 | 4.60 | 4.86 | 1.00 | 1.00 | 1.00 | 1.00 |
| 50 | 3.67 | 3.72 | 3.73 | 3.83 | 4.58 | 4.62 | 4.62 | 4.79 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75 | 3.66 | 3.70 | 3.70 | 3.76 | 4.60 | 4.63 | 4.64 | 4.70 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100 | 3.66 | 3.66 | 3.66 | 3.66 | 4.62 | 4.62 | 4.62 | 4.62 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S10: Entropy, Diversity, and Overlap results for the Pd (warm) subset of TM23

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 4.01 | 4.12 | 4.23 | 4.36 | 4.64 | 4.78 | 4.87 | 5.16 | 0.98 | 0.98 | 0.98 | 0.98 |
| 25 | 4.07 | 4.18 | 4.23 | 4.37 | 4.84 | 4.97 | 5.00 | 5.25 | 0.99 | 0.99 | 1.00 | 1.00 |
| 50 | 4.10 | 4.18 | 4.20 | 4.29 | 4.96 | 5.05 | 5.06 | 5.21 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75 | 4.10 | 4.15 | 4.15 | 4.21 | 5.02 | 5.07 | 5.07 | 5.13 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100 | 4.10 | 4.10 | 4.10 | 4.10 | 5.05 | 5.05 | 5.05 | 5.05 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S11: Entropy, Diversity, and Overlap results for the Ti (warm) subset of TM23

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 6.02 | 6.01 | 6.06 | 6.12 | 6.54 | 6.53 | 6.60 | 6.73 | 0.94 | 0.94 | 0.94 | 0.94 |
| 25 | 6.09 | 6.11 | 6.15 | 6.20 | 6.78 | 6.80 | 6.85 | 7.01 | 0.98 | 0.98 | 0.98 | 0.98 |
| 50 | 6.15 | 6.16 | 6.17 | 6.22 | 6.97 | 6.98 | 6.99 | 7.13 | 0.99 | 0.99 | 0.99 | 1.00 |
| 75 | 6.15 | 6.17 | 6.16 | 6.20 | 7.04 | 7.06 | 7.06 | 7.13 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100 | 6.16 | 6.16 | 6.16 | 6.16 | 7.09 | 7.09 | 7.09 | 7.09 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S12: Force errors for a SevenNet model trained on subsampled datasets of the Ag (warm) subset of TM23. The models were tested on the original test splits corresponding to the subset.

| | Cold Error (eV/Å) | | | | Warm Error (eV/Å) | | | | Melt Error (eV/Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 0.08 | 0.06 | 0.06 | 0.02 | 0.23 | 0.20 | 0.22 | 0.05 | 0.58 | 0.50 | 0.54 | 0.17 |
| 25 | 0.03 | 0.03 | 0.03 | 0.02 | 0.10 | 0.12 | 0.10 | 0.10 | 0.31 | 0.35 | 0.31 | 0.29 |
| 50 | 0.02 | 0.02 | 0.02 | 0.02 | 0.07 | 0.08 | 0.09 | 0.07 | 0.22 | 0.25 | 0.26 | 0.22 |
| 75 | 0.02 | 0.02 | 0.02 | 0.02 | 0.06 | 0.08 | 0.07 | 0.07 | 0.21 | 0.25 | 0.24 | 0.21 |
| 100 | 0.01 | 0.01 | 0.01 | 0.01 | 0.07 | 0.07 | 0.07 | 0.07 | 0.21 | 0.21 | 0.21 | 0.21 |

**Table** S13: Force errors for a SevenNet model trained on subsampled datasets of the Au (warm) subset of TM23. The models were tested on the original test splits corresponding to the subset.

| | Cold Error (eV/Å) | | | | Warm Error (eV/Å) | | | | Melt Error (eV/Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 0.14 | 0.25 | 0.14 | 0.07 | 0.30 | 0.50 | 0.29 | 0.13 | 0.54 | 0.81 | 0.53 | 0.21 |
| 25 | 0.09 | 0.09 | 0.10 | 0.09 | 0.19 | 0.22 | 0.21 | 0.20 | 0.33 | 0.40 | 0.37 | 0.35 |
| 50 | 0.07 | 0.07 | 0.08 | 0.07 | 0.16 | 0.15 | 0.17 | 0.15 | 0.28 | 0.26 | 0.30 | 0.25 |
| 75 | 0.06 | 0.06 | 0.06 | 0.06 | 0.13 | 0.12 | 0.13 | 0.13 | 0.23 | 0.21 | 0.23 | 0.23 |
| 100 | 0.05 | 0.05 | 0.05 | 0.05 | 0.11 | 0.11 | 0.11 | 0.11 | 0.20 | 0.20 | 0.20 | 0.20 |

**Table** S14: Force errors for a SevenNet model trained on subsampled datasets of the Cd (warm) subset of TM23. The models were tested on the original test splits corresponding to the subset.

| | Cold Error (eV/Å) | | | | Warm Error (eV/Å) | | | | Melt Error (eV/Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 0.12 | 0.10 | 0.11 | 0.04 | 0.20 | 0.17 | 0.18 | 0.06 | 0.36 | 0.30 | 0.31 | 0.12 |
| 25 | 0.05 | 0.05 | 0.05 | 0.05 | 0.07 | 0.08 | 0.07 | 0.07 | 0.15 | 0.16 | 0.14 | 0.15 |
| 50 | 0.04 | 0.04 | 0.04 | 0.04 | 0.07 | 0.06 | 0.07 | 0.07 | 0.14 | 0.12 | 0.14 | 0.14 |
| 75 | 0.04 | 0.04 | 0.04 | 0.04 | 0.06 | 0.06 | 0.06 | 0.07 | 0.13 | 0.11 | 0.13 | 0.13 |
| 100 | 0.04 | 0.04 | 0.04 | 0.04 | 0.06 | 0.06 | 0.06 | 0.06 | 0.13 | 0.13 | 0.13 | 0.13 |

**Table** S15: Force errors for a SevenNet model trained on subsampled datasets of the Co (warm) subset of TM23. The models were tested on the original test splits corresponding to the subset.

| | Cold Error (eV/Å) | | | | Warm Error (eV/Å) | | | | Melt Error (eV/Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 0.32 | 0.29 | 0.31 | 0.18 | 0.64 | 0.63 | 0.61 | 0.30 | 1.22 | 1.26 | 1.12 | 0.61 |
| 25 | 0.21 | 0.20 | 0.21 | 0.20 | 0.38 | 0.36 | 0.39 | 0.34 | 0.80 | 0.70 | 0.79 | 0.66 |
| 50 | 0.18 | 0.18 | 0.18 | 0.18 | 0.31 | 0.33 | 0.30 | 0.31 | 0.61 | 0.65 | 0.54 | 0.58 |
| 75 | 0.18 | 0.18 | 0.18 | 0.18 | 0.29 | 0.28 | 0.29 | 0.28 | 0.56 | 0.51 | 0.59 | 0.49 |
| 100 | 0.17 | 0.17 | 0.17 | 0.17 | 0.28 | 0.28 | 0.28 | 0.28 | 0.53 | 0.53 | 0.53 | 0.53 |

**Table** S16: Force errors for a SevenNet model trained on subsampled datasets of the Ir (warm) subset of TM23. The models were tested on the original test splits corresponding to the subset.

| | Cold Error (eV/Å) | | | | Warm Error (eV/Å) | | | | Melt Error (eV/Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 0.41 | 0.42 | 0.34 | 0.20 | 0.82 | 0.84 | 0.63 | 0.34 | 1.54 | 1.46 | 1.18 | 0.64 |
| 25 | 0.25 | 0.23 | 0.24 | 0.26 | 0.51 | 0.49 | 0.54 | 0.51 | 0.96 | 0.94 | 1.07 | 1.00 |
| 50 | 0.21 | 0.21 | 0.21 | 0.21 | 0.39 | 0.39 | 0.43 | 0.39 | 0.74 | 0.73 | 0.82 | 0.72 |
| 75 | 0.20 | 0.20 | 0.20 | 0.20 | 0.36 | 0.37 | 0.40 | 0.40 | 0.63 | 0.68 | 0.76 | 0.80 |
| 100 | 0.19 | 0.19 | 0.19 | 0.19 | 0.36 | 0.36 | 0.36 | 0.36 | 0.71 | 0.71 | 0.71 | 0.71 |

**Table** S17: Force errors for a SevenNet model trained on subsampled datasets of the Pd (warm) subset of TM23. The models were tested on the original test splits corresponding to the subset.

| | Cold Error (eV/Å) | | | | Warm Error (eV/Å) | | | | Melt Error (eV/Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 0.18 | 0.21 | 0.20 | 0.07 | 0.52 | 0.54 | 0.50 | 0.17 | 1.14 | 1.18 | 1.15 | 0.48 |
| 25 | 0.09 | 0.09 | 0.09 | 0.09 | 0.31 | 0.27 | 0.27 | 0.26 | 0.79 | 0.69 | 0.69 | 0.68 |
| 50 | 0.07 | 0.07 | 0.07 | 0.07 | 0.19 | 0.20 | 0.20 | 0.20 | 0.52 | 0.54 | 0.53 | 0.54 |
| 75 | 0.06 | 0.07 | 0.07 | 0.07 | 0.16 | 0.15 | 0.17 | 0.15 | 0.44 | 0.42 | 0.47 | 0.44 |
| 100 | 0.06 | 0.06 | 0.06 | 0.06 | 0.16 | 0.16 | 0.16 | 0.16 | 0.48 | 0.48 | 0.48 | 0.48 |

**Table** S18: Force errors for a SevenNet model trained on subsampled datasets of the Ti (warm) subset of TM23. The models were tested on the original test splits corresponding to the subset.

| | Cold Error (eV/Å) | | | | Warm Error (eV/Å) | | | | Melt Error (eV/Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10 | 0.36 | 0.39 | 0.37 | 0.26 | 0.55 | 0.64 | 0.59 | 0.37 | 1.22 | 1.28 | 1.27 | 0.93 |
| 25 | 0.31 | 0.31 | 0.31 | 0.30 | 0.46 | 0.44 | 0.46 | 0.44 | 1.10 | 1.03 | 1.08 | 1.00 |
| 50 | 0.27 | 0.28 | 0.27 | 0.26 | 0.41 | 0.40 | 0.39 | 0.40 | 1.08 | 1.00 | 0.92 | 1.00 |
| 75 | 0.25 | 0.25 | 0.25 | 0.25 | 0.39 | 0.39 | 0.42 | 0.39 | 1.01 | 1.10 | 1.12 | 0.98 |
| 100 | 0.24 | 0.24 | 0.24 | 0.24 | 0.37 | 0.37 | 0.37 | 0.37 | 0.88 | 0.88 | 0.88 | 0.88 |

## S3.3. Results of dataset compression for 64 different datasets obtained from ColabFit

**Table** S19: Names of the 64 single-component datasets obtained from the ColabFit repository. The dataset numbers provided here are used in Figs. S25–S32 for brevity.

| Number | Name in ColabFit | DOI |
|---|---|---|
| 0 | 23-Single-Element-DNPs RSCDD 2023-Ag | 10.60732/7cf58a2f [1] |
| 1 | 23-Single-Element-DNPs RSCDD 2023-Al | 10.60732/ef6f5966 [2] |
| 2 | 23-Single-Element-DNPs RSCDD 2023-Au | 10.60732/7c0dfbc8 [3] |
| 3 | 23-Single-Element-DNPs RSCDD 2023-Co | 10.60732/15bb1dca [4] |
| 4 | 23-Single-Element-DNPs RSCDD 2023-Cu | 10.60732/e0a72dd8 [5] |
| 5 | 23-Single-Element-DNPs RSCDD 2023-Ge | 10.60732/33767c4f [6] |
| 6 | 23-Single-Element-DNPs RSCDD 2023-I | 10.60732/57c54149 [7] |
| 7 | 23-Single-Element-DNPs RSCDD 2023-Kr | 10.60732/6a060a77 [8] |
| 8 | 23-Single-Element-DNPs RSCDD 2023-Li | 10.60732/f5cc2a19 [9] |
| 9 | 23-Single-Element-DNPs RSCDD 2023-Mg | 10.60732/f4d2f0b8 [10] |
| 10 | 23-Single-Element-DNPs RSCDD 2023-Mo | 10.60732/b6ece7fd [11] |
| 11 | 23-Single-Element-DNPs RSCDD 2023-Nb | 10.60732/2146db76 [12] |
| 12 | 23-Single-Element-DNPs RSCDD 2023-Ni | 10.60732/b184fffc [13] |
| 13 | 23-Single-Element-DNPs RSCDD 2023-Os | 10.60732/1ec0df98 [14] |
| 14 | 23-Single-Element-DNPs RSCDD 2023-Pb | 10.60732/bddd3245 [15] |
| 15 | 23-Single-Element-DNPs RSCDD 2023-Pd | 10.60732/8d5bdb05 [16] |
| 16 | 23-Single-Element-DNPs RSCDD 2023-Pt | 10.60732/49b97320 [17] |
| 17 | 23-Single-Element-DNPs RSCDD 2023-Re | 10.60732/1e0997f8 [18] |
| 18 | 23-Single-Element-DNPs RSCDD 2023-Sb | 10.60732/7980ece8 [19] |
| 19 | 23-Single-Element-DNPs RSCDD 2023-Sr | 10.60732/69de7b6b [20] |
| 20 | 23-Single-Element-DNPs RSCDD 2023-Ti | 10.60732/f08eba7c [21] |
| 21 | 23-Single-Element-DNPs RSCDD 2023-Zn | 10.60732/dfa1e792 [22] |
| 22 | 23-Single-Element-DNPs RSCDD 2023-Zr | 10.60732/efb626b6 [23] |
| 23 | Ag-PBE MSMSE 2021 | 10.60732/93adbc95 [24] |
| 24 | Au-PBE MSMSE 2021 | 10.60732/c4492535 [25] |
| 25 | CA-9 | 10.60732/8b765383 [26] |
| 26 | CA-9 BB | 10.60732/f3bbbd36 [27] |
| 27 | CA-9 BR | 10.60732/07b7d297 [28] |
| 28 | CA-9 RR | 10.60732/4096ff5c [29] |
| 29 | CGM-MLP natcomm2023 screening amorphous carbon | 10.60732/f340d1d9 [30] |
| 30 | CGM-MLP natcomm2023 screening graphite train | 10.60732/85590078 [31] |
| 31 | C-NPJ2020 | 10.60732/e65112ef [32] |
| 32 | DP-GEN Cu | 10.60732/2060021e [33] |
| 33 | Fe nanoparticles PRB 2023 | 10.60732/20ba88af [34] |

Continued on next page

**Table S19 (continued)**

| Number | Name in ColabFit | DOI |
|---|---|---|
| 34 | FitSNAP Fe NPJ 2021 | 10.60732/fe28ef5e [35] |
| 35 | Mg edmonds 2022 | 10.60732/4b13be86 [36] |
| 36 | Mo PRM2019 | 10.60732/31dbb6ee [37] |
| 37 | NDSC TUT 2022 | 10.60732/965983d1 [38] |
| 38 | NEP qHPF | 10.60732/906d79f3 [39] |
| 39 | Nb PRM2019 | 10.60732/a90f7f6e [40] |
| 40 | Si JCP 2017 | 10.60732/68b1a5ad [41] |
| 41 | Si PRX GAP | 10.60732/8e9bc5b0 [42] |
| 42 | Sn-SCAN PRM 2023 | 10.60732/7d8a06fe [43] |
| 43 | Ta Linear JCP2014 | 10.60732/da9afef7 [44] |
| 44 | Ta PINN 2021 | 10.60732/7f6cac29 [45] |
| 45 | Ta PRM2019 | 10.60732/43837a12 [46] |
| 46 | Ti NPJCM 2021 | 10.60732/85e47ff3 [47] |
| 47 | V PRM2019 | 10.60732/aad06a25 [48] |
| 48 | W-14 | 10.60732/8d093f34 [49] |
| 49 | W LML-retrain bulk MD test | 10.60732/9d48595f [50] |
| 50 | W PRB2019 | 10.60732/6367ea51 [51] |
| 51 | Zn MTP CMS2023 | 10.60732/54902e18 [52] |
| 52 | aC JCP 2023 | 10.60732/eeb61a0d [53] |
| 53 | defected phosphorene ACS 2023 | 10.60732/87b2341a [54] |
| 54 | discrepancies and error metrics NPJ 2023 enhanced validation set | 10.60732/9c77bb8c [55] |
| 55 | discrepancies and error metrics NPJ 2023 interstitial enhanced training set | 10.60732/f0a44294 [56] |
| 56 | discrepancies and error metrics NPJ 2023 vacancy enhanced training set | 10.60732/5a780a3a [57] |
| 57 | mlearn Cu | 10.60732/49de06ae [58] |
| 58 | mlearn Ge | 10.60732/4552d3fd [59] |
| 59 | mlearn Li | 10.60732/63ab9206 [60] |
| 60 | mlearn Mo | 10.60732/3827e5e1 [61] |
| 61 | mlearn Ni | 10.60732/9a25df21 [62] |
| 62 | mlearn Si | 10.60732/c2471ffc [63] |
| 63 | pure magnesium DFT PRM2020 | 10.60732/28f038f2 [64] |

**Table** S20: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Ag

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.04 | 6.65 | 6.71 | 7.25 | 7.19 | 7.56 | 7.57 | 7.77 | 0.89 | 0.96 | 0.95 | 0.97 |
| 25.0 | 6.19 | 6.54 | 6.65 | 7.34 | 7.52 | 7.69 | 7.73 | 7.78 | 0.98 | 1.00 | 1.00 | 1.00 |
| 50.0 | 6.23 | 6.45 | 6.51 | 7.14 | 7.66 | 7.71 | 7.73 | 7.76 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.28 | 6.39 | 6.39 | 6.74 | 7.70 | 7.71 | 7.72 | 7.72 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.26 | 6.26 | 6.26 | 6.26 | 7.71 | 7.71 | 7.71 | 7.71 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S21: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Al

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.34 | 7.58 | 7.68 | 7.96 | 7.91 | 8.09 | 8.13 | 8.45 | 0.67 | 0.69 | 0.67 | 0.59 |
| 25.0 | 7.79 | 8.07 | 8.17 | 8.49 | 8.56 | 8.70 | 8.72 | 9.00 | 0.86 | 0.89 | 0.88 | 0.89 |
| 50.0 | 7.94 | 8.33 | 8.38 | 8.51 | 8.87 | 9.04 | 9.06 | 9.15 | 0.94 | 0.98 | 0.97 | 1.00 |
| 75.0 | 7.98 | 8.21 | 8.21 | 8.29 | 9.03 | 9.13 | 9.13 | 9.14 | 0.98 | 1.00 | 1.00 | 1.00 |
| 100.0 | 8.02 | 8.02 | 8.02 | 8.02 | 9.12 | 9.12 | 9.12 | 9.12 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S22: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Au

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.56 | 7.16 | 7.41 | 7.81 | 7.26 | 7.68 | 7.84 | 8.02 | 0.92 | 0.96 | 0.95 | 0.96 |
| 25.0 | 6.80 | 7.24 | 7.26 | 7.71 | 7.58 | 7.96 | 8.00 | 8.01 | 0.98 | 1.00 | 1.00 | 1.00 |
| 50.0 | 6.92 | 7.11 | 7.13 | 7.55 | 7.77 | 7.93 | 7.94 | 7.95 | 0.99 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.94 | 7.07 | 7.07 | 7.34 | 7.82 | 7.90 | 7.91 | 7.90 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.96 | 6.96 | 6.96 | 6.96 | 7.88 | 7.88 | 7.88 | 7.88 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S23: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Co

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.53 | 7.86 | 7.80 | 8.17 | 7.98 | 8.22 | 8.18 | 8.46 | 0.53 | 0.50 | 0.46 | 0.35 |
| 25.0 | 8.00 | 8.34 | 8.37 | 8.78 | 8.61 | 8.88 | 8.90 | 9.10 | 0.78 | 0.80 | 0.76 | 0.87 |
| 50.0 | 8.20 | 8.36 | 8.40 | 8.84 | 8.92 | 9.10 | 9.13 | 9.16 | 0.94 | 0.97 | 0.97 | 0.99 |
| 75.0 | 8.29 | 8.39 | 8.39 | 8.76 | 9.07 | 9.13 | 9.15 | 9.16 | 0.98 | 0.99 | 1.00 | 1.00 |
| 100.0 | 8.35 | 8.35 | 8.35 | 8.35 | 9.14 | 9.14 | 9.14 | 9.14 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S24: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Cu

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.72 | 8.21 | 8.28 | 8.51 | 8.28 | 8.53 | 8.61 | 8.86 | 0.62 | 0.64 | 0.58 | 0.43 |
| 25.0 | 8.08 | 8.72 | 8.87 | 9.11 | 8.85 | 9.20 | 9.27 | 9.49 | 0.80 | 0.84 | 0.82 | 0.81 |
| 50.0 | 8.30 | 8.92 | 8.96 | 9.13 | 9.27 | 9.57 | 9.58 | 9.63 | 0.92 | 0.97 | 0.97 | 0.99 |
| 75.0 | 8.41 | 8.61 | 8.61 | 8.92 | 9.48 | 9.61 | 9.62 | 9.61 | 0.97 | 1.00 | 1.00 | 1.00 |
| 100.0 | 8.43 | 8.43 | 8.43 | 8.43 | 9.59 | 9.59 | 9.59 | 9.59 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S25: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Ge

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.79 | 7.29 | 7.53 | 8.23 | 8.05 | 8.26 | 8.33 | 8.83 | 0.88 | 0.90 | 0.89 | 0.85 |
| 25.0 | 6.75 | 7.81 | 8.02 | 8.35 | 8.45 | 8.86 | 8.93 | 9.05 | 0.95 | 0.98 | 0.98 | 0.99 |
| 50.0 | 6.92 | 7.65 | 7.72 | 7.93 | 8.75 | 9.03 | 9.03 | 9.01 | 0.98 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.88 | 7.38 | 7.38 | 7.50 | 8.88 | 8.99 | 8.99 | 8.98 | 0.99 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.90 | 6.90 | 6.90 | 6.90 | 8.97 | 8.97 | 8.97 | 8.97 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S26: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-I

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.56 | 5.59 | 5.79 | 6.25 | 6.18 | 6.51 | 6.60 | 6.70 | 0.97 | 0.99 | 0.99 | 1.00 |
| 25.0 | 4.73 | 5.58 | 5.64 | 6.11 | 6.43 | 6.66 | 6.67 | 6.70 | 0.99 | 1.00 | 1.00 | 1.00 |
| 50.0 | 4.74 | 5.26 | 5.29 | 5.71 | 6.51 | 6.63 | 6.63 | 6.65 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 4.73 | 4.97 | 5.01 | 5.25 | 6.57 | 6.61 | 6.61 | 6.62 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 4.76 | 4.76 | 4.76 | 4.76 | 6.61 | 6.61 | 6.61 | 6.61 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S27: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Kr

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 3.15 | 4.54 | 4.70 | 5.10 | 1.65 | 1.72 | 1.73 | 1.74 | 0.99 | 1.00 | 1.00 | 1.00 |
| 25.0 | 3.28 | 4.38 | 4.49 | 4.86 | 1.69 | 1.73 | 1.73 | 1.74 | 0.99 | 1.00 | 1.00 | 1.00 |
| 50.0 | 3.32 | 4.05 | 4.07 | 4.55 | 1.71 | 1.73 | 1.73 | 1.73 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 3.31 | 3.49 | 3.49 | 3.99 | 1.73 | 1.73 | 1.73 | 1.73 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 3.32 | 3.32 | 3.32 | 3.32 | 1.73 | 1.73 | 1.73 | 1.73 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S28: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Li

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.15 | 7.73 | 7.95 | 8.56 | 7.96 | 8.28 | 8.37 | 8.75 | 0.77 | 0.78 | 0.75 | 0.63 |
| 25.0 | 7.25 | 8.05 | 8.18 | 8.73 | 8.40 | 8.87 | 8.92 | 9.10 | 0.88 | 0.91 | 0.91 | 0.98 |
| 50.0 | 7.44 | 7.98 | 8.01 | 8.42 | 8.79 | 9.06 | 9.08 | 9.11 | 0.95 | 0.99 | 0.99 | 1.00 |
| 75.0 | 7.48 | 7.84 | 7.86 | 8.10 | 8.96 | 9.08 | 9.09 | 9.10 | 0.98 | 1.00 | 1.00 | 1.00 |
| 100.0 | 7.53 | 7.53 | 7.53 | 7.53 | 9.09 | 9.09 | 9.09 | 9.09 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S29: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Mg

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 5.71 | 6.24 | 6.43 | 6.44 | 6.77 | 7.01 | 7.05 | 7.30 | 0.89 | 0.94 | 0.94 | 0.89 |
| 25.0 | 5.88 | 6.33 | 6.48 | 6.58 | 7.11 | 7.29 | 7.33 | 7.45 | 0.97 | 0.99 | 0.99 | 1.00 |
| 50.0 | 5.97 | 6.30 | 6.35 | 6.67 | 7.30 | 7.38 | 7.40 | 7.45 | 0.99 | 1.00 | 1.00 | 1.00 |
| 75.0 | 5.99 | 6.19 | 6.19 | 6.47 | 7.37 | 7.41 | 7.41 | 7.42 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.99 | 5.99 | 5.99 | 5.99 | 7.41 | 7.41 | 7.41 | 7.41 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S30: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Mo

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.16 | 7.50 | 7.44 | 8.11 | 7.76 | 7.95 | 7.89 | 8.32 | 0.65 | 0.68 | 0.61 | 0.68 |
| 25.0 | 7.59 | 7.82 | 7.86 | 8.38 | 8.25 | 8.45 | 8.49 | 8.57 | 0.88 | 0.94 | 0.91 | 0.96 |
| 50.0 | 7.70 | 7.81 | 7.88 | 8.37 | 8.47 | 8.55 | 8.58 | 8.60 | 0.99 | 0.99 | 1.00 | 0.99 |
| 75.0 | 7.71 | 7.74 | 7.75 | 8.22 | 8.54 | 8.56 | 8.57 | 8.59 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 7.74 | 7.74 | 7.74 | 7.74 | 8.58 | 8.58 | 8.58 | 8.58 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S31: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Nb

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.62 | 7.07 | 7.18 | 7.56 | 1.98 | 2.02 | 2.03 | 2.06 | 0.78 | 0.86 | 0.84 | 0.86 |
| 25.0 | 6.80 | 7.22 | 7.36 | 7.73 | 2.03 | 2.06 | 2.07 | 2.08 | 0.93 | 0.98 | 0.98 | 0.99 |
| 50.0 | 6.92 | 7.19 | 7.25 | 7.62 | 2.06 | 2.07 | 2.07 | 2.08 | 0.99 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.91 | 7.12 | 7.12 | 7.32 | 2.07 | 2.07 | 2.07 | 2.07 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.90 | 6.90 | 6.90 | 6.90 | 2.07 | 2.07 | 2.07 | 2.07 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S32: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Ni

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.49 | 7.75 | 7.86 | 8.39 | 8.08 | 8.18 | 8.23 | 8.65 | 0.56 | 0.59 | 0.50 | 0.41 |
| 25.0 | 7.94 | 8.57 | 8.79 | 9.23 | 8.75 | 9.09 | 9.19 | 9.47 | 0.77 | 0.80 | 0.76 | 0.76 |
| 50.0 | 8.17 | 8.84 | 8.90 | 9.16 | 9.14 | 9.54 | 9.57 | 9.57 | 0.90 | 0.97 | 0.97 | 1.00 |
| 75.0 | 8.29 | 8.62 | 8.65 | 8.85 | 9.39 | 9.55 | 9.56 | 9.55 | 0.96 | 1.00 | 1.00 | 1.00 |
| 100.0 | 8.35 | 8.35 | 8.35 | 8.35 | 9.54 | 9.54 | 9.54 | 9.54 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S33: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Os

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.47 | 7.06 | 7.23 | 7.74 | 7.58 | 7.91 | 7.99 | 8.24 | 0.86 | 0.91 | 0.91 | 0.96 |
| 25.0 | 6.72 | 7.05 | 7.15 | 7.88 | 7.99 | 8.15 | 8.19 | 8.28 | 0.96 | 0.98 | 0.98 | 1.00 |
| 50.0 | 6.72 | 6.95 | 6.98 | 7.63 | 8.13 | 8.20 | 8.23 | 8.26 | 0.99 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.73 | 6.90 | 6.92 | 7.31 | 8.19 | 8.21 | 8.22 | 8.23 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.73 | 6.73 | 6.73 | 6.73 | 8.22 | 8.22 | 8.22 | 8.22 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S34: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Pb

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 5.20 | 5.98 | 6.15 | 6.35 | 6.25 | 6.71 | 6.76 | 6.98 | 0.97 | 0.99 | 0.98 | 1.00 |
| 25.0 | 5.32 | 5.96 | 5.96 | 6.15 | 6.57 | 6.90 | 6.92 | 6.93 | 0.99 | 1.00 | 1.00 | 1.00 |
| 50.0 | 5.35 | 5.79 | 5.80 | 5.88 | 6.74 | 6.88 | 6.88 | 6.88 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 5.38 | 5.54 | 5.54 | 5.75 | 6.80 | 6.85 | 6.85 | 6.85 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.39 | 5.39 | 5.39 | 5.39 | 6.83 | 6.83 | 6.83 | 6.83 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S35: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Pd

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.07 | 7.59 | 7.89 | 8.78 | 8.14 | 8.36 | 8.46 | 9.00 | 0.79 | 0.80 | 0.77 | 0.75 |
| 25.0 | 7.26 | 8.22 | 8.46 | 9.00 | 8.67 | 9.05 | 9.13 | 9.32 | 0.90 | 0.95 | 0.93 | 0.97 |
| 50.0 | 7.40 | 8.15 | 8.19 | 8.57 | 8.96 | 9.29 | 9.30 | 9.25 | 0.97 | 1.00 | 1.00 | 1.00 |
| 75.0 | 7.43 | 7.75 | 7.77 | 8.07 | 9.12 | 9.27 | 9.27 | 9.23 | 0.99 | 1.00 | 1.00 | 1.00 |
| 100.0 | 7.43 | 7.43 | 7.43 | 7.43 | 9.22 | 9.22 | 9.22 | 9.22 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S36: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Pt

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.46 | 7.00 | 7.22 | 7.62 | 7.23 | 7.54 | 7.65 | 8.05 | 0.80 | 0.83 | 0.80 | 0.71 |
| 25.0 | 6.82 | 7.59 | 7.73 | 7.92 | 7.74 | 8.29 | 8.35 | 8.43 | 0.92 | 0.96 | 0.96 | 0.99 |
| 50.0 | 6.99 | 7.35 | 7.36 | 7.69 | 8.06 | 8.41 | 8.43 | 8.39 | 0.97 | 1.00 | 1.00 | 1.00 |
| 75.0 | 7.01 | 7.15 | 7.15 | 7.39 | 8.21 | 8.39 | 8.39 | 8.36 | 0.99 | 1.00 | 1.00 | 1.00 |
| 100.0 | 7.03 | 7.03 | 7.03 | 7.03 | 8.35 | 8.35 | 8.35 | 8.35 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S37: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Re

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.81 | 8.14 | 8.15 | 8.33 | 8.26 | 8.50 | 8.50 | 8.71 | 0.68 | 0.72 | 0.66 | 0.63 |
| 25.0 | 8.12 | 8.25 | 8.34 | 8.55 | 8.69 | 8.90 | 8.96 | 9.07 | 0.90 | 0.94 | 0.91 | 0.97 |
| 50.0 | 8.20 | 8.16 | 8.18 | 8.58 | 8.89 | 9.05 | 9.07 | 9.08 | 0.98 | 0.99 | 1.00 | 1.00 |
| 75.0 | 8.27 | 8.25 | 8.26 | 8.54 | 8.99 | 9.05 | 9.06 | 9.06 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 8.29 | 8.29 | 8.29 | 8.29 | 9.04 | 9.04 | 9.04 | 9.04 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S38: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Sb

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.41 | 6.82 | 6.92 | 7.14 | 7.08 | 7.37 | 7.40 | 7.70 | 0.96 | 0.98 | 0.96 | 0.99 |
| 25.0 | 6.54 | 6.86 | 6.84 | 7.04 | 7.36 | 7.59 | 7.61 | 7.70 | 0.99 | 0.99 | 0.99 | 1.00 |
| 50.0 | 6.57 | 6.84 | 6.85 | 6.90 | 7.48 | 7.66 | 7.67 | 7.66 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.59 | 6.74 | 6.75 | 6.78 | 7.56 | 7.63 | 7.63 | 7.63 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.62 | 6.62 | 6.62 | 6.62 | 7.61 | 7.61 | 7.61 | 7.61 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S39: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Sr

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.14 | 4.50 | 4.67 | 4.68 | 5.06 | 5.36 | 5.39 | 5.37 | 0.98 | 1.00 | 0.99 | 1.00 |
| 25.0 | 4.16 | 4.51 | 4.54 | 4.66 | 5.27 | 5.39 | 5.40 | 5.40 | 1.00 | 1.00 | 1.00 | 1.00 |
| 50.0 | 4.18 | 4.43 | 4.45 | 4.60 | 5.34 | 5.38 | 5.38 | 5.39 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 4.17 | 4.22 | 4.22 | 4.46 | 5.37 | 5.38 | 5.38 | 5.38 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 4.16 | 4.16 | 4.16 | 4.16 | 5.38 | 5.38 | 5.38 | 5.38 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S40: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Ti

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 8.35 | 8.62 | 8.81 | 9.27 | 8.76 | 8.98 | 9.08 | 9.50 | 0.62 | 0.64 | 0.57 | 0.47 |
| 25.0 | 8.69 | 9.26 | 9.39 | 9.59 | 9.33 | 9.70 | 9.77 | 9.97 | 0.83 | 0.88 | 0.83 | 0.85 |
| 50.0 | 8.87 | 9.35 | 9.36 | 9.50 | 9.69 | 10.02 | 10.04 | 10.05 | 0.93 | 0.99 | 0.99 | 1.00 |
| 75.0 | 8.94 | 9.18 | 9.18 | 9.32 | 9.88 | 10.03 | 10.03 | 10.03 | 0.98 | 1.00 | 1.00 | 1.00 |
| 100.0 | 8.98 | 8.98 | 8.98 | 8.98 | 10.01 | 10.01 | 10.01 | 10.01 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S41: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Zn

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.26 | 7.02 | 7.16 | 7.42 | 2.00 | 2.04 | 2.05 | 2.07 | 0.89 | 0.94 | 0.93 | 0.95 |
| 25.0 | 6.39 | 6.99 | 7.13 | 7.51 | 2.04 | 2.07 | 2.07 | 2.08 | 0.97 | 0.99 | 0.99 | 1.00 |
| 50.0 | 6.50 | 6.95 | 7.03 | 7.46 | 2.06 | 2.07 | 2.08 | 2.08 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.54 | 6.74 | 6.75 | 7.19 | 2.07 | 2.07 | 2.07 | 2.08 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.56 | 6.56 | 6.56 | 6.56 | 2.07 | 2.07 | 2.07 | 2.07 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S42: Entropy, Diversity, and Overlap results for dataset 23-Single-Element-DNPs RSCDD 2023-Zr

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.49 | 7.00 | 7.08 | 7.24 | 7.26 | 7.63 | 7.67 | 7.89 | 0.90 | 0.93 | 0.90 | 0.94 |
| 25.0 | 6.67 | 7.00 | 7.04 | 7.21 | 7.55 | 7.88 | 7.91 | 7.93 | 0.97 | 0.99 | 0.99 | 0.99 |
| 50.0 | 6.75 | 6.82 | 6.85 | 7.10 | 7.77 | 7.91 | 7.93 | 7.94 | 0.99 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.74 | 6.88 | 6.88 | 7.03 | 7.85 | 7.92 | 7.93 | 7.94 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.74 | 6.74 | 6.74 | 6.74 | 7.92 | 7.92 | 7.92 | 7.92 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S43: Entropy, Diversity, and Overlap results for dataset Ag-PBE MSMSE 2021

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.68 | 6.92 | 8.04 | 8.57 | 7.65 | 8.32 | 8.49 | 8.80 | 0.85 | 0.91 | 0.91 | 0.83 |
| 25.0 | 4.95 | 7.48 | 8.34 | 8.70 | 8.36 | 8.97 | 9.05 | 9.14 | 0.93 | 0.98 | 0.99 | 0.97 |
| 50.0 | 4.98 | 6.65 | 6.67 | 7.07 | 8.75 | 9.13 | 9.13 | 9.13 | 0.97 | 1.00 | 1.00 | 1.00 |
| 75.0 | 4.97 | 5.55 | 5.55 | 5.97 | 8.97 | 9.13 | 9.13 | 9.13 | 0.99 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.00 | 5.00 | 5.00 | 5.00 | 9.13 | 9.13 | 9.13 | 9.13 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S44: Entropy, Diversity, and Overlap results for dataset Au-PBE MSMSE 2021

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.75 | 7.90 | 8.39 | 8.75 | 2.15 | 2.17 | 2.18 | 2.21 | 0.90 | 0.91 | 0.91 | 0.89 |
| 25.0 | 6.84 | 8.21 | 8.54 | 8.79 | 2.19 | 2.22 | 2.22 | 2.23 | 0.97 | 0.99 | 0.99 | 0.99 |
| 50.0 | 6.87 | 6.90 | 6.91 | 8.76 | 2.21 | 2.23 | 2.23 | 2.23 | 0.99 | 1.00 | 1.00 | 1.00 |
| 75.0 | 6.90 | 6.90 | 6.91 | 8.09 | 2.22 | 2.23 | 2.23 | 2.23 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 6.92 | 6.92 | 6.92 | 6.92 | 2.23 | 2.23 | 2.23 | 2.23 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S45: Entropy, Diversity, and Overlap results for dataset CA-9

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 8.40 | 9.94 | 10.16 | 10.97 | 10.86 | 11.45 | 11.53 | 11.71 | 0.80 | 0.84 | 0.85 | 0.87 |
| 25.0 | 8.58 | 9.45 | 9.46 | 10.67 | 11.48 | 11.80 | 11.82 | 12.02 | 0.88 | 0.93 | 0.92 | 0.98 |
| 50.0 | 8.66 | 9.01 | 8.99 | 10.48 | 11.84 | 11.93 | 11.96 | 12.09 | 0.96 | 0.97 | 0.97 | 1.00 |
| 75.0 | 8.69 | 8.90 | 8.90 | 9.58 | 11.99 | 12.01 | 12.03 | 12.08 | 0.99 | 0.99 | 0.99 | 1.00 |
| 100.0 | 8.71 | 8.71 | 8.71 | 8.71 | 12.08 | 12.08 | 12.08 | 12.08 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S46: Entropy, Diversity, and Overlap results for dataset CA-9 BB

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 9.40 | 10.37 | 10.48 | 11.02 | 10.80 | 11.20 | 11.24 | 11.37 | 0.65 | 0.67 | 0.67 | 0.69 |
| 25.0 | 9.66 | 10.17 | 10.12 | 11.04 | 11.43 | 11.69 | 11.70 | 11.88 | 0.78 | 0.83 | 0.82 | 0.90 |
| 50.0 | 9.78 | 9.67 | 9.63 | 11.00 | 11.79 | 11.87 | 11.90 | 12.03 | 0.92 | 0.93 | 0.93 | 1.00 |
| 75.0 | 9.83 | 9.57 | 9.56 | 10.64 | 11.95 | 11.96 | 11.98 | 12.04 | 0.98 | 0.98 | 0.98 | 1.00 |
| 100.0 | 9.87 | 9.87 | 9.87 | 9.87 | 12.04 | 12.04 | 12.04 | 12.04 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S47: Entropy, Diversity, and Overlap results for dataset CA-9 BR

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 9.37 | 10.35 | 10.48 | 10.99 | 10.78 | 11.19 | 11.24 | 11.37 | 0.65 | 0.67 | 0.67 | 0.69 |
| 25.0 | 9.63 | 10.16 | 10.13 | 11.05 | 11.40 | 11.68 | 11.70 | 11.88 | 0.78 | 0.83 | 0.82 | 0.90 |
| 50.0 | 9.77 | 9.66 | 9.63 | 10.99 | 11.78 | 11.87 | 11.90 | 12.03 | 0.92 | 0.93 | 0.93 | 1.00 |
| 75.0 | 9.83 | 9.57 | 9.56 | 10.64 | 11.94 | 11.96 | 11.98 | 12.04 | 0.98 | 0.98 | 0.98 | 1.00 |
| 100.0 | 9.87 | 9.87 | 9.87 | 9.87 | 12.04 | 12.04 | 12.04 | 12.04 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S48: Entropy, Diversity, and Overlap results for dataset CA-9 RR

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 8.69 | 9.88 | 10.30 | 10.89 | 10.49 | 11.05 | 11.19 | 11.34 | 0.76 | 0.78 | 0.78 | 0.80 |
| 25.0 | 8.87 | 9.69 | 9.81 | 10.76 | 11.15 | 11.57 | 11.62 | 11.82 | 0.83 | 0.89 | 0.89 | 0.95 |
| 50.0 | 9.00 | 9.22 | 9.23 | 10.47 | 11.58 | 11.76 | 11.79 | 11.93 | 0.92 | 0.94 | 0.94 | 1.00 |
| 75.0 | 9.05 | 8.97 | 8.96 | 9.71 | 11.80 | 11.85 | 11.86 | 11.92 | 0.97 | 0.97 | 0.98 | 1.00 |
| 100.0 | 9.08 | 9.08 | 9.08 | 9.08 | 11.92 | 11.92 | 11.92 | 11.92 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S49: Entropy, Diversity, and Overlap results for dataset CGM-MLP natcomm2023 screening amorphous carbon

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 9.92 | 9.97 | 9.66 | 10.05 | 2.29 | 2.30 | 2.27 | 2.31 | 0.06 | 0.06 | 0.05 | 0.06 |
| 25.0 | 10.81 | 10.79 | 10.60 | 10.97 | 2.38 | 2.38 | 2.36 | 2.40 | 0.17 | 0.16 | 0.14 | 0.18 |
| 50.0 | 11.49 | 11.44 | 11.37 | 11.63 | 2.44 | 2.44 | 2.43 | 2.45 | 0.39 | 0.35 | 0.33 | 0.40 |
| 75.0 | 11.89 | 11.85 | 11.84 | 11.97 | 2.48 | 2.47 | 2.47 | 2.48 | 0.60 | 0.57 | 0.56 | 0.61 |
| 100.0 | 12.18 | 12.18 | 12.18 | 12.18 | 2.50 | 2.50 | 2.50 | 2.50 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S50: Entropy, Diversity, and Overlap results for dataset CGM-MLP natcomm2023 screening graphite train

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 2.29 | 2.31 | 2.30 | 2.31 | 2.29 | 2.32 | 2.30 | 2.31 | 0.28 | 0.23 | 0.23 | 0.28 |
| 25.0 | 2.97 | 3.10 | 3.10 | 3.10 | 2.99 | 3.10 | 3.10 | 3.11 | 0.44 | 0.47 | 0.59 | 0.57 |
| 50.0 | 3.55 | 3.74 | 3.74 | 3.75 | 3.57 | 3.74 | 3.74 | 3.75 | 0.70 | 0.91 | 0.90 | 0.95 |
| 75.0 | 3.90 | 4.03 | 4.06 | 4.08 | 3.93 | 4.04 | 4.07 | 4.08 | 0.93 | 0.99 | 1.00 | 1.00 |
| 100.0 | 4.17 | 4.17 | 4.17 | 4.17 | 4.19 | 4.19 | 4.19 | 4.19 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S51: Entropy, Diversity, and Overlap results for dataset C NPJ2020

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 5.69 | 6.68 | 6.85 | 8.27 | 7.62 | 8.15 | 8.22 | 8.71 | 0.94 | 0.96 | 0.96 | 0.99 |
| 25.0 | 5.74 | 6.61 | 6.66 | 7.85 | 8.00 | 8.45 | 8.45 | 8.66 | 0.97 | 0.99 | 0.99 | 1.00 |
| 50.0 | 5.81 | 6.32 | 6.35 | 6.95 | 8.28 | 8.51 | 8.52 | 8.55 | 0.99 | 1.00 | 1.00 | 1.00 |
| 75.0 | 5.86 | 6.04 | 6.03 | 6.26 | 8.46 | 8.53 | 8.53 | 8.53 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.83 | 5.83 | 5.83 | 5.83 | 8.53 | 8.53 | 8.53 | 8.53 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S52: Entropy, Diversity, and Overlap results for dataset DP-GEN Cu

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.75 | 8.32 | 9.37 | 9.95 | 9.05 | 9.48 | 9.66 | 9.99 | 0.73 | 0.76 | 0.75 | 0.51 |
| 25.0 | 7.03 | 9.31 | 10.04 | 10.55 | 9.79 | 10.34 | 10.44 | 10.67 | 0.81 | 0.88 | 0.89 | 0.83 |
| 50.0 | 7.14 | 9.61 | 9.84 | 9.90 | 10.31 | 10.78 | 10.80 | 10.80 | 0.89 | 0.99 | 1.00 | 1.00 |
| 75.0 | 7.24 | 8.29 | 8.30 | 8.49 | 10.60 | 10.80 | 10.80 | 10.80 | 0.95 | 1.00 | 1.00 | 1.00 |
| 100.0 | 7.30 | 7.30 | 7.30 | 7.30 | 10.80 | 10.80 | 10.80 | 10.80 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S53: Entropy, Diversity, and Overlap results for dataset Fe nanoparticles PRB 2023

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.64 | 5.82 | 5.43 | 5.99 | 6.90 | 6.10 | 5.81 | 6.21 | 0.64 | 0.38 | 0.25 | 0.27 |
| 25.0 | 6.89 | 6.33 | 6.11 | 7.17 | 7.36 | 6.82 | 6.62 | 7.51 | 0.80 | 0.61 | 0.55 | 0.82 |
| 50.0 | 7.08 | 6.72 | 6.68 | 7.41 | 7.70 | 7.41 | 7.39 | 7.93 | 0.91 | 0.82 | 0.81 | 0.96 |
| 75.0 | 7.23 | 7.02 | 7.05 | 7.38 | 7.95 | 7.79 | 7.82 | 8.06 | 0.97 | 0.93 | 0.94 | 1.00 |
| 100.0 | 7.26 | 7.26 | 7.26 | 7.26 | 8.07 | 8.07 | 8.07 | 8.07 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S54: Entropy, Diversity, and Overlap results for dataset FitSNAP Fe NPJ 2021

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.46 | 6.92 | 5.95 | 6.58 | 7.84 | 7.06 | 6.00 | 7.48 | 0.77 | 0.19 | 0.01 | 0.11 |
| 25.0 | 7.72 | 7.33 | 6.96 | 8.03 | 8.17 | 7.51 | 7.07 | 8.42 | 0.92 | 0.38 | 0.09 | 0.77 |
| 50.0 | 7.77 | 7.92 | 7.73 | 8.20 | 8.35 | 8.19 | 8.02 | 8.55 | 0.97 | 0.80 | 0.60 | 0.99 |
| 75.0 | 7.85 | 8.14 | 8.14 | 7.96 | 8.49 | 8.56 | 8.59 | 8.55 | 0.99 | 0.99 | 0.98 | 1.00 |
| 100.0 | 7.88 | 7.88 | 7.88 | 7.88 | 8.56 | 8.56 | 8.56 | 8.56 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S55: Entropy, Diversity, and Overlap results for dataset Mg edmonds 2022

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.20 | 7.58 | 7.43 | 8.10 | 7.87 | 7.88 | 7.66 | 8.41 | 0.41 | 0.42 | 0.19 | 0.20 |
| 25.0 | 7.77 | 8.44 | 8.42 | 9.10 | 8.62 | 8.75 | 8.67 | 9.35 | 0.59 | 0.65 | 0.50 | 0.48 |
| 50.0 | 8.18 | 9.02 | 9.10 | 9.42 | 9.18 | 9.36 | 9.39 | 9.66 | 0.78 | 0.89 | 0.89 | 0.97 |
| 75.0 | 8.41 | 9.19 | 9.21 | 9.33 | 9.48 | 9.68 | 9.69 | 9.70 | 0.90 | 0.99 | 1.00 | 1.00 |
| 100.0 | 8.55 | 8.55 | 8.55 | 8.55 | 9.70 | 9.70 | 9.70 | 9.70 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S56: Entropy, Diversity, and Overlap results for dataset Mo PRM2019

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.46 | 6.57 | 7.36 | 6.59 | 6.85 | 7.48 | 7.59 | 8.74 | 0.79 | 0.81 | 0.77 | 0.97 |
| 25.0 | 4.91 | 6.67 | 7.30 | 6.04 | 7.73 | 7.92 | 8.01 | 8.77 | 0.86 | 0.87 | 0.86 | 1.00 |
| 50.0 | 5.03 | 6.76 | 6.95 | 5.28 | 8.25 | 8.44 | 8.43 | 8.77 | 0.92 | 0.93 | 0.93 | 1.00 |
| 75.0 | 5.16 | 6.44 | 6.52 | 5.20 | 8.57 | 8.71 | 8.72 | 8.77 | 0.96 | 0.98 | 0.98 | 1.00 |
| 100.0 | 5.14 | 5.14 | 5.14 | 5.14 | 8.77 | 8.77 | 8.77 | 8.77 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S57: Entropy, Diversity, and Overlap results for dataset NDSC TUT 2022

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 0.16 | 0.20 | 0.29 | 0.51 | 0.16 | 0.20 | 0.30 | 0.51 | 1.00 | 1.00 | 1.00 | 1.00 |
| 25.0 | 0.16 | 0.23 | 0.26 | 0.41 | 0.17 | 0.23 | 0.27 | 0.41 | 1.00 | 1.00 | 1.00 | 1.00 |
| 50.0 | 0.18 | 0.22 | 0.22 | 0.31 | 0.19 | 0.22 | 0.22 | 0.31 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 0.20 | 0.20 | 0.20 | 0.24 | 0.21 | 0.21 | 0.21 | 0.25 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 | 0.20 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S58: Entropy, Diversity, and Overlap results for dataset NEP qHPF

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.91 | 6.96 | 6.99 | 7.91 | 7.38 | 7.46 | 7.52 | 7.96 | 0.53 | 0.49 | 0.52 | 0.31 |
| 25.0 | 7.34 | 7.71 | 7.80 | 8.51 | 8.08 | 8.29 | 8.36 | 8.67 | 0.69 | 0.69 | 0.69 | 0.59 |
| 50.0 | 7.66 | 8.08 | 8.18 | 8.74 | 8.60 | 8.81 | 8.86 | 9.09 | 0.85 | 0.87 | 0.88 | 0.86 |
| 75.0 | 7.79 | 8.21 | 8.22 | 8.48 | 8.87 | 9.05 | 9.06 | 9.15 | 0.93 | 0.97 | 0.97 | 0.99 |
| 100.0 | 7.91 | 7.91 | 7.91 | 7.91 | 9.08 | 9.08 | 9.08 | 9.08 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S59: Entropy, Diversity, and Overlap results for dataset Nb PRM2019

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 5.14 | 6.55 | 7.18 | 6.24 | 7.22 | 7.41 | 7.46 | 8.52 | 0.83 | 0.82 | 0.80 | 0.98 |
| 25.0 | 5.19 | 6.22 | 7.02 | 5.73 | 7.75 | 7.70 | 7.85 | 8.55 | 0.86 | 0.89 | 0.89 | 1.00 |
| 50.0 | 4.87 | 6.46 | 6.70 | 4.92 | 8.11 | 8.24 | 8.25 | 8.54 | 0.93 | 0.93 | 0.95 | 1.00 |
| 75.0 | 4.91 | 6.13 | 6.19 | 4.89 | 8.40 | 8.51 | 8.51 | 8.54 | 0.96 | 0.99 | 0.99 | 1.00 |
| 100.0 | 4.83 | 4.83 | 4.83 | 4.83 | 8.54 | 8.54 | 8.54 | 8.54 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S60: Entropy, Diversity, and Overlap results for dataset Si JCP 2017

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.16 | 7.36 | 7.66 | 8.48 | 8.10 | 8.12 | 8.20 | 8.54 | 0.48 | 0.49 | 0.49 | 0.38 |
| 25.0 | 7.74 | 8.13 | 8.23 | 9.08 | 8.82 | 8.85 | 8.86 | 9.19 | 0.71 | 0.70 | 0.70 | 0.73 |
| 50.0 | 8.11 | 8.57 | 8.67 | 9.35 | 9.26 | 9.32 | 9.34 | 9.57 | 0.88 | 0.89 | 0.89 | 0.95 |
| 75.0 | 8.28 | 8.81 | 8.86 | 9.22 | 9.50 | 9.57 | 9.58 | 9.66 | 0.95 | 0.98 | 0.98 | 1.00 |
| 100.0 | 8.35 | 8.35 | 8.35 | 8.35 | 9.65 | 9.65 | 9.65 | 9.65 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S61: Entropy, Diversity, and Overlap results for dataset Si PRX GAP

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.89 | 5.65 | 5.80 | 6.72 | 7.11 | 7.13 | 6.95 | 8.62 | 0.93 | 0.91 | 0.86 | 0.95 |
| 25.0 | 5.06 | 5.94 | 6.43 | 5.63 | 7.81 | 7.79 | 7.84 | 8.76 | 0.96 | 0.94 | 0.93 | 0.99 |
| 50.0 | 5.18 | 6.06 | 6.12 | 5.15 | 8.33 | 8.63 | 8.66 | 8.78 | 0.98 | 0.99 | 0.99 | 1.00 |
| 75.0 | 5.22 | 5.57 | 5.55 | 5.27 | 8.61 | 8.76 | 8.76 | 8.77 | 0.99 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.20 | 5.20 | 5.20 | 5.20 | 8.77 | 8.77 | 8.77 | 8.77 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S62: Entropy, Diversity, and Overlap results for dataset Sn-SCAN PRM 2023

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.06 | 7.11 | 7.30 | 7.77 | 7.63 | 7.67 | 7.71 | 8.14 | 0.92 | 0.92 | 0.89 | 0.87 |
| 25.0 | 7.16 | 7.46 | 7.58 | 7.83 | 7.92 | 8.04 | 8.07 | 8.33 | 0.97 | 0.98 | 0.98 | 0.96 |
| 50.0 | 7.24 | 7.60 | 7.64 | 7.76 | 8.12 | 8.23 | 8.24 | 8.33 | 0.99 | 1.00 | 1.00 | 1.00 |
| 75.0 | 7.26 | 7.58 | 7.58 | 7.60 | 8.22 | 8.29 | 8.28 | 8.29 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 7.26 | 7.26 | 7.26 | 7.26 | 8.28 | 8.28 | 8.28 | 8.28 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S63: Entropy, Diversity, and Overlap results for dataset Ta Linear JCP2014

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 3.60 | 4.71 | 4.70 | 5.39 | 4.32 | 5.12 | 4.93 | 5.82 | 0.79 | 0.74 | 0.29 | 0.60 |
| 25.0 | 4.02 | 5.08 | 5.12 | 5.10 | 5.25 | 5.78 | 5.79 | 6.02 | 0.88 | 0.95 | 0.94 | 1.00 |
| 50.0 | 4.29 | 4.55 | 4.52 | 4.20 | 5.91 | 6.03 | 6.03 | 6.02 | 0.96 | 1.00 | 1.00 | 1.00 |
| 75.0 | 4.09 | 4.25 | 4.25 | 4.06 | 5.98 | 6.03 | 6.03 | 6.02 | 0.99 | 1.00 | 1.00 | 1.00 |
| 100.0 | 3.95 | 3.95 | 3.95 | 3.95 | 6.02 | 6.02 | 6.02 | 6.02 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S64: Entropy, Diversity, and Overlap results for dataset Ta PINN 2021

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 5.94 | 6.63 | 6.92 | 8.26 | 8.00 | 7.63 | 7.45 | 8.81 | 0.87 | 0.78 | 0.71 | 0.94 |
| 25.0 | 6.05 | 7.30 | 7.67 | 8.28 | 8.43 | 8.39 | 8.41 | 8.94 | 0.94 | 0.93 | 0.92 | 0.99 |
| 50.0 | 5.83 | 7.26 | 7.72 | 7.77 | 8.67 | 8.85 | 8.88 | 8.93 | 0.98 | 0.99 | 0.99 | 1.00 |
| 75.0 | 5.73 | 7.15 | 7.21 | 7.18 | 8.83 | 8.94 | 8.94 | 8.93 | 0.99 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.78 | 5.78 | 5.78 | 5.78 | 8.92 | 8.92 | 8.92 | 8.92 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S65: Entropy, Diversity, and Overlap results for dataset Ta PRM2019

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.15 | 6.38 | 7.23 | 6.22 | 1.88 | 1.98 | 2.01 | 2.14 | 0.82 | 0.83 | 0.81 | 0.98 |
| 25.0 | 4.52 | 6.42 | 6.97 | 5.71 | 2.00 | 2.05 | 2.06 | 2.14 | 0.88 | 0.88 | 0.89 | 1.00 |
| 50.0 | 4.64 | 6.39 | 6.66 | 4.89 | 2.08 | 2.10 | 2.11 | 2.14 | 0.93 | 0.95 | 0.95 | 1.00 |
| 75.0 | 4.76 | 6.09 | 6.18 | 4.87 | 2.12 | 2.14 | 2.14 | 2.14 | 0.96 | 0.99 | 0.99 | 1.00 |
| 100.0 | 4.81 | 4.81 | 4.81 | 4.81 | 2.14 | 2.14 | 2.14 | 2.14 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S66: Entropy, Diversity, and Overlap results for dataset Ti NPJCM 2021

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 7.85 | 8.11 | 8.34 | 9.05 | 8.69 | 8.73 | 8.64 | 9.16 | 0.66 | 0.60 | 0.48 | 0.46 |
| 25.0 | 8.11 | 8.73 | 9.04 | 9.48 | 9.29 | 9.43 | 9.43 | 9.77 | 0.81 | 0.79 | 0.78 | 0.79 |
| 50.0 | 8.24 | 9.03 | 9.25 | 9.56 | 9.70 | 9.86 | 9.87 | 10.06 | 0.91 | 0.93 | 0.94 | 0.93 |
| 75.0 | 8.30 | 9.07 | 9.17 | 9.31 | 9.93 | 10.05 | 10.06 | 10.09 | 0.97 | 0.99 | 0.99 | 1.00 |
| 100.0 | 8.34 | 8.34 | 8.34 | 8.34 | 10.09 | 10.09 | 10.09 | 10.09 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S67: Entropy, Diversity, and Overlap results for dataset V PRM2019

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.97 | 6.72 | 7.51 | 7.00 | 6.98 | 7.51 | 7.66 | 8.92 | 0.77 | 0.77 | 0.69 | 0.95 |
| 25.0 | 5.13 | 6.86 | 7.29 | 6.49 | 7.67 | 7.93 | 8.03 | 8.98 | 0.83 | 0.83 | 0.83 | 1.00 |
| 50.0 | 5.61 | 7.08 | 7.25 | 5.83 | 8.49 | 8.51 | 8.55 | 8.97 | 0.91 | 0.91 | 0.91 | 1.00 |
| 75.0 | 5.43 | 6.93 | 6.97 | 5.64 | 8.67 | 8.91 | 8.91 | 8.97 | 0.94 | 0.98 | 0.98 | 1.00 |
| 100.0 | 5.58 | 5.58 | 5.58 | 5.58 | 8.97 | 8.97 | 8.97 | 8.97 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S68: Entropy, Diversity, and Overlap results for dataset W-14

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.59 | 5.35 | 6.06 | 6.15 | 6.80 | 6.87 | 7.08 | 7.68 | 0.97 | 0.97 | 0.97 | 0.99 |
| 25.0 | 4.64 | 5.60 | 6.04 | 5.88 | 7.18 | 7.34 | 7.42 | 7.73 | 0.99 | 0.99 | 0.99 | 1.00 |
| 50.0 | 4.66 | 5.71 | 5.86 | 5.65 | 7.40 | 7.57 | 7.57 | 7.64 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 4.69 | 5.51 | 5.57 | 5.08 | 7.52 | 7.61 | 7.61 | 7.61 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 4.69 | 4.69 | 4.69 | 4.69 | 7.60 | 7.60 | 7.60 | 7.60 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S69: Entropy, Diversity, and Overlap results for dataset W LML-retrain bulk MD test

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 2.97 | 2.89 | 1.25 | 4.71 | 3.45 | 3.35 | 2.07 | 4.83 | 0.84 | 0.85 | 0.72 | 0.62 |
| 25.0 | 4.21 | 4.22 | 3.56 | 4.22 | 4.78 | 4.82 | 4.74 | 4.82 | 0.95 | 0.95 | 0.94 | 0.95 |
| 50.0 | 4.33 | 3.45 | 3.06 | 4.36 | 5.13 | 4.74 | 4.71 | 5.15 | 0.99 | 0.95 | 0.96 | 1.00 |
| 75.0 | 3.78 | 3.14 | 3.59 | 3.86 | 5.07 | 4.74 | 5.10 | 5.11 | 0.99 | 0.96 | 1.00 | 1.00 |
| 100.0 | 3.52 | 3.52 | 3.52 | 3.52 | 5.08 | 5.08 | 5.08 | 5.08 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S70: Entropy, Diversity, and Overlap results for dataset W PRB2019

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 5.13 | 6.71 | 7.35 | 6.35 | 7.19 | 7.52 | 7.63 | 8.70 | 0.76 | 0.81 | 0.81 | 0.98 |
| 25.0 | 5.27 | 6.59 | 7.09 | 5.91 | 7.79 | 7.81 | 7.96 | 8.72 | 0.86 | 0.87 | 0.87 | 1.00 |
| 50.0 | 5.03 | 6.90 | 7.09 | 5.14 | 8.14 | 8.38 | 8.39 | 8.71 | 0.88 | 0.93 | 0.93 | 1.00 |
| 75.0 | 4.82 | 6.59 | 6.64 | 5.12 | 8.33 | 8.68 | 8.68 | 8.71 | 0.95 | 0.99 | 0.98 | 1.00 |
| 100.0 | 5.06 | 5.06 | 5.06 | 5.06 | 8.71 | 8.71 | 8.71 | 8.71 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S71: Entropy, Diversity, and Overlap results for dataset Zn MTP CMS2023

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 5.04 | 6.33 | 7.41 | 8.23 | 8.06 | 8.36 | 8.66 | 9.47 | 0.92 | 0.91 | 0.86 | 0.99 |
| 25.0 | 5.19 | 7.20 | 7.45 | 7.03 | 8.63 | 9.47 | 9.47 | 9.46 | 0.95 | 1.00 | 0.99 | 1.00 |
| 50.0 | 5.31 | 6.23 | 6.27 | 6.25 | 9.02 | 9.48 | 9.48 | 9.46 | 0.97 | 1.00 | 1.00 | 1.00 |
| 75.0 | 5.35 | 5.92 | 5.93 | 5.65 | 9.27 | 9.46 | 9.46 | 9.46 | 0.99 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.38 | 5.38 | 5.38 | 5.38 | 9.46 | 9.46 | 9.46 | 9.46 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S72: Entropy, Diversity, and Overlap results for dataset aC JCP 2023

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 12.79 | 12.79 | 12.80 | 12.80 | 12.79 | 12.79 | 12.80 | 12.80 | 0.08 | 0.08 | 0.08 | 0.07 |
| 25.0 | 13.69 | 13.70 | 13.70 | 13.71 | 13.69 | 13.70 | 13.70 | 13.71 | 0.22 | 0.22 | 0.22 | 0.21 |
| 50.0 | 14.35 | 14.37 | 14.38 | 14.39 | 14.36 | 14.38 | 14.38 | 14.39 | 0.46 | 0.46 | 0.46 | 0.45 |
| 75.0 | 14.73 | 14.76 | 14.76 | 14.78 | 14.74 | 14.77 | 14.77 | 14.78 | 0.71 | 0.70 | 0.70 | 0.70 |
| 100.0 | 14.99 | 14.99 | 14.99 | 14.99 | 15.01 | 15.01 | 15.01 | 15.01 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S73: Entropy, Diversity, and Overlap results for dataset defected phosphorene ACS 2023

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 3.43 | 3.47 | 3.56 | 3.75 | 6.16 | 6.19 | 6.33 | 6.88 | 1.00 | 1.00 | 1.00 | 1.00 |
| 25.0 | 3.45 | 3.51 | 3.56 | 3.72 | 6.50 | 6.61 | 6.66 | 7.10 | 1.00 | 1.00 | 1.00 | 1.00 |
| 50.0 | 3.46 | 3.53 | 3.54 | 3.65 | 6.73 | 6.84 | 6.83 | 7.05 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 3.46 | 3.52 | 3.52 | 3.57 | 6.84 | 6.90 | 6.90 | 6.97 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 3.46 | 3.46 | 3.46 | 3.46 | 6.91 | 6.91 | 6.91 | 6.91 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S74: Entropy, Diversity, and Overlap results for dataset discrepencies and error metrics NPJ 2023 enhanced validation set

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.31 | 5.43 | 4.93 | 5.67 | 4.90 | 5.54 | 5.39 | 5.69 | 0.67 | 0.53 | 0.59 | 0.41 |
| 25.0 | 4.55 | 5.61 | 5.80 | 6.45 | 5.51 | 6.18 | 6.32 | 6.55 | 0.77 | 0.78 | 0.80 | 0.83 |
| 50.0 | 5.13 | 5.78 | 5.77 | 5.93 | 6.28 | 6.66 | 6.65 | 6.69 | 0.84 | 0.93 | 0.93 | 0.95 |
| 75.0 | 5.17 | 5.38 | 5.38 | 5.54 | 6.57 | 6.71 | 6.72 | 6.75 | 0.92 | 0.97 | 0.97 | 0.98 |
| 100.0 | 5.22 | 5.22 | 5.22 | 5.22 | 6.77 | 6.77 | 6.77 | 6.77 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S75: Entropy, Diversity, and Overlap results for dataset discrepencies and error metrics NPJ 2023 interstitial enhanced training set

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.45 | 4.19 | 4.35 | 6.44 | 5.81 | 5.87 | 6.01 | 6.81 | 0.85 | 0.82 | 0.78 | 0.81 |
| 25.0 | 4.53 | 4.81 | 5.05 | 5.67 | 6.34 | 6.72 | 6.91 | 7.02 | 0.90 | 0.91 | 0.91 | 0.95 |
| 50.0 | 4.53 | 5.03 | 5.09 | 5.25 | 6.70 | 7.10 | 7.11 | 7.15 | 0.95 | 0.98 | 0.98 | 0.97 |
| 75.0 | 4.54 | 4.85 | 4.86 | 4.91 | 6.94 | 7.18 | 7.18 | 7.20 | 0.97 | 0.99 | 0.99 | 1.00 |
| 100.0 | 4.62 | 4.62 | 4.62 | 4.62 | 7.20 | 7.20 | 7.20 | 7.20 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S76: Entropy, Diversity, and Overlap results for dataset discrepencies and error metrics NPJ 2023 vacancy enhanced training set

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 3.76 | 4.13 | 4.29 | 6.54 | 5.30 | 5.86 | 5.98 | 6.82 | 0.81 | 0.84 | 0.80 | 0.81 |
| 25.0 | 4.20 | 4.81 | 5.13 | 5.65 | 5.91 | 6.69 | 6.90 | 6.99 | 0.92 | 0.93 | 0.93 | 0.93 |
| 50.0 | 4.53 | 5.05 | 5.06 | 5.24 | 6.59 | 7.05 | 7.06 | 7.12 | 0.96 | 0.98 | 0.98 | 0.98 |
| 75.0 | 4.59 | 4.88 | 4.88 | 4.92 | 6.89 | 7.12 | 7.13 | 7.15 | 0.98 | 1.00 | 1.00 | 1.00 |
| 100.0 | 4.62 | 4.62 | 4.62 | 4.62 | 7.15 | 7.15 | 7.15 | 7.15 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S77: Entropy, Diversity, and Overlap results for dataset mlearn Cu

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.83 | 4.25 | 4.26 | 7.68 | 6.85 | 6.33 | 6.34 | 7.87 | 0.75 | 0.76 | 0.74 | 0.50 |
| 25.0 | 5.02 | 5.07 | 5.93 | 7.94 | 7.48 | 7.48 | 7.93 | 8.42 | 0.86 | 0.85 | 0.87 | 0.94 |
| 50.0 | 5.30 | 6.65 | 6.68 | 7.22 | 8.01 | 8.45 | 8.45 | 8.48 | 0.92 | 0.99 | 0.99 | 1.00 |
| 75.0 | 5.23 | 6.17 | 6.17 | 6.19 | 8.27 | 8.47 | 8.47 | 8.47 | 0.96 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.28 | 5.28 | 5.28 | 5.28 | 8.47 | 8.47 | 8.47 | 8.47 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S78: Entropy, Diversity, and Overlap results for dataset mlearn Ge

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 3.88 | 3.96 | 4.45 | 6.97 | 6.13 | 6.03 | 6.17 | 7.04 | 0.63 | 0.62 | 0.62 | 0.53 |
| 25.0 | 4.83 | 5.84 | 6.10 | 7.59 | 7.04 | 7.27 | 7.33 | 7.69 | 0.72 | 0.79 | 0.81 | 0.72 |
| 50.0 | 4.85 | 6.99 | 7.03 | 7.33 | 7.44 | 7.87 | 7.88 | 7.93 | 0.86 | 0.98 | 0.98 | 0.96 |
| 75.0 | 5.20 | 6.04 | 6.03 | 5.99 | 7.76 | 7.93 | 7.93 | 7.93 | 0.97 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.26 | 5.26 | 5.26 | 5.26 | 7.93 | 7.93 | 7.93 | 7.93 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S79: Entropy, Diversity, and Overlap results for dataset mlearn Li

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 6.12 | 6.32 | 6.14 | 6.92 | 6.36 | 6.61 | 6.48 | 7.00 | 0.51 | 0.41 | 0.39 | 0.21 |
| 25.0 | 6.77 | 7.06 | 7.07 | 7.63 | 7.12 | 7.37 | 7.37 | 7.72 | 0.68 | 0.71 | 0.71 | 0.63 |
| 50.0 | 7.12 | 7.47 | 7.51 | 7.83 | 7.63 | 7.86 | 7.88 | 8.05 | 0.85 | 0.92 | 0.92 | 0.97 |
| 75.0 | 7.33 | 7.50 | 7.52 | 7.57 | 7.93 | 8.04 | 8.05 | 8.07 | 0.95 | 0.99 | 0.99 | 1.00 |
| 100.0 | 7.38 | 7.38 | 7.38 | 7.38 | 8.06 | 8.06 | 8.06 | 8.06 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S80: Entropy, Diversity, and Overlap results for dataset mlearn Mo

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.09 | 4.22 | 4.31 | 6.58 | 5.98 | 5.92 | 5.65 | 6.76 | 0.60 | 0.60 | 0.60 | 0.46 |
| 25.0 | 4.59 | 5.24 | 5.62 | 7.45 | 6.69 | 6.88 | 7.03 | 7.64 | 0.76 | 0.75 | 0.74 | 0.78 |
| 50.0 | 4.76 | 6.65 | 6.74 | 7.33 | 7.21 | 7.74 | 7.75 | 7.83 | 0.85 | 0.97 | 0.97 | 1.00 |
| 75.0 | 5.21 | 6.14 | 6.13 | 6.10 | 7.65 | 7.82 | 7.82 | 7.82 | 0.95 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.30 | 5.30 | 5.30 | 5.30 | 7.82 | 7.82 | 7.82 | 7.82 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S81: Entropy, Diversity, and Overlap results for dataset mlearn Ni

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.29 | 3.80 | 3.87 | 7.76 | 6.46 | 6.10 | 6.02 | 7.91 | 0.83 | 0.81 | 0.78 | 0.56 |
| 25.0 | 4.60 | 4.89 | 5.70 | 7.67 | 7.39 | 7.48 | 7.86 | 8.37 | 0.87 | 0.88 | 0.90 | 0.96 |
| 50.0 | 4.82 | 6.20 | 6.23 | 6.65 | 7.91 | 8.38 | 8.38 | 8.40 | 0.92 | 0.99 | 1.00 | 1.00 |
| 75.0 | 4.94 | 5.67 | 5.66 | 5.69 | 8.20 | 8.39 | 8.39 | 8.39 | 0.96 | 1.00 | 1.00 | 1.00 |
| 100.0 | 4.89 | 4.89 | 4.89 | 4.89 | 8.39 | 8.39 | 8.39 | 8.39 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S82: Entropy, Diversity, and Overlap results for dataset mlearn Si

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 4.37 | 4.82 | 5.15 | 7.09 | 6.30 | 6.41 | 6.43 | 7.14 | 0.67 | 0.68 | 0.68 | 0.54 |
| 25.0 | 5.50 | 5.96 | 6.25 | 7.92 | 7.33 | 7.43 | 7.53 | 8.00 | 0.77 | 0.78 | 0.79 | 0.77 |
| 50.0 | 5.11 | 7.06 | 7.22 | 7.36 | 7.65 | 8.22 | 8.27 | 8.28 | 0.83 | 0.97 | 0.99 | 0.97 |
| 75.0 | 5.34 | 6.21 | 6.19 | 6.17 | 8.05 | 8.27 | 8.27 | 8.27 | 0.92 | 1.00 | 1.00 | 1.00 |
| 100.0 | 5.41 | 5.41 | 5.41 | 5.41 | 8.27 | 8.27 | 8.27 | 8.27 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table** S83: Entropy, Diversity, and Overlap results for dataset pure magnesium DFT PRM2020

| | Entropy (nats) | | | | Diversity (nats) | | | | Overlap (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size (%) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) | Random | K-means | Mean FPS | MSC (ours) |
| 10.0 | 0.90 | 2.15 | 2.12 | 2.05 | 3.20 | 3.80 | 3.69 | 4.08 | 0.98 | 0.98 | 0.98 | 0.99 |
| 25.0 | 1.04 | 2.00 | 1.99 | 2.03 | 3.52 | 4.17 | 4.21 | 4.27 | 0.99 | 1.00 | 1.00 | 1.00 |
| 50.0 | 1.25 | 1.63 | 1.69 | 1.82 | 4.00 | 4.25 | 4.24 | 4.24 | 1.00 | 1.00 | 1.00 | 1.00 |
| 75.0 | 1.32 | 1.56 | 1.47 | 1.57 | 4.13 | 4.24 | 4.23 | 4.22 | 1.00 | 1.00 | 1.00 | 1.00 |
| 100.0 | 1.40 | 1.40 | 1.40 | 1.40 | 4.22 | 4.22 | 4.22 | 4.22 | 1.00 | 1.00 | 1.00 | 1.00 |

# References


[1] Andolina, C. M. & Saidi, W. A. 23-Single-Element-DNPs RSCDD 2023-Ag (2023). URL https://materials.colabfit.org/id/DS_q4h7q8q0fnve_0.

[2] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Al (2023). URL https://materials.colabfit.org/id/DS_8y775we7um7w_0.

[3] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Au (2023). URL https://materials.colabfit.org/id/DS_iie3c31ar46x_0.

[4] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Co (2023). URL https://materials.colabfit.org/id/DS_jt0lax9yd15r_0.

[5] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Cu (2023). URL https://materials.colabfit.org/id/DS_dc3o40aou2le_0.

[6] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Ge (2023). URL https://materials.colabfit.org/id/DS_90fnsmavv1am_0.

[7] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 I (2023). URL https://materials.colabfit.org/id/DS_gc1y80tpyylb_0.

[8] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Kr (2023). URL https://materials.colabfit.org/id/DS_omnl1yy49sdh_0.

[9] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Li (2023). URL https://materials.colabfit.org/id/DS_wyo91w20wlgm_0.

[10] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Mg (2023). URL https://materials.colabfit.org/id/DS_mevqyitwxukc_0.

[11] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Mo (2023). URL https://materials.colabfit.org/id/DS_bjkftn0ug9r3_0.

[12] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Nb (2023). URL https://materials.colabfit.org/id/DS_zbxayq0diq6l_0.

[13] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Ni (2023). URL https://materials.colabfit.org/id/DS_nue14kckbkdh_0.

[14] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Os (2023). URL https://materials.colabfit.org/id/DS_098x6q7kbeat_0.

[15] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Pb (2023). URL https://materials.colabfit.org/id/DS_k065jfggbq43_0.

[16] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Pd (2023). URL https://materials.colabfit.org/id/DS_g0sb0h7usqw7_0.

[17] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Pt (2023). URL https://materials.colabfit.org/id/DS_0zgz34a90a6i_0.

[18] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Re (2023). URL https://materials.colabfit.org/id/DS_gzbvdicu231p_0.

[19] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Sb (2023). URL https://materials.colabfit.org/id/DS_z90lfjg88fzo_0.

[20] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Sr (2023). URL https://materials.colabfit.org/id/DS_o3itca7mk80r_0.

[21] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Ti (2023). URL https://materials.colabfit.org/id/DS_ofpcyxez6xsc_0.

[22] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Zn (2023). URL https://materials.colabfit.org/id/DS_xsad0btc0bsn_0.

[23] Andolina, C. M. & Saidi, W. A. 23 Single Element DNPs RSCDD 2023 Zr (2023). URL https://materials.colabfit.org/id/DS_h9w8v7289utq_0.

[24] Yinan Wang, Xiaoyang Wang, Linfeng Zhang, Xu, B. & Wang, H. Ag PBE MSMSE 2021 (2023). URL https://materials.colabfit.org/id/DS_6e5ljaqa3cfr_0.

[25] Yinan Wang, Xiaoyang Wang, Linfeng Zhang, Xu, B. & Wang, H. Au PBE MSMSE 2021 (2023). URL https://materials.colabfit.org/id/DS_tydu7u0h0hhz_0.

[26] Hedman, D. *et al.* CA 9 training (2023). URL https://materials.colabfit.org/id/DS_9cdtww9pcjqd_0.

[27] Hedman, D. *et al.* CA 9 BB training (2023). URL https://materials.colabfit.org/id/DS_l7inbtql4ea9_0.

[28] Hedman, D. *et al.* CA 9 BR training (2023). URL https://materials.colabfit.org/id/DS_fw7m5d8b0fa9_0.

[29] Hedman, D. *et al.* CA 9 RR training (2023). URL https://materials.colabfit.org/id/DS_tag5zubl21w8_0.

[30] Zhang, D., Peiyun Yi, Xinmin Lai, Linfa Peng & Li, H. CGM MLP natcomm2023 screening amorphous carbon train (2024). URL https://materials.colabfit.org/id/DS_taxwl7jo4f8a_0.

[31] Zhang, D., Peiyun Yi, Xinmin Lai, Linfa Peng & Li, H. CGM MLP natcomm2023 screening graphite train (2024). URL https://materials.colabfit.org/id/DS_jasbxoigo7r4_0.

[32] Mingjian Wen & Ellad B. Tadmor. C NPJ2020 (2023). URL https://materials.colabfit.org/id/DS_qph0akhjv9kv_0.

[33] Yuzhi Zhang *et al.* DP GEN Cu (2023). URL https://materials.colabfit.org/id/DS_101uk70asqhq_0.

[34] Jana, R. & Caro, M. A. Fe nanoparticles PRB 2023 (2023). URL https://materials.colabfit.org/id/DS_5h3810yhu4wj_0.

[35] Nikolov, S. *et al.* FitSNAP Fe NPJ 2021 (2023). URL https://materials.colabfit.org/id/DS_ej5u0tuycoph_0.

[36] Poul, M. Mg edmonds 2022 (2023). URL https://materials.colabfit.org/id/DS_caktb6z8yiy7_0.

[37] Byggmästar, J., Nordlund, K. & Flyura Djurabekova. Mo PRM2019 (2023). URL https://materials.colabfit.org/id/DS_5aeg7va6k305_0.

[38] Allen, C. & Bartok, A. P. NDSC TUT 2022 (2023). URL https://materials.colabfit.org/id/DS_oqqwzogut1on_0.

[39] Penghua Ying. NEP qHPF train (2023). URL https://materials.colabfit.org/id/DS_umliq1qh50gy_0.

[40] Byggmästar, J., Nordlund, K. & Flyura Djurabekova. Nb PRM2019 (2023). URL https://materials.colabfit.org/id/DS_tbpx3pawtlgt_0.

[41] Cubuk, E. D., Malone, B. D., Onat, B., Waterland, A. & Kaxiras, E. Si JCP 2017 (2023). URL https://materials.colabfit.org/id/DS_paehju6qhaym_0.

[42] Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Si PRX GAP (2023). URL https://materials.colabfit.org/id/DS_u9wd92plbetq_0.

[43] Chen, T. *et al.* Sn SCAN PRM 2023 (2023). URL https://materials.colabfit.org/id/DS_h1i3plesx7bc_0.

[44] Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Garritt J. Tucker. Ta Linear JCP2014 (2023). URL https://materials.colabfit.org/id/DS_rgtu2lkgv5rq_0.

[45] Yi-Shen Lin, Pun, G. P. P. & Mishin, Y. Ta PINN 2021 (2023). URL https://materials.colabfit.org/id/DS_r6c6gt2s98xm_0.

[46] Byggmästar, J., Nordlund, K. & Flyura Djurabekova. Ta PRM2019 (2023). URL https://materials.colabfit.org/id/DS_40zw467dnc6d_0.

[47] Tongqi Wen *et al.* Ti NPJCM 2021 (2023). URL https://materials.colabfit.org/id/DS_aaocsvp2x40m_0.

[48] Byggmästar, J., Nordlund, K. & Flyura Djurabekova. V PRM2019 (2023). URL https://materials.colabfit.org/id/DS_ouwlietscprn_0.

[49] Szlachta, W. J., Bartók, A. P. & Csányi, G. W 14 (2023). URL https://materials.colabfit.org/id/DS_kk4a8u1eo42m_0.

[50] Onat, B., Ortner, C. & Kermode, J. R. W LML retrain bulk MD test (2023). URL https://materials.colabfit.org/id/DS_0od3fns1ap8a_0.

[51] Byggmästar, J., Hamedani, A., Nordlund, K. & Flyura Djurabekova. W PRB2019 (2023). URL https://materials.colabfit.org/id/DS_gh9s2sopu064_0.

[52] Haojie Mei *et al.* Zn MTP CMS2023 (2024). URL https://materials.colabfit.org/id/DS_58y020ce6b6j_0.

[53] Minamitani, E., Ippei Obayashi, Shimizu, K. & Watanabe, S. aC JCP 2023 (2023). URL https://materials.colabfit.org/id/DS_bmjfal3bj4ah_0.

[54] Lukáš Kývala, Angeletti, A., Franchini, C. & Dellago, C. defected phosphorene ACS 2023 (2023). URL https://materials.colabfit.org/id/DS_k059wtxqsksu_0.

[55] Yunsheng Liu, Xingfeng He & Yifei Mo. discrepencies and error metrics NPJ 2023 enhanced validation set (2023). URL https://materials.colabfit.org/id/DS_q6e3bvq4y67a_0.

[56] Yunsheng Liu, Xingfeng He & Yifei Mo. discrepencies and error metrics NPJ 2023 interstitial enhanced training set (2023). URL https://materials.colabfit.org/id/DS_nublbp38wse0_0.

[57] Yunsheng Liu, Xingfeng He & Yifei Mo. discrepencies and error metrics NPJ 2023 vacancy enhanced training set (2023). URL https://materials.colabfit.org/id/DS_qxd7wv9yabtp_0.

[58] Yunxing Zuo *et al.* mlearn Cu train (2023). URL https://materials.colabfit.org/id/DS_3pv3hck35iy6_0.

[59] Yunxing Zuo *et al.* mlearn Ge train (2023). URL https://materials.colabfit.org/id/DS_ot3m0rxle8fs_0.

[60] Yunxing Zuo *et al.* mlearn Li train (2023). URL https://materials.colabfit.org/id/DS_h7zao9ya9sd8_0.

[61] Yunxing Zuo *et al.* mlearn Mo train (2023). URL https://materials.colabfit.org/id/DS_ytoet4uyc32k_0.

[62] Yunxing Zuo *et al.* mlearn Ni train (2023). URL https://materials.colabfit.org/id/DS_lfyd4jv627cr_0.

[63] Yunxing Zuo *et al.* mlearn Si train (2023). URL https://materials.colabfit.org/id/DS_eltotrjoqonr_0.

[64] Binglun Yin, Stricker, M. & W. A. Curtin. pure magnesium DFT PRM2020 (2024). URL https://materials.colabfit.org/id/DS_48yn17885mdi_0.