

Supplementary Material: Surprisingly High Redundancy in Electronic Structure Data Across Materials Explained by Low Intrinsic Dimensionality

Sazzad Hossain ¹, Ponkrshnan Thiagarajan ², Shashank Pathrudkar ¹, Stephanie Taylor ³, Abhijeet Sadashiv Gangan ³, Amartya S. Banerjee ^{3,*} and Susanta Ghosh ^{1,4,†}

¹*Department of Mechanical and Aerospace Engineering,
Michigan Technological University, Houghton, MI, USA.*

²*Hopkins Extreme Materials Institute, Johns Hopkins University, Baltimore, MD, USA.*

³*Department of Materials Science and Engineering,
University of California, Los Angeles, CA, USA.*

⁴*Center for Artificial Intelligence, Michigan Technological University, Houghton, MI, USA.*

(Dated: September 22, 2026)

I. ERROR METRICS

The normalized root mean square error (NRMSE) of the ML predicted electron density is defined as

$$NRMSE = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (\rho_{DFT}^i - \rho_{ML}^i)^2}}{\rho_{DFT}^{max} - \rho_{DFT}^{min}} \quad (S1)$$

where N is the number of grid-points for each snapshot. At i -th grid point, ρ_{DFT}^i represents the electron density calculated using KS-DFT while ρ_{ML}^i represents the electron density predicted by the ML models. ρ_{DFT}^{max} and ρ_{DFT}^{min} are the maximum and minimum values of electron density across all grid-points of that snapshot.

Let u be the error field of the ML-predicted electron density, given by $u(x) = \rho_{DFT}(x) - \rho_{ML}(x)$, with $x \in \Omega$. The H^1 seminorm of the scalar error field $u(x)$ over a domain Ω is defined as $|u|_{H^1(\Omega)} = (\int_{\Omega} |\nabla u(x)|^2 dx)^{1/2}$. The H^1 norm is defined as $\|u\|_{H^1(\Omega)} = \left(\|u\|_{L^2(\Omega)}^2 + |u|_{H^1(\Omega)}^2 \right)^{1/2}$. These quantities are well defined whenever, u lies in the Sobolev space $H^1(\Omega)$.

In the paper, the energy prediction errors are defined as the absolute differences (in Hartree per atom) between KS-DFT energies and the energies obtained by postprocessing the ML-predicted electron density.

The relative L_{∞} norm is defined as:

$$L_{\infty} = \frac{\max_{i=1}^N |\rho_i^{ML} - \rho_i^{DFT}|}{\max_i |\rho_i^{DFT}|} \quad (S2)$$

where ρ_i^{ML} and ρ_i^{DFT} are the ML-predicted and KS-DFT electron densities at grid point i , respectively, and N is the number of grid-points for each snapshot.

II. POSTPROCESSING OF ML PREDICTED ELECTRON DENSITY

For application purposes, it is of interest to compute downstream properties from the ML-predicted electron density. Since the Kohn-Sham total ground-state energy is a key property that can be computed from the electron density, it can serve as a useful test for the downstream accuracy of the ML model. We postprocess the density predictions to obtain the corresponding total ground-state energies as follows. The predicted electron density is first rescaled by the total number of electrons; this serves to ensure that the total electronic charge is consistent [1–4]. The scaled electron density ρ^{scaled} at grid point $\mathbf{r}$ is obtained from the predicted electron density ρ^{ML} via:

$$\rho^{\text{scaled}}(\mathbf{r}) = \rho^{\text{ML}}(\mathbf{r}) \frac{N_e}{\int_{\Omega} \rho^{\text{ML}}(\mathbf{r}) d\mathbf{r}}. \quad (S3)$$

Here Ω is the periodic supercell used in the calculations, and N_e is the number of electrons in the system. The rescaled electron density ρ^{scaled} is then used to set up the Kohn-Sham Hamiltonian and a single diagonalization step is performed by using the same electronic structure calculation framework, as was used in the data generation process [5–7]. Thereafter, the energy of the system is computed using the Harris-Foulkes formula [8, 9]. The difference between the energy obtained from the ML-predicted electron density and that obtained from the original DFT density serves as the energy error reported in this work.

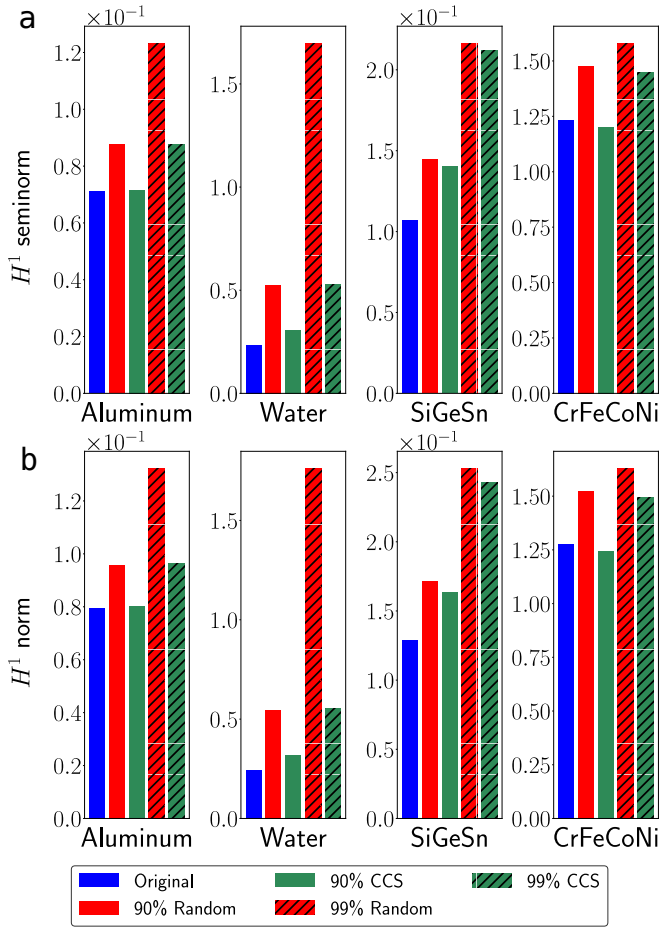
III. SMOOTHNESS OF THE ELECTRON DENSITY PREDICTION

To demonstrate that the ML models not only capture the electron density fields but also smooth spatial variations in them, we plot the H^1 seminorm and H^1 norm of the error in electron density fields in Figure 1. The ML models reproduce both the electron density fields and their smooth spatial variations, with smoothness errors

* asbanerjee@ucla.edu

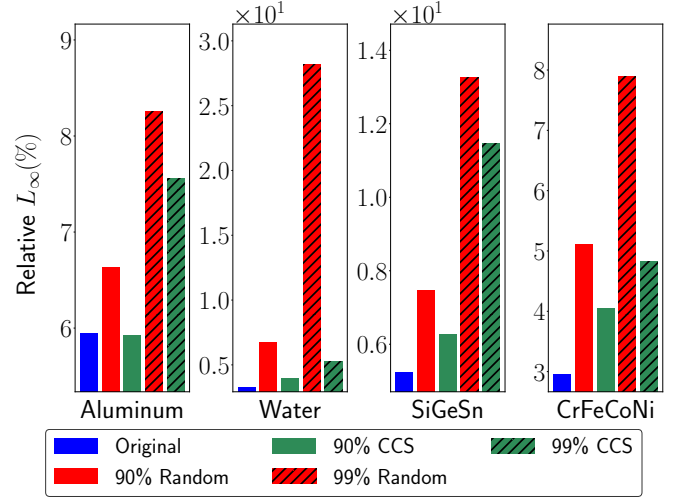
† susantag@mtu.edu

following trends similar to those observed in the electron density itself. For all four material systems, the 90% CCS-pruned dataset achieves a similar degree of spatial smoothness as the original dataset (i.e., similar H^1 seminorm and H^1 norm values), whereas the 99% CCS-pruned dataset shows a slight increase. The CCS method achieves better smoothness than random for 90% and 99% pruning, for all four systems.



Supplementary Figure 1. **Error in smoothness of the predicted electron density:**(a) H^1 seminorm and (b) H^1 norm of the error for the ML-predicted electron density obtained from the original dataset, 90%, and 99% randomly and CCS based pruned dataset.

IV. ROBUSTNESS OF PRUNING



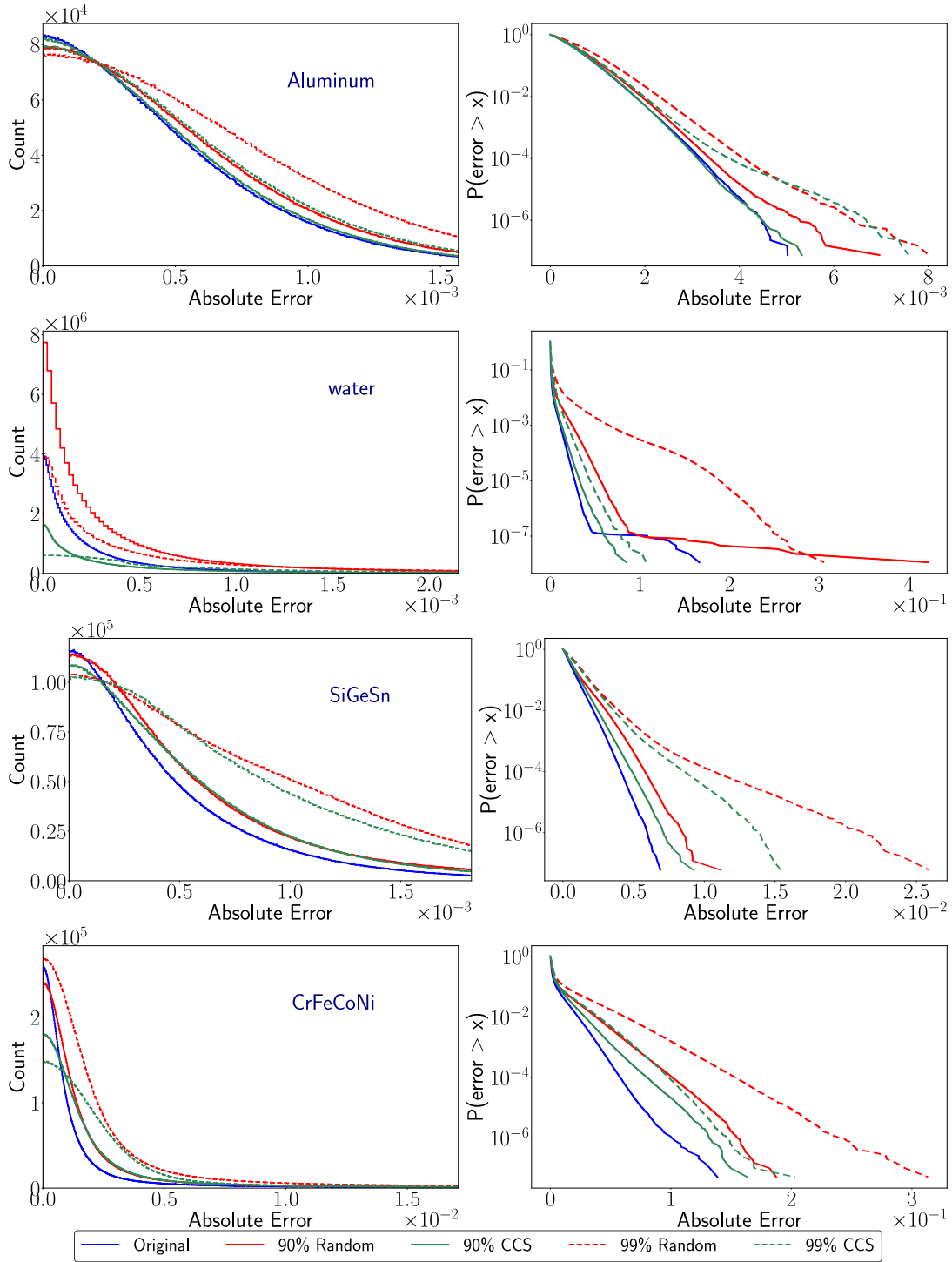
Supplementary Figure 2. Relative L_∞ error in the ML-predicted electron density for the original dataset, 90% and 99% randomly, and CCS-based pruned datasets.

To assess the maximum error upon pruning, we evaluate the L_∞ norm of the error and the absolute error distribution of the ML prediction. Figure 2 shows that after 90% pruning, there is a marginal increase in errors. At 99% pruning, the error increase is larger. However, CCS consistently outperforms random pruning across all systems. Figure 3 shows the absolute error count distribution (left panels) and the probability of the pointwise error exceeding a given threshold (right panels) for all grid points across all test snapshots, for all four systems. For Aluminum and SiGeSn, the 90% CCS pruning's error distributions nearly overlap with the original, indicating negligible degradation at this pruning level. For CrFeCoNi, 90% CCS tracks the original closely. Notably, for water, the tail of the 90% CCS model is comparable to or better than the original. Across all systems, random pruning consistently produces the heaviest tail, while 99% CCS remains between the original and 99% Random, confirming that CCS-based pruning introduces significantly fewer large local errors than random pruning at high pruning factors.

V. EXPERIMENTAL DETAILS

Hardware Details: All experiments were run on NVIDIA A100 (40GB) GPUs available through NERSC Perlmutter. For the CrFeCoNi experiments, we have used 4 GPUs but for other experiments, we have used single GPU.

Dataset details: We use sufficiently large datasets to achieve high accuracy in ML for all four material systems. The KS-DFT dataset for the aluminum system consists



Supplementary Figure 3. Comparison among the error distribution from the ML-predicted density obtained from the original dataset, 90%, and 99% randomly and CCS based dataset. (left) Absolute Error Distribution for four systems. (right) Probability of a pointwise error exceeding a given threshold.

of 200 training snapshots ($216,000$ grid points each), totaling 43.2×10^6 grid points, obtained from a 32-atom simulation cell; the test set includes 60 snapshots from six temperatures (298 – 1800 K). The dataset for water includes 112 training snapshots ($456,533$ grid points

each, totaling 51.1×10^6 grid points), and 193 test snapshots from four temperatures, obtained from a 32-atom simulation cell. The dataset for SiGeSn system, taken from 45 compositions, consists of 216 training snapshots (totaling 38.6×10^6 grid points) and test set includes

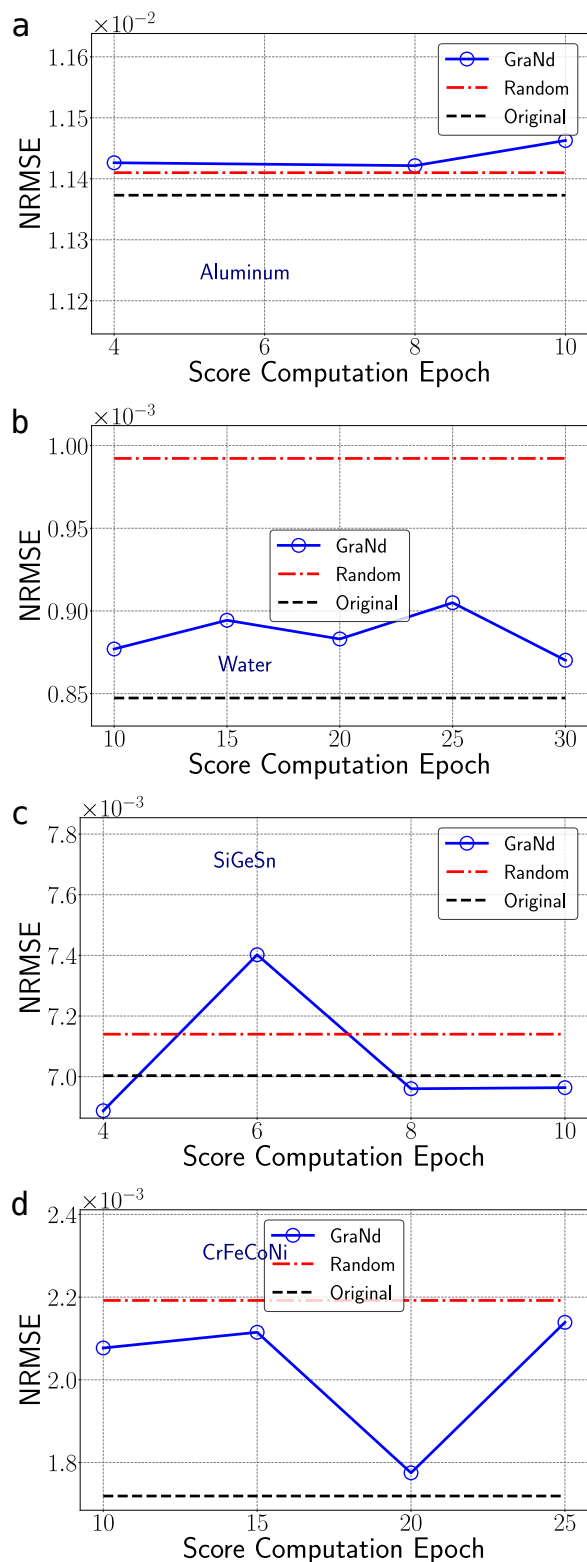
90 test snapshots from 45 compositions, obtained from a 64-atom simulation cell. The dataset for CrFeCoNi system, taken from 69 compositions, consists of 552 training snapshots (totaling 161.4×10^6 grid points) and test set includes 207 test snapshots across four different temperatures and 69 compositions, also from a 32-atom simulation cell.

Descriptor Size: We have chosen descriptor size so that the original model meets the chemical accuracy level of predictions, across all systems. We have kept the all descriptor fixed throughout the pruning experiments.

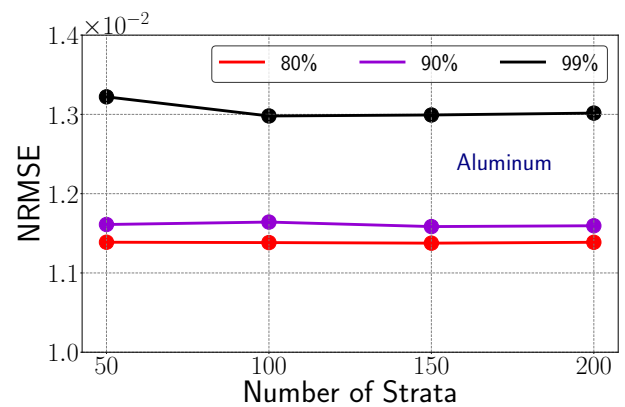
Neural Network: We employ a deep neural network with 12 hidden layers, each containing 900 nodes, using GELU activation [10] in the hidden layers and a ReLU activation in the output layer. A starting learning rate of 0.001 is used across all cases, with the AdamW optimizer [11, 12]. ReduceLROnPlateau type learning rate scheduler was used with a minimum learning rate of 10^{-5} . Training was not terminated at a fixed number of epochs; instead, a convergence criterion was used; training was stopped when the validation loss did not improve over a window of 35 epochs. We also want to note that all models trained on pruned datasets are trained from scratch with independent random initialization; no weights are transferred from the original model.

GraNd score computation: The epoch 10, 30, 10 and 20 were chosen for GraNd score computation for Aluminum, water, SiGeSn and CrFeCoNi respectively. Figure 4 shows the errors on the test dataset for Aluminum, water, SiGeSn, and CrFeCoNi for pruned subsets, for scores computed at different times in training. The Figure 4 shows that computing the scores early in training is sufficient to identify important examples, aligning with the findings of [13]. For GraNd score computation, we use randomly initialized score-computation models that are separate from the models used for final training on the original dataset. The score-computation models are trained only to early checkpoints rather than to full convergence [13]. The computational cost of these score-computation runs is included in the full-pipeline timing comparison reported in Section IX and Supplementary Figure 17.

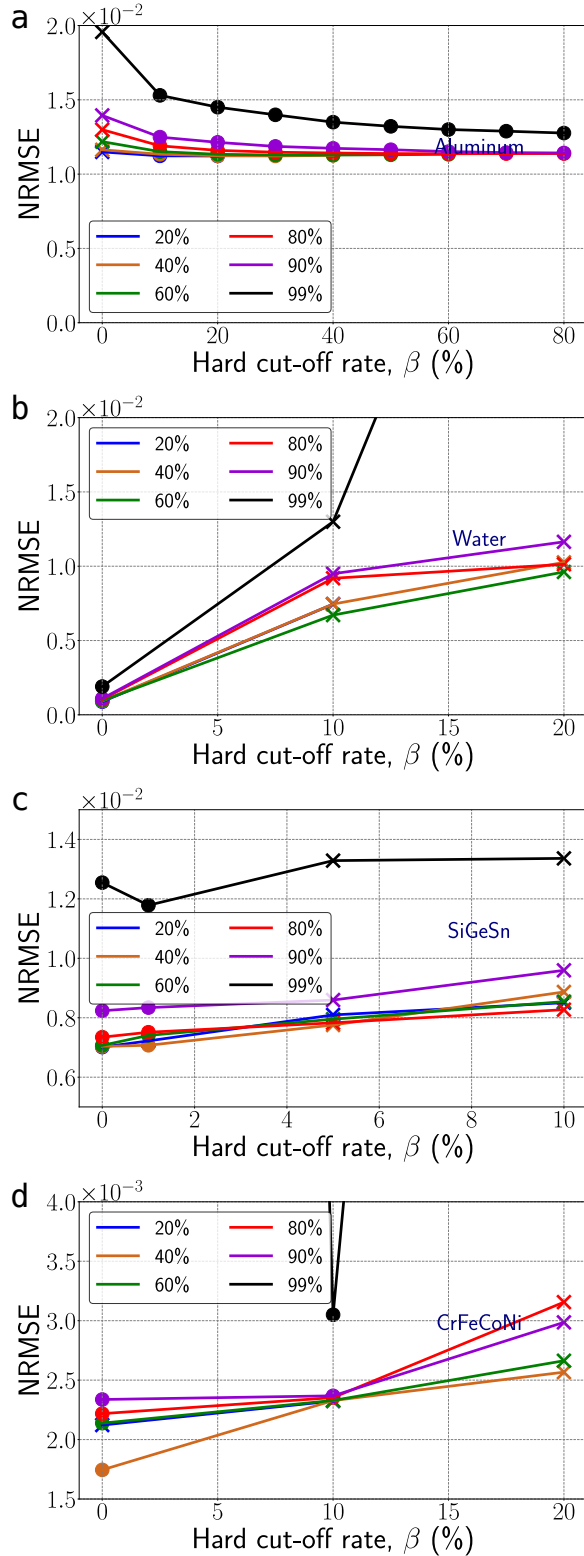
Hyper-parameters for CCS: There are two hyper-parameters for the CCS method. a) the hard cut-off rate (β) and b) the number of strata (k). The optimal β generally increases with the pruning rate. The performance of various CCS-based coresets constructed from varying β is shown in Figure 6. The error in charge density prediction for various numbers of strata, k is shown in Figure 5. The error is not very sensitive to the number of strata (k), which is consistent with the findings of Zheng et al. [14] for image classification datasets. Hence, in this work, the number of strata (k) being used is 100 for all four material systems for all pruning factors.



Supplementary Figure 4. **Choice of score computation epoch:** The error in electron density prediction for (a) Aluminum(20% pruning), (b) Water(40% pruning) (c) SiGeSn(20% pruning) and (d) CrFeCoNi(40% pruning) after computing the GraNd score at different times in training.



Supplementary Figure 5. **Choice of number of strata(k):** The error in electron density prediction with various numbers of strata for Aluminum system for various pruning factors for CCS method.



Supplementary Figure 6. **Choice of hard cut-off rate(β):** The error in electron density prediction with varying β for (a) Aluminum, (b) water (c) SiGeSn (d) CrFeCoNi for the CCS based pruning. It is observed that the obtained NRMSE with pruned dataset varies with the choice of β . For higher pruning rate, a higher β may be optimal. The points with $\times$ markers failed to beat random pruning.

VI. INVESTIGATION INTO TEMPORAL CORRELATION

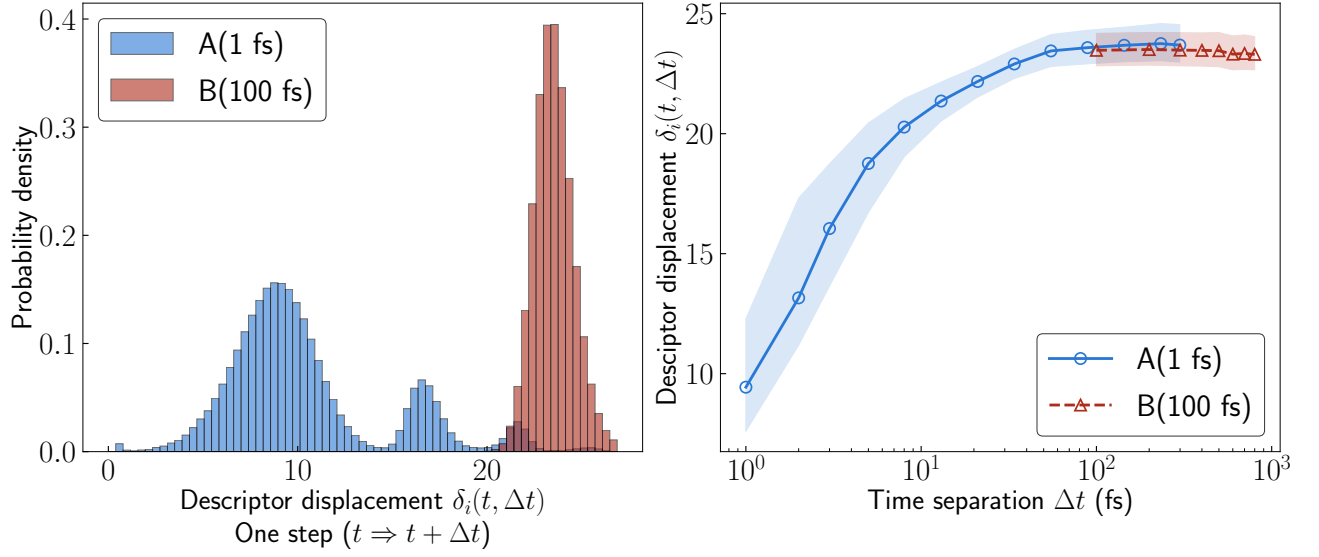
There will be high structural similarity (or strong correlation) between adjacent snapshots taken from an AIMD/MD trajectory. We have generated data so that it is (nearly) free from any structural similarity or temporal correlations present in AIMD/MD trajectory of a single simulation. To ensure that the data is structurally dissimilar (or uncorrelated) we have adopted the following three strategies in data generation:

- **Large time interval between snapshots:** the training snapshots are extracted at intervals of 100 timesteps (1 femtosecond (fs) timestep, i.e., 100 fs spacing) from AIMD/MD trajectories. This interval is sufficiently large compared to the typical vibrational periods in solids and liquids that fall in the range of ~ 10 –100 fs. Such spacing between snapshots is commonly used for obtaining uncorrelated configurations from MD trajectories [15, 16].
- **Simulations at different temperatures:** data at different temperatures are generated through entirely independent simulations, with a diverse set of initial configurations. Snapshots at different temperatures are therefore not points along a single MD trajectory and hence, carry little to no temporal correlation with one another. Higher-temperature simulations might give a better sampling of the configuration space [1] and hence, various high-temperature simulations are included.
- **Simulations for several alloy compositions:** for the alloy systems (SiGeSn and CrFeCoNi), each composition is simulated independently — 69 compositions for CrFeCoNi and 45 compositions for SiGeSn, at various temperatures. Snapshots from different compositions at different temperatures thus originate from entirely distinct initial conditions and follow distinct trajectories, hence, are structurally dissimilar or uncorrelated.

Temporal Correlation Analysis: In the following, we analyze the temporal correlation between snapshots taken from AIMD/MD trajectories. AIMD/MD simulations are used to generate the trajectory of snapshots from a 32-atom cell of $\text{Cr}_{25}\text{Fe}_{25}\text{Co}_{25}\text{Ni}_{25}$ system at 5000 K. We consider two datasets extracted from this simulation:

- **Dataset A:** Snapshots collected at every timestep (1 fs intervals).
- **Dataset B:** Snapshots collected every 100 timesteps (100 fs intervals).

In the following, we quantify the temporal correlation between snapshots directly in the descriptor space, by measuring how much the descriptor at a fixed grid point changes between two snapshots separated by a time lag



Supplementary Figure 7. **Temporal correlation analysis for a representative system (equiatomic $\text{Cr}_{25}\text{Fe}_{25}\text{Co}_{25}\text{Ni}_{25}$, 32 atoms, 5000 K)**: Dataset A contains snapshots at every timestep (1 fs spacing) whereas dataset B contains snapshots taken at 100-timestep (100 fs spacing). (Left) Distribution of the descriptor displacement over one sampling step, i.e. $\Delta t = 1$ fs (dataset A) and 100 fs (dataset B). (Right) Descriptor displacement as a function of time separation Δt averaged over all t and grid points.

Δt . Small changes indicate that the two snapshots supply nearly duplicate inputs; large changes indicate decorrelated snapshots. Let $\mathbf{D}_i(t) \in \mathbb{R}^{5M}$ denote the atomic-neighborhood descriptor (Eq.(1) of the main document) evaluated at the i -th grid point, at time t . Each descriptor component is standardized using its mean and standard deviation. The standardized descriptor is denoted as $\tilde{\mathbf{D}}_i(t)$. The *descriptor displacement* over a time separation Δt is then defined as the Euclidean distance between the standardized descriptors of the *same* grid point evaluated at two snapshots separated by Δt as

$$\delta_i(t, \Delta t) = \left\| \tilde{\mathbf{D}}_i(t + \Delta t) - \tilde{\mathbf{D}}_i(t) \right\|_2. \quad (\text{S4})$$

Since the grid is fixed in the simulation cell while the atoms evolve, δ_i measures how much the local atomic environment of a given spatial location changes over Δt . The smaller the value of the *descriptor displacement*, the greater the correlation between the snapshots. Figure 7(left) shows the probability density of $\delta_i(t, \Delta t)$ for $\Delta t = 1$ fs (dataset A) and $\Delta t = 100$ fs (dataset B). These distributions clearly show that the grid points in dataset A change only slightly between consecutive snapshots, whereas those in dataset B exhibit substantially larger changes. This observation further confirms that consecutive configurations in dataset A are temporally correlated, while those in dataset B are effectively uncorrelated. Figure 7 (right) shows the descriptor distance, $\delta_i(t, \Delta t)$, as a function of the time separation, Δt . For dataset A, the descriptor distance increases with Δt and converges after approximately 90 fs indicating that the descriptors have become temporally uncorrelated. In contrast, for dataset B, the descriptor distance reaches

the same converged value after a single sampling step, demonstrating that the selected data are already temporally uncorrelated.

Furthermore, uncorrelated datasets (such as dataset B shown above) are constructed from independent simulations performed at different temperatures and compositions, further reducing temporal correlations within the dataset. For all systems considered in the present work, snapshots are separated by a minimum interval of 100 fs. Therefore, our analysis indicates that the high redundancy in the data does not arise from temporal correlations.

Temporal correlations may exist within a dataset, and the degree of redundancy may depend on the extent of these correlations. However, quantifying redundancy as a function of dataset correlations is beyond the scope of the present work.

VII. QUANTIFYING THE COVERAGE OF A SPECIFIC DISTRIBUTION BY A SUBSET

The GraNd score and other state-of-the-art importance-based pruning methods often perform poorly at higher pruning rates because of poor data coverage.

We have evaluated the AUC_{pr} by computing the expected minimum distance from the data in the validation dataset to the data in coreset.

From the Table 1, we can see that, for 90% and 99% pruning percentage, the CCS has lower AUC_{pr} than the GraNd method. It confirms that the CCS method based

Aluminum				
Pruning Rate	AUC _{pr}		NRMSE	
	GraNd	CCS	GraNd	CCS
90%	8.396	8.30	1.58×10^{-2}	1.15×10^{-2}
99%	8.727	8.584	2.55×10^{-2}	1.3×10^{-2}
Water				
Pruning Rate	AUC _{pr}		NRMSE	
	GraNd	CCS	GraNd	CCS
90%	23.18	22.93	3.1×10^{-2}	1.09×10^{-3}
99%	25.29	24.89	1.11×10^{-1}	1.9×10^{-3}
SiGeSn				
Pruning Rate	AUC _{pr}		NRMSE	
	GraNd	CCS	GraNd	CCS
90%	76.57	76.30	2.55×10^{-2}	8.27×10^{-3}
99%	81.91	80.537	8.27×10^{-2}	1.17×10^{-2}
CrFeCoNi				
Pruning Rate	AUC _{pr}		NRMSE	
	GraNd	CCS	GraNd	CCS
90%	13.074	13.039	1.7×10^{-2}	2.3×10^{-3}
99%	13.416	13.335	7.45×10^{-2}	3×10^{-3}

Supplementary Table 1. Comparison of AUC_{pr} and NRMSE between GraNd and CCS method for different pruning percentages

coreset indeed selects a subset that ensures a better data coverage, as evidenced by its lower errors (NRMSE).

VIII. GENERALIZABILITY

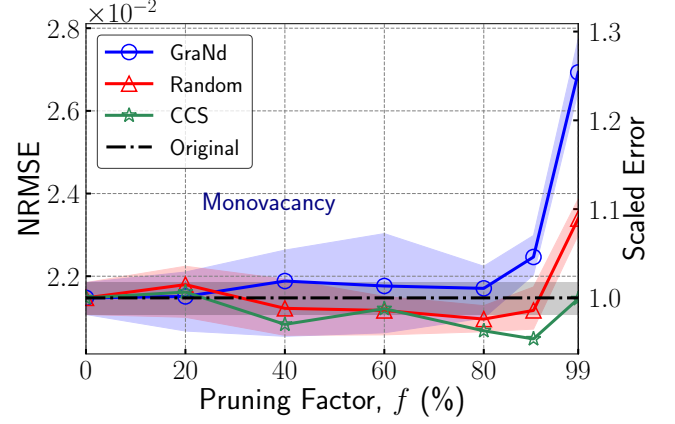
In this section, we provide additional results that demonstrate the generalizability of ML models trained on pruned datasets.

A. Structural Defects and Ordered Configurations

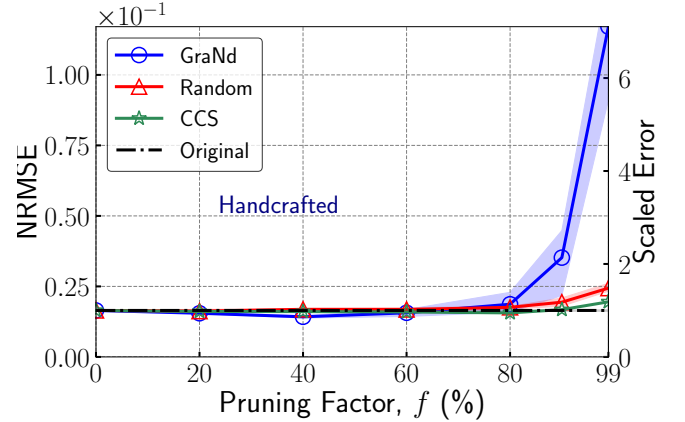
The error in electron density prediction for defects in the 32-atoms aluminum system is shown in Figure 8. The generalization error of the GraNd method increases rapidly beyond 80% pruning, whereas random pruning maintains high accuracy up to 90% pruning. The CCS method supports a remarkably high pruning factor (99%) without any significant loss in generalization performance.

The errors in electron density prediction for handcrafted checkerboard SiGeSn alloy systems is shown in Figure 9. Checkerboard systems of 216 atoms and 64 atoms were considered. The generalization performance for GraNd worsens rapidly beyond 60%, whereas random pruning maintains high accuracy up to 80% pruning factor. However, CCS maintains near-original performance until 99% pruning.

For CrFeCoNi system with defects, the errors in electron density prediction for defects is shown in Figure 10. The generalization performance for GraNd

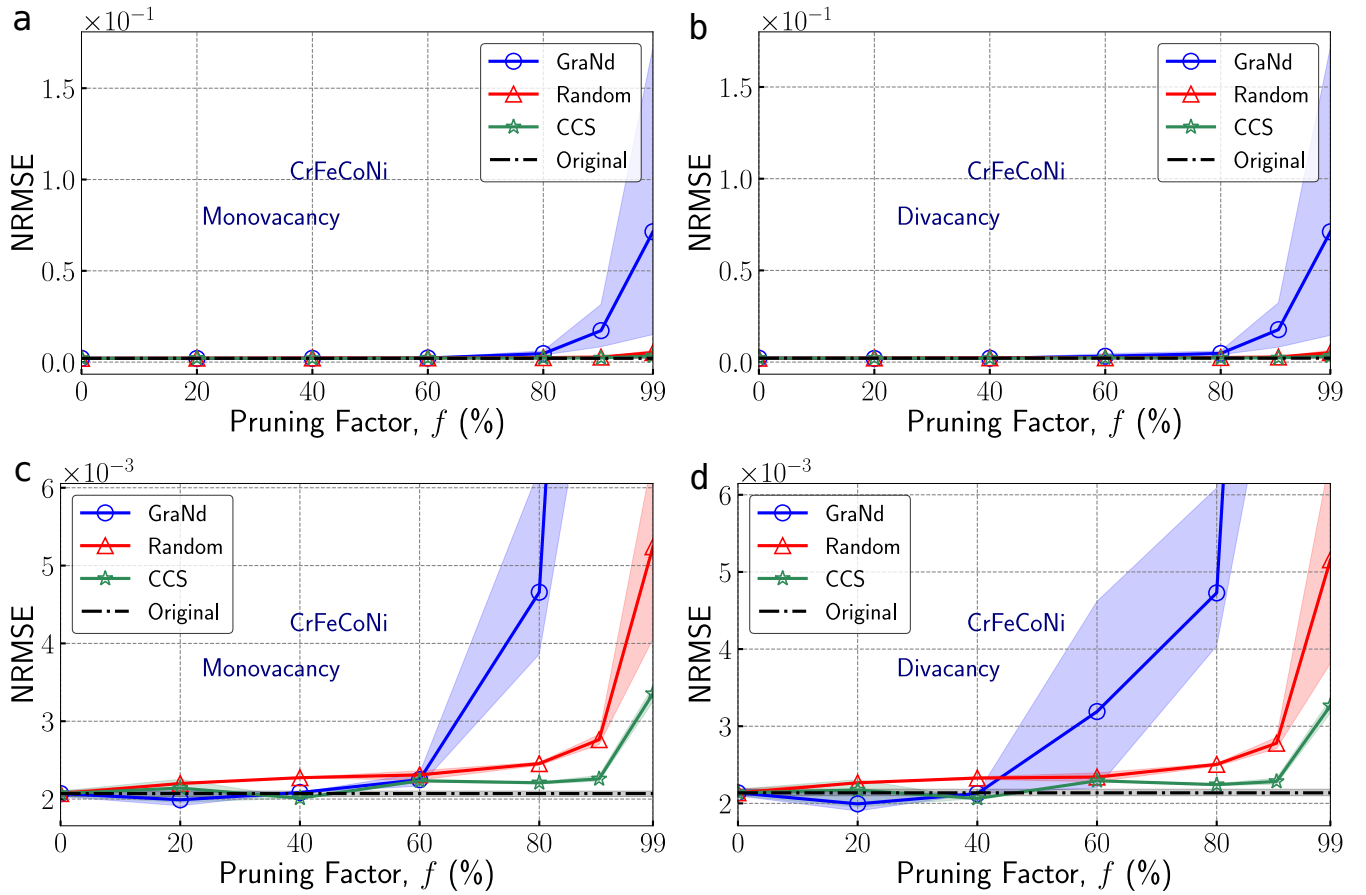


Supplementary Figure 8. **Structural Generalizability tests for Aluminum system:** The error in the electron density prediction for various pruning factors for monovacancies for 32-atoms Aluminum system. The shaded region represents the range (maximum to minimum) of three ML models around the mean, shown by a solid line.



Supplementary Figure 9. **Structural Generalizability tests for SiGeSn system:** The error in the electron density prediction for various pruning factors for SiGeSn handcrafted checkerboard systems. The shaded region represents the range (maximum to minimum) of three ML models around the mean, shown by a solid line.

worsen rapidly beyond 60%, whereas random pruning maintains high accuracy up to 80% pruning factors. However, CCS maintains near-original performance until 99% pruning and show very little degradation at 99% pruning compared to the other methods. We also compute the total energy of the system from the ML-predicted electron density, via postprocessing, for these generalizability tests. The total energy errors (Hartree/atom) relative to the KS-DFT calculations, are shown in Figure 11, which shows that the 90% pruning through CCS performs in the same level as the original



Supplementary Figure 10. **Structural Generalizability tests for CrFeCoNi system:** The error in the electron density prediction for various pruning factors for (a) mono- (b) di-vacancy defects for 32-atoms CrFeCoNi system. The shaded region represents the range (maximum to minimum) of three ML models around the mean, shown by a solid line. Figures (c) and (d) show the magnified views of Figures (a) and (b).

dataset, achieving close to the chemical accuracy. However, 99% pruning through CCS exceeds chemically accurate threshold.

B. Generalization across compositions

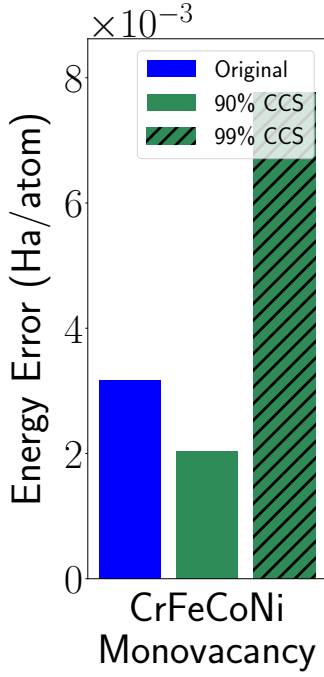
To test extrapolation across composition space, we evaluate the ML models for CrFeCoNi on 36 unseen compositions. The training and test compositions are shown in the composition space of CrFeCoNi in the Figure. 12. This constitutes a stringent test for generalizability. The results of pruning are reported in Figure 13. The 90% CCS-pruned model shows a slight increase in error relative to the original model, while remaining chemically accurate. The performance of the 99% CCS-pruned model degrades further, exceeding the chemical-accuracy threshold. Such increase in error for high factors of pruning is the expected for unseen compositions.

C. Temperature Extrapolation

Generalizability beyond the temperature range of training data is tested. The aluminum models are trained on configurations sampled at six simulation temperatures (298, 600, 900, 1200, 1500, and 1800 K) and tested on unseen dataset generated at 2100K. As shown in Figure 14, the errors in the electron density prediction and energy are very close for the original, 90% CCS, and 99% CCS datasets, and all three energy errors remain well within chemical accuracy. The pruned-data models therefore reproduce the temperature extrapolation behavior of the full-dataset model with minimal degradation even at 99% pruning.

D. Size Extrapolation

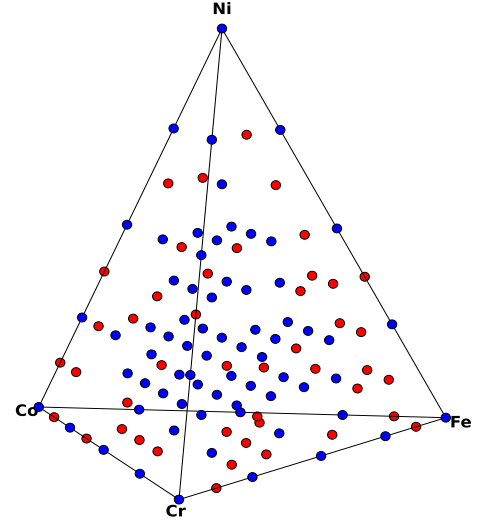
Generalizability beyond the size of training system is tested. We evaluate the SiGeSn models on 216-atom systems whereas only 64-atom systems are used in training.



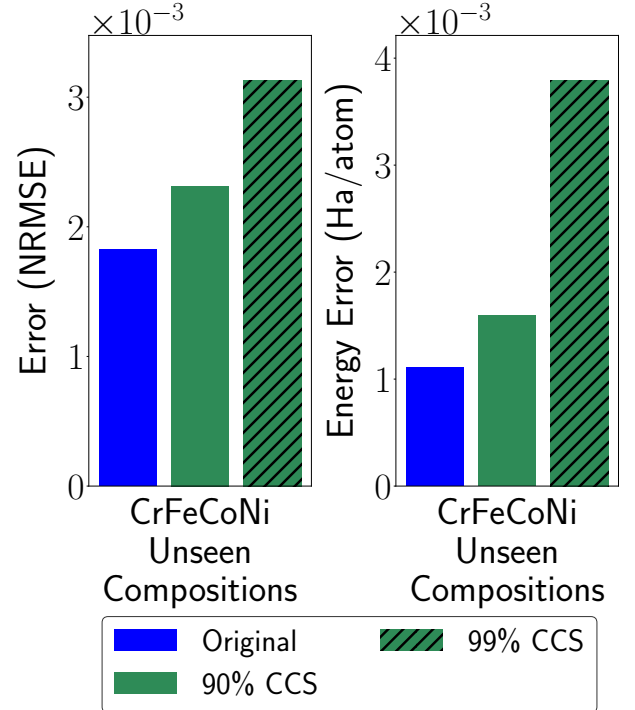
Supplementary Figure 11. **Structural Generalizability tests for CrFeCoNi**

Monovacancy: The error in the energy from the ML-predicted electron density for the original dataset, 90% and 99% CCS-based dataset.

We have used three compositions of SiGeSn at 2400 K. The density prediction error increases with pruning fraction slightly; however, the 90% CCS model maintains chemical accuracy in energy, and the 99% CCS model remains close to the chemical-accuracy threshold, as shown in Figure 15.

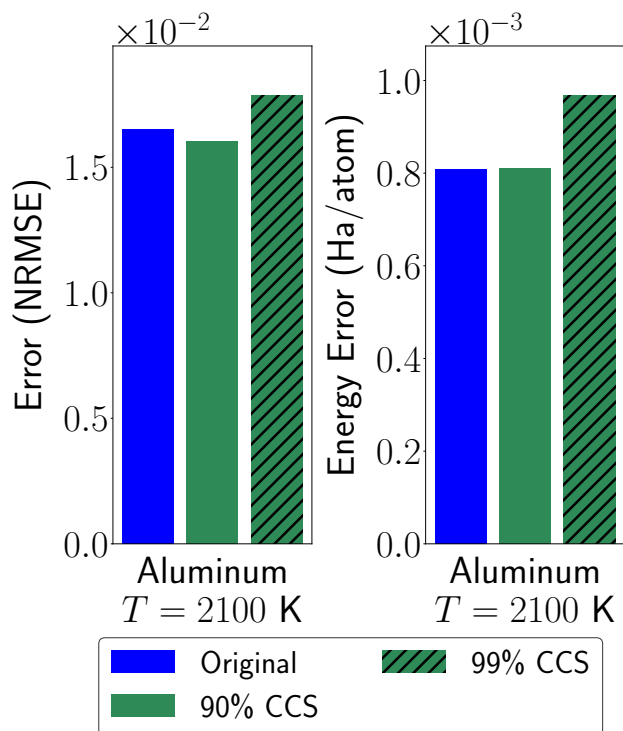


Supplementary Figure 12. Composition space of the CrFeCoNi alloy systems. We have used 69 compositions (blue) during training; and tested on 36 unseen compositions (red) for generalization.

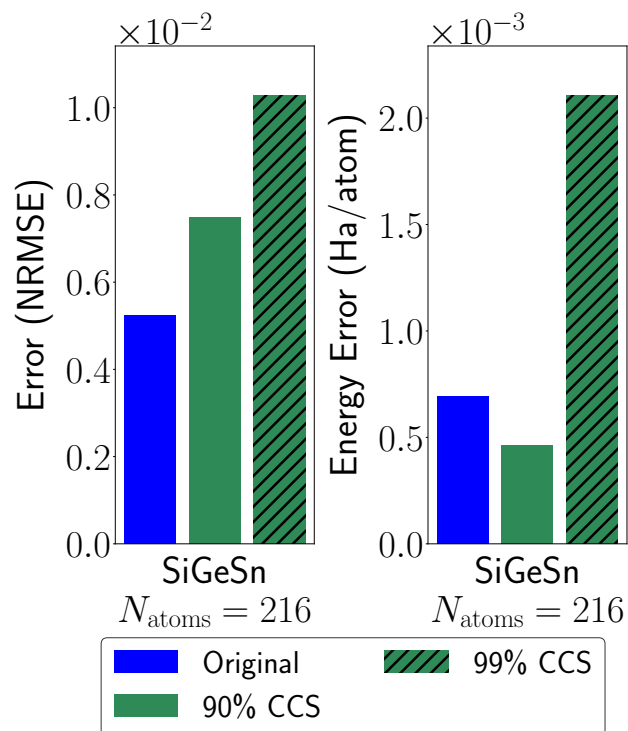


Supplementary Figure 13. **Generalization tests for unseen compositions of CrFeCoNi:** (left)

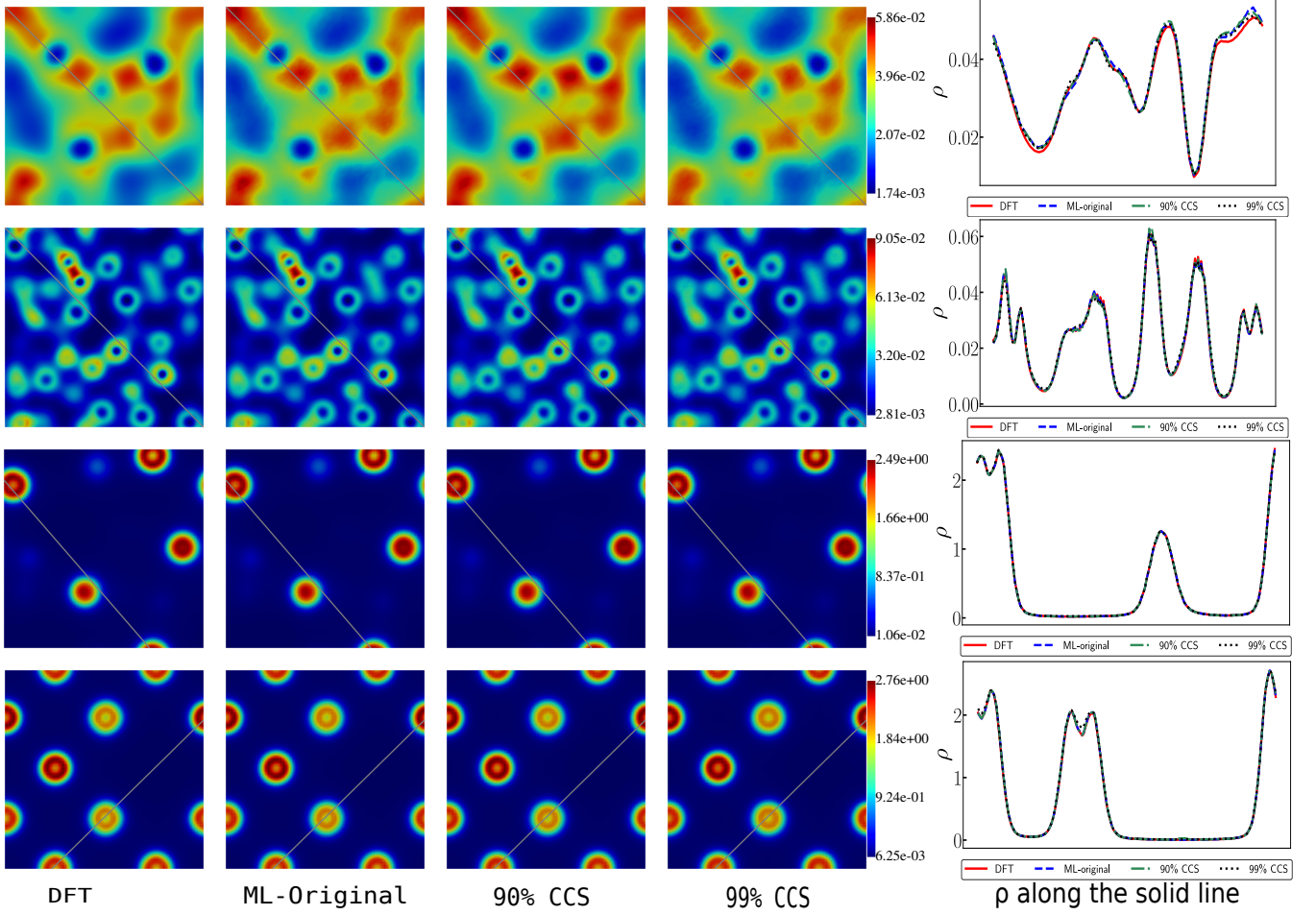
The error in the electron density prediction and (right) the energy error, for models trained on the original, 90% CCS, and 99% CCS pruned datasets.



Supplementary Figure 14. **Generalization tests for temperature extrapolation for aluminum:** (left) The error in electron density prediction and (right) the energy error for Aluminum 32-atoms system at 2100K, for models trained on the original, 90% CCS, and 99% CCS pruned datasets.

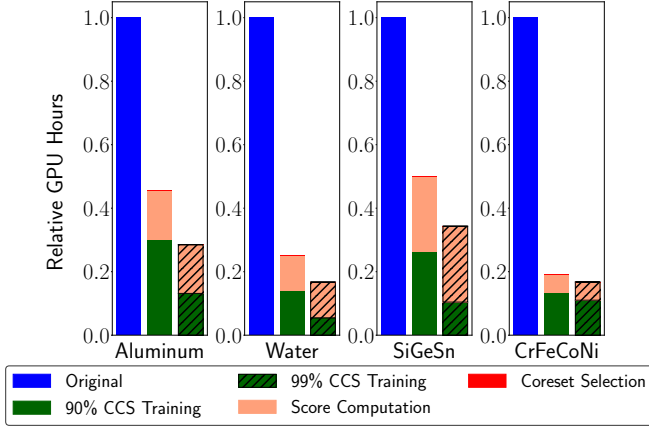


Supplementary Figure 15. **Generalization tests for size extrapolation for SiGeSn:** (left) The error in electron density prediction and (right) the energy error for SiGeSn 216-atoms system at 2400K, for models trained on the original, 90% CCS, and 99% CCS pruned datasets.



Supplementary Figure 16. **Model generalizability to unseen configurations:** 2D slices of electron density obtained by KS-DFT, ML model trained on the original dataset, 90% CCS, and 99% CCS based pruned datasets. The electron density(ρ) along the solid line is also compared with KS-DFT, showing remarkable agreement. (*First row*) Temperature extrapolation test for 32-atoms aluminum systems. (*Second row*) System size extrapolation test for SiGeSn system. (*Third row*) Unseen compositions test for the CrFeCoNi system (*Fourth row*) Mono-vacancy defect for 32-atoms CrFeCoNi systems. The unit of electron density is Bohr⁻³ in atomic units.

IX. TIME COMPARISON



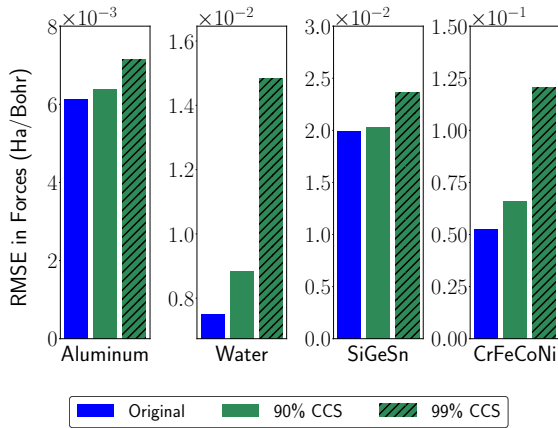
Supplementary Figure 17. Time comparison between the ML model trained on the original dataset, 90% CCS and 99% CCS based pruned dataset. The coreset selection time is small and hence is not visible.

The full pipeline cost, including GraNd score computation, coreset selection, and training on the pruned dataset, relative to training on the original dataset are shown in Figure 17. The cost of coreset selection step through CCS is negligible across all systems, taking 30 to 60 seconds. The GraNd score computation costs significant fraction of computational overhead. Even accounting for this cost, the total computational cost is substantially lower than training on the original dataset for all systems: at 99% CCS pruning, corresponding to speedups of roughly $3\times$ to $6\times$ over the original training time. At 90% CCS pruning, the speedups range from approximately $2\times$ to $5\times$.

X. VISUALIZATION OF REDUNDANCY

To visualize which grid-points were kept or pruned, we have plotted the spatial distribution of retained and pruned grid points for Aluminum, CrFeCoNi and SiGeSn at 90% CCS pruning, as shown in Figure 18. The figures reveal that CCS retains grid points in regions of high and rapidly varying electron density — e.g. near atoms and bonding regions — while predominantly pruning the low-density interstitial regions. For aluminum, whose density variation is comparatively less rapid, the retained points are correspondingly more uniformly distributed. In all cases, a small number of scattered interstitial points are also retained, consistent with CCS’s stratified sampling across the importance scores. The CrFeCoNi system does show strong clustering of retained data points near the atomic nuclei, very likely due to the strongly localized pseudocharges associated with the elements in this system, arising from the inclusion of semi-core states in the corresponding pseudopotentials. Notably, the spatial pattern of data retention reflects the underlying electronic structure of the material. Importantly, this physically meaningful spatial selection pattern emerges naturally from the ML-based importance scores, without any explicit physics-informed assumption about which spatial regions are informative.

XI. VALIDATION OF FORCES

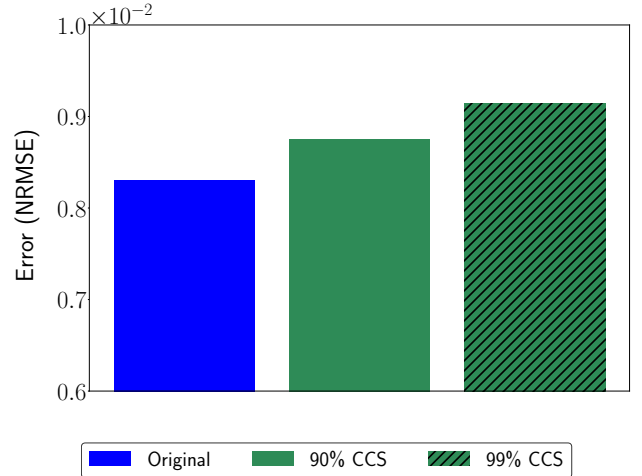


Supplementary Figure 19. Error(RMSE) in atomic forces, as obtained from the ML-predicted electron density for the original dataset, 90% CCS, and 99% CCS-based pruned datasets.

It should be noted that the ML model is not trained using force data. Consequently, the forces, which are computed solely from the ML predicted electron densities, may be predicted less accurately than the electron densities themselves. We have carried out this analysis of force calculation, and our key finding is that, like the energies, the force errors do not necessarily get much worse,

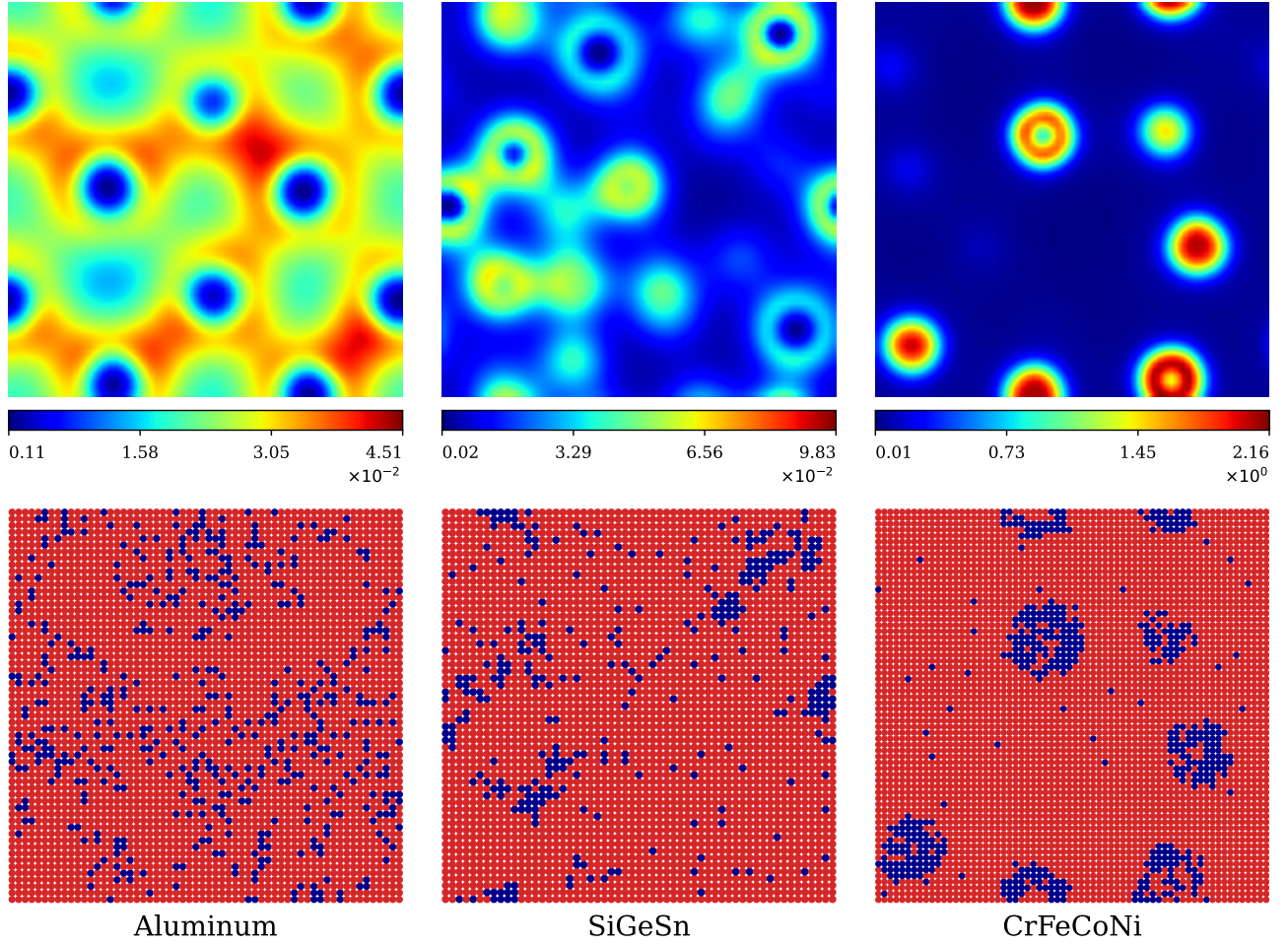
depending on the level of pruning, i.e., the pruned ML models have similar force errors (with respect to KS-DFT references), as do the forces computed from the original ML model. Specifically, from Figure 19 above, we see that the errors in the forces of the ML models, with respect to reference KS-DFT results (force errors with respect to KS-DFT results for each snapshot were computed as the L^2 norm of the difference, over all atoms and all force components). Although the errors in the forces remain low for the 90% pruning case, they increase significantly for the 99% pruning case for some cases. We find it pertinent to point out, that in general, the errors in the forces of the ML models are somewhat worse than errors in the energies. However, this is to be expected, since force errors are known to be far more sensitive to self-consistency errors in the electron density [17], as compared to the energy, which benefits from a variational structure and the quadratically correct Harris-Foulkes formula [8]. Notably, in terms of force errors, the performance of the original ML model is the worst for the CrFeCoNi system, due to the hard pseudopotentials associated with the elements present. Therefore, the error in forces does not increase significantly up to 90% pruning.

XII. PRUNING USING THE SOAP DESCRIPTOR



Supplementary Figure 20. **Redundancy using the SOAP descriptor for the SiGeSn system:** Error in electron density prediction for the ML models trained on the original, 90% and 99% CCS pruned datasets.

We note that the pruning approaches themselves (GraNd and CCS) are descriptor-agnostic by construction, as they operate on training dynamics and stratified sampling across importance scores. To test whether the observed redundancy is an artifact of our descriptor choice, we have repeated the pruning experiments for a



Supplementary Figure 18. Spatial distribution of pruned and retained grid points for Aluminum, SiGeSn, and CrFeCoNi at 90% CCS pruning. For each material, the top figure shows the KS-DFT electron density on a representative 2D slice, and the bottom figure shows the corresponding spatial map of pruned (red) and retained (blue) grid points. The unit of electron density is Bohr^{-3} in atomic units.

representative system (SiGeSn) using the Smooth Overlap of Atomic Positions (SOAP) descriptors [18] in place of the currently used descriptors, keeping everything else the same. The results are shown in Figure 20. The redundancy findings persist under the SOAP descriptor as well. The models trained on 90% CCS-pruned data achieve near-original performance, and even at 99% pruning, the error increase is very small. Such findings imply that the observed redundancies are generalizable to other descriptors.

XIII. ESTIMATING INTRINSIC DIMENSIONS

We estimated the intrinsic dimension (ID) using several additional estimators, namely Gride [19], MiND_{ML} [20], iterative-based intrinsic Dimension (IID), and the modified intrinsic dimension (MID) estimator [21]. The results are summarized in Table 2. The

Gride, MiND_{ML}, and IID estimators produce ID values that corroborate closely with the maximum likelihood estimator (MLE) values reported in the manuscript for each system. Although there is no universally accepted method for accurately estimating the intrinsic dimension of high-dimensional nonlinear manifolds for real datasets, these independent estimators show close agreements. However, the manifold nonlinearity may lead to underestimation of the intrinsic dimension depending on the dataset’s high dimensionality. To account for this effect, we also employed the modified intrinsic dimension (MID) estimator, which yields the most conservative (largest) estimates among the methods considered. The MID estimator yields systematically larger IDs: 131 for aluminum, 196 for SiGeSn, 200 for water, and 201 for CrFeCoNi. Importantly, even the largest estimates remain below the corresponding PCA dimensions required to explain 99% of the variance (206–259) and are substantially lower than the ambient descriptor dimensions (236–295) for all systems considered. Therefore,

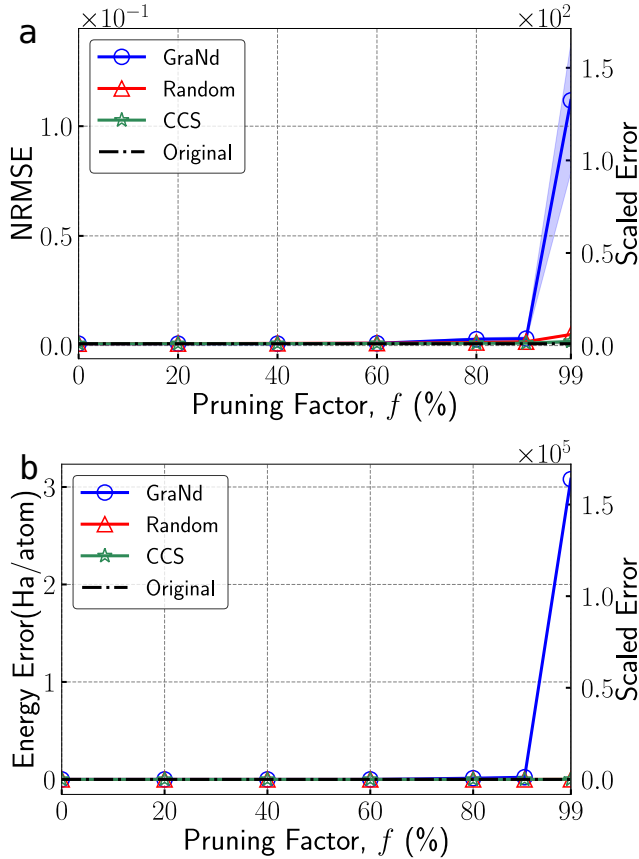
System	MLE	IID	GRIDE	MiND _{ML}	MID	PCA (99%)	Ambient
Aluminum	60 - 65	72	55 - 60	68 - 70	131	206	236
Water	83 - 92	95	75 - 83	90 - 93	200	247	295
SiGeSn	78 - 89	97	69 - 78	89 - 94	196	243	295
CrFeCoNi	78 - 88	96	68 - 80	89 - 93	201	259	295

Supplementary Table 2. Intrinsic dimension (ID) estimates for the four material systems using several estimators, compared with the number of principal components required to capture 99% of the variance (PCA) and the ambient descriptor dimension. The neighbourhood-based estimators (MLE, GRIDE, MiND_{ML}) are reported as the range obtained by considering the following number of neighbors: 100, 250, 500, 750, 1000. The IID as well as the MID sweep across 100 – 1000 nearest neighbors for their estimation.

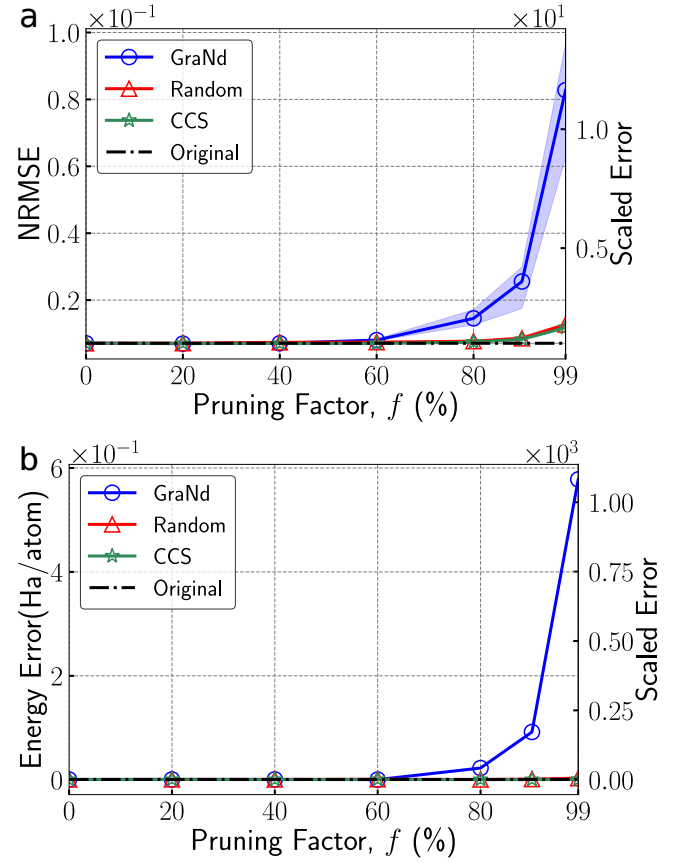
our conclusions are robust with respect to the choice of intrinsic-dimension estimator. Regardless of the estimation method, the electronic-structure descriptors occupy a nonlinear manifold whose intrinsic dimension is significantly smaller than the ambient descriptor space. At the same time, the variation among the estimators highlights the inherent uncertainty in assigning a single numerical value to the intrinsic dimension of complex, high-dimensional nonlinear manifolds and suggests that the reported IDs should be interpreted as method-dependent estimates rather than exact quantities.

XIV. ADDITIONAL RESULTS

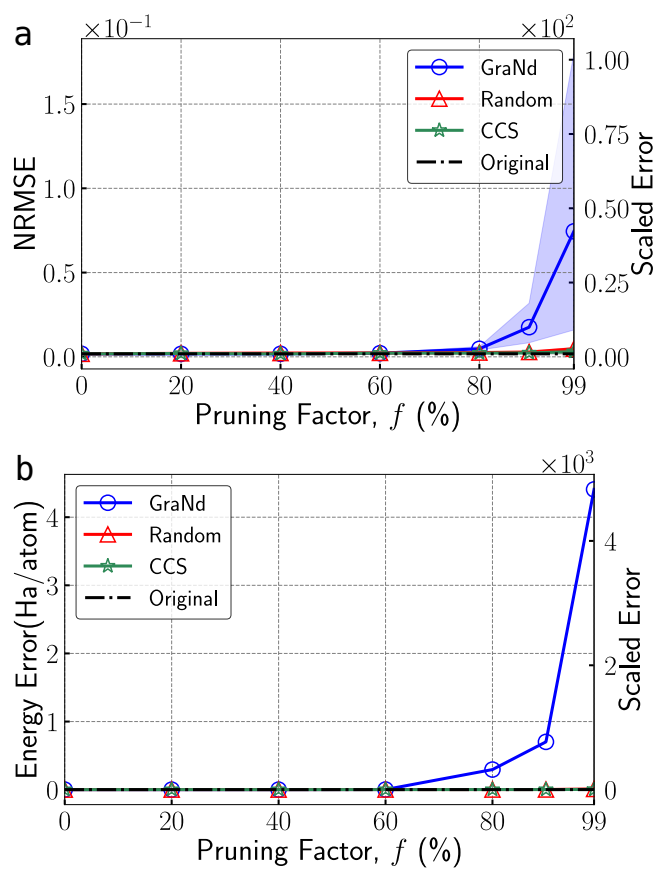
A complete and unscaled view of the error versus pruning factor for water, SiGeSn, and CrFeCoNi, (corresponding to Figure 4 of the main text) is provided in Figure 21, Figure 22, and Figure 23 for completeness. It shows that the GraNd method becomes too erroneous beyond 60% pruning for water, SiGeSn and CrFeCoNi. However, the random pruning provides high accuracy up to 90%.



Supplementary Figure 21. **Water system:** (a) Error in the predicted electron density for various pruning factors. (b) Error in energy obtained from the predicted electron density for various pruning factors.



Supplementary Figure 22. **SiGeSn system:** (a) Error in the predicted electron density for various pruning factors. (b) Error in energy obtained from the predicted electron density for various pruning factors.

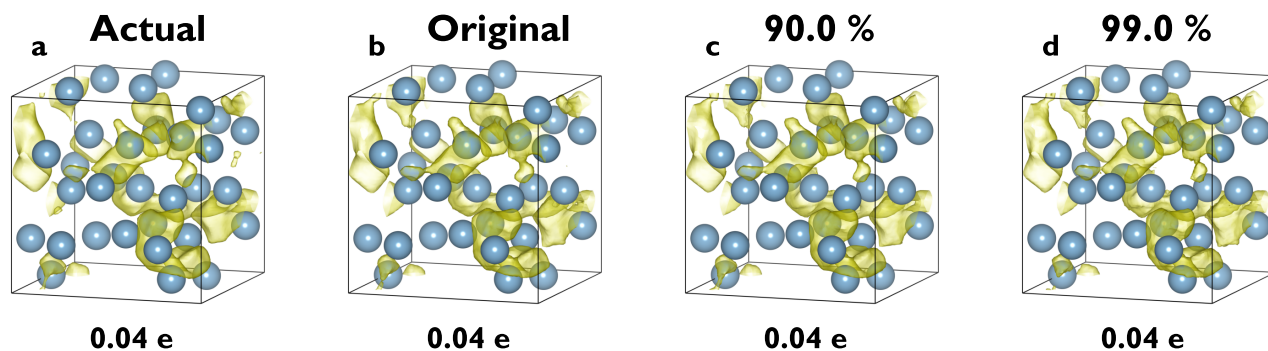


Supplementary Figure 23. **CrFeCoNi system:** (a) Error in the predicted electron density for various pruning factors. (b) Error in energy obtained from the predicted electron density for various pruning factors.

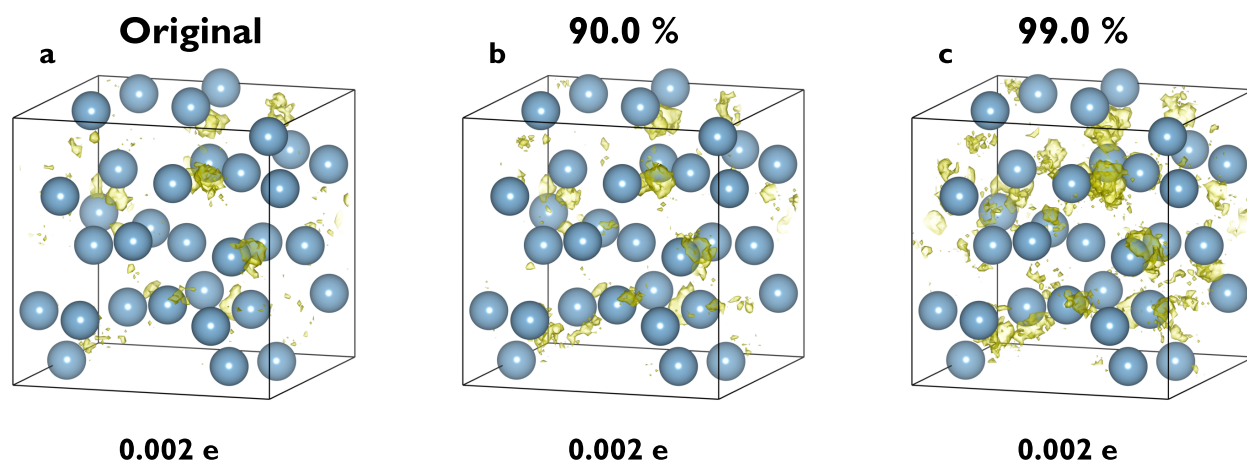
A. Isosurface Plots

The 3D electron density fields are shown as iso-surface plots for the four materials in Figures 24, 26, 28, and 30. The overall pattern of electron density iso-surfaces is well captured by the original dataset, 90%, and 99% CCS based pruned datasets.

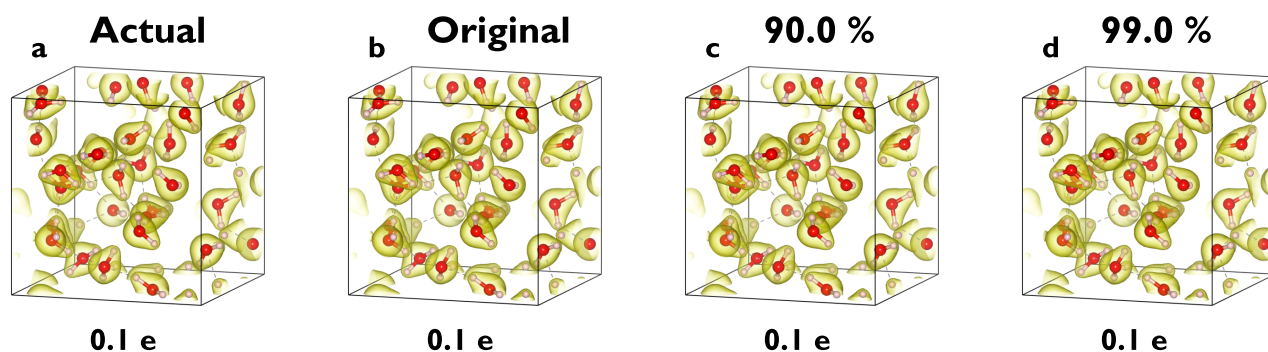
The corresponding 3D error fields are shown in Figures 25, 27, 29, and 31. Although the error iso-surfaces closely resemble those of the original datasets even after 90% and 99% pruning, a slight yet progressive increase in error is observed as pruning levels increase. These iso-surface visualizations are generated with the VESTA [22] software.



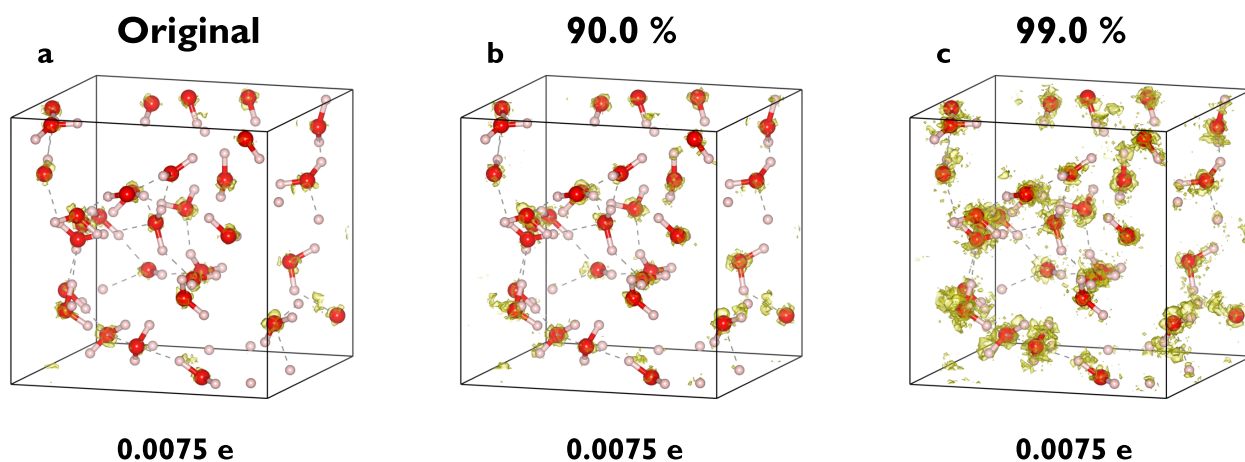
Supplementary Figure 24. **Aluminum system at 1500 K.** Iso-surfaces of electron density obtained by KS-DFT (a) and ML model for original (unpruned) (b), 90% CCS (c), and 99% CCS (d).



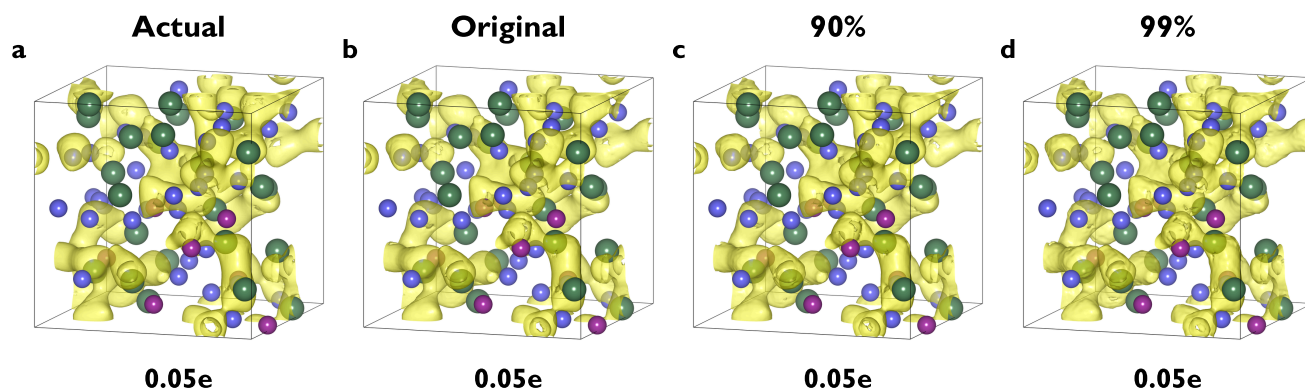
Supplementary Figure 25. **Aluminum system at 1500 K.** Iso-surfaces of errors in electron density obtained by ML model for original (a), 90% CCS (b), and 99% CCS (c) respectively.



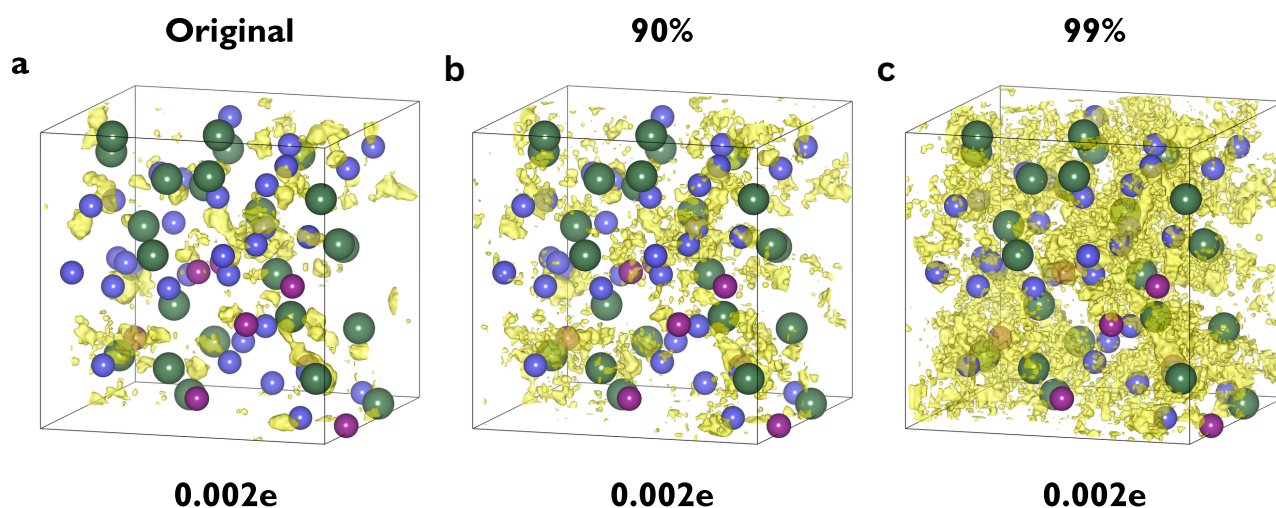
Supplementary Figure 26. **Water system at 600K:** Iso-surfaces of electron density (at 0.1 e) obtained by DFT (a) and ML model for original (b), 90% CCS (c), and 99% CCS (d). The red and pink spheres represent the Oxygen and Hydrogen atoms.



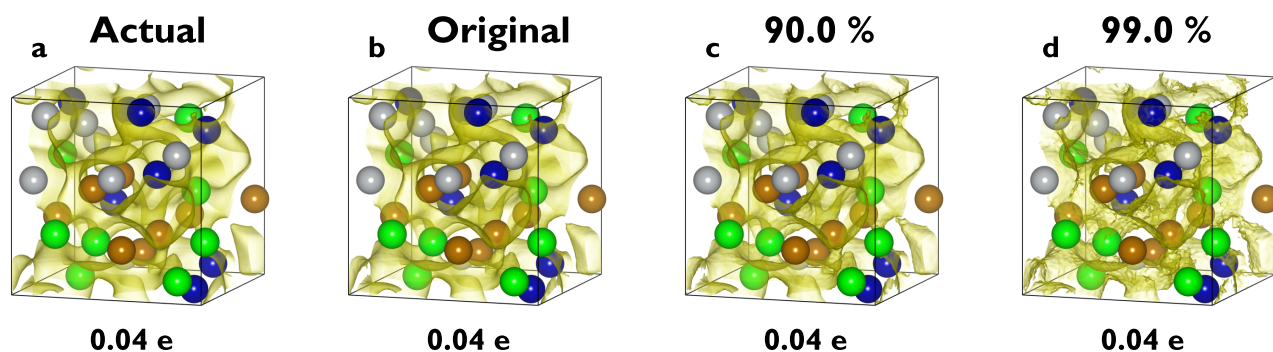
Supplementary Figure 27. **Water system at 600K**: Iso-surfaces of errors in electron density (at 0.0075 e) obtained by ML model for original (unpruned) (a), 90% CCS (b), and 99% CCS (c) respectively.



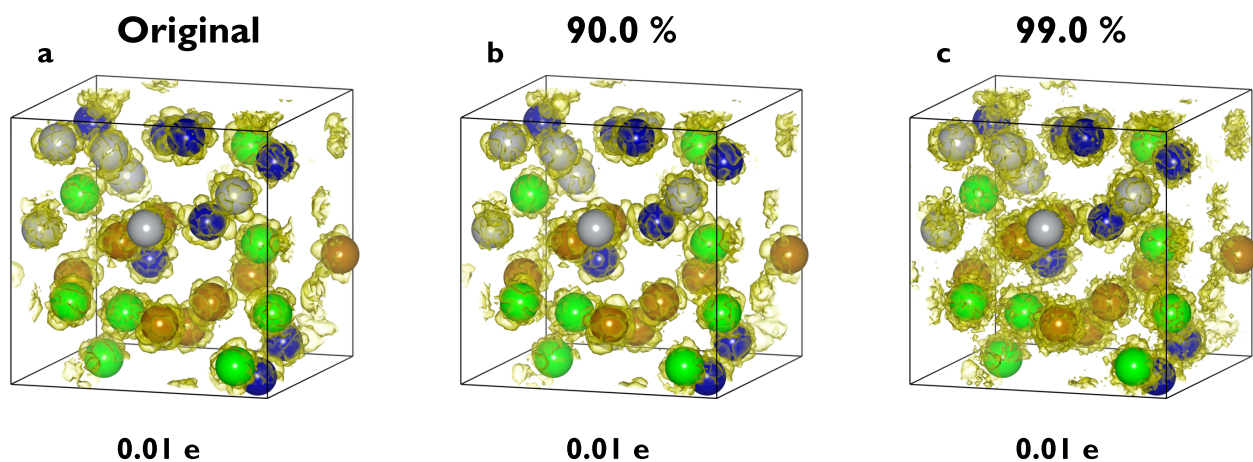
Supplementary Figure 28. **$\text{Si}_{50}\text{Ge}_{12.5}\text{Sn}_{37.5}$ system at 2400K**: Iso-surfaces of electron density obtained by DFT (a) and ML model for original (unpruned) (b), 90% CCS (c), and 99% CCS (d). The blue, purple, and green spheres represent the Si, Ge, and Sn atoms.



Supplementary Figure 29. **$\text{Si}_{50}\text{Ge}_{12.5}\text{Sn}_{37.5}$ system at 2400K**: Iso-surfaces of errors in electron density obtained by ML model for original (a), 90% CCS (b), and 99% CCS (c) respectively.



Supplementary Figure 30. $\text{Cr}_{25}\text{Fe}_{25}\text{Co}_{25}\text{Ni}_{25}$ system at 5000K: Iso-surfaces of electron density obtained by DFT (a) and ML model for original (unpruned) (b), 90% CCS (c), and 99% CCS (d). The green, orange, blue, and gray spheres represent the Cr, Fe, Co, and Ni atoms.



Supplementary Figure 31. $\text{Cr}_{25}\text{Fe}_{25}\text{Co}_{25}\text{Ni}_{25}$ system at 5000K: Iso-surfaces of errors in electron density obtained by ML model for original (a), 90% CCS (b), and 99% CCS (c) respectively.

-
- [1] S. Pathrudkar, P. Thiagarajan, S. Agarwal, A. S. Banerjee, and S. Ghosh, *npj Computational Materials* **10**, 175 (2024).
 - [2] K. R. Briling, A. Fabrizio, and C. Corminboeuf, *The Journal of Chemical Physics* **155** (2021).
 - [3] J. M. Alred, K. V. Bets, Y. Xie, and B. I. Yakobson, *Composites Science and Technology* **166**, 3 (2018).
 - [4] S. Pathrudkar, H. M. Yu, S. Ghosh, and A. S. Banerjee, *Physical Review B* **105**, 195141 (2022).
 - [5] Q. Xu, A. Sharma, B. Comer, H. Huang, E. Chow, A. J. Medford, J. E. Pask, and P. Suryanarayana, *SoftwareX* **15**, 100709 (2021).
 - [6] Q. Xu, A. Sharma, and P. Suryanarayana, *SoftwareX* **11**, 100423 (2020).
 - [7] S. Ghosh and P. Suryanarayana, *Computer Physics Communications* **212**, 189 (2017).
 - [8] J. Harris, *Physical Review B* **31**, 1770 (1985).
 - [9] W. M. C. Foulkes and R. Haydock, *Physical review B* **39**, 12520 (1989).
 - [10] D. Hendrycks and K. Gimpel, *arXiv preprint arXiv:1606.08415* (2016).
 - [11] D. P. Kingma and J. Ba, *arXiv preprint arXiv:1412.6980* (2014).
 - [12] I. Loshchilov and F. Hutter, *arXiv preprint arXiv:1711.05101* (2017).
 - [13] M. Paul, S. Ganguli, and G. K. Dziugaite, *Advances in neural information processing systems* **34**, 20596 (2021).
 - [14] H. Zheng, R. Liu, F. Lai, and A. Prakash, in *The Eleventh International Conference on Learning Representations* (2023).
 - [15] T. Hsu, B. Sadigh, V. Bulatov, and F. Zhou, *Journal of Chemical Theory and Computation* **20**, 2335 (2024).
 - [16] R. Galindo-Murillo, D. R. Roe, and T. E. Cheatham III, *Biochimica et Biophysica Acta (BBA)-General Subjects* **1850**, 1041 (2015).
 - [17] M.-C. Kim, E. Sim, and K. Burke, *The Journal of chemical physics* **140** (2014).
 - [18] A. P. Bartók, R. Kondor, and G. Csányi, *Physical Review B* **87**, 184115 (2013).
 - [19] F. Denti, D. Doimo, A. Laio, and A. Mira, *Scientific Reports* **12**, 20005 (2022).
 - [20] A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli, *Machine learning* **89**, 37 (2012).
 - [21] H. Qiu, Y. Yang, and H. Pan, *Pattern Recognition* **140**, 109580 (2023).
 - [22] K. Momma and F. Izumi, *Journal of Applied crystallography* **41**, 653 (2008).