coatings

MDPI

Article

Stacking Ensemble Learning with Genetic Algorithm Optimization for Multi-Property Prediction and Inverse Design of BCC-Type V-Based Hydrogen Storage Alloys

Yishen Wu and Xiaofang Chen *

College of Chemistry, Chemical Engineering and Materials Science, Soochow University, Suzhou 215000, China; 2309404101@stu.suda.edu.cn

* Correspondence: xfchen75@suda.edu.cn

Abstract

Accurate prediction of hydrogen storage properties is needed for accelerating the design of body-centered cubic (BCC)-type V-based alloys, where the composition–property space is too large for trial-and-error experimentation alone. Here we report a stacking ensemble framework that combines XGBoost, Random Forest, Extra Trees, and Gradient Boosting as base learners with linear or ridge meta-learners, with hyperparameters tuned by a genetic algorithm (GA). Three descriptor strategies are compared across seven target properties: composition-only (C), property-only (P), and composition–property fusion (CP). On target-specific subsets containing 95–901 experimental records and 17 alloying elements (the smallest subset, 95 records, corresponds to maximum hydrogen capacity C_{max}, which may limit model stability for that property), the best model for each target gives R^2 values from 0.865 to 0.981 (all metrics are from five-fold cross-validation; no independent external test set was employed); the P-type model for desorption plateau pressure reaches $R^2 = 0.981$. SHapley Additive exPlanations (SHAP) analysis shows that physically derived descriptors, including valence electron concentration, atomic size mismatch, and electronegativity difference, dominate in P and CP models, whereas Ti and Cr contents are the leading compositional features. A Non-dominated Sorting Genetic Algorithm II (NSGA-II) optimizer then ranks 3000 final candidate alloy compositions in six alloy families using four objectives: high predicted retention, high predicted cycle count, high predicted BCC phase ratio, and a low cost proxy; these candidates represent model-based predictions and await experimental synthesis and cycling validation. The V–Ti–Cr–Fe–Mn system contains the largest number of candidates with predicted retention above 99%, while Ti-free V–Cr–Fe–Mn–Al alloys provide low-cost alternatives in the model search space. First-principles calculations on four representative alloys only show that BCC structures are lower in energy than FCC structures by about 0.08–0.13 eV/atom and that hydrogenated structures exhibit clear charge accumulation around H sites, supporting the physical plausibility of the data-driven screening results, though density functional theory (DFT) validation does not replace experimental measurement of PCI curves and cyclic stability.

Keywords: hydrogen storage alloys; machine learning; stacking ensemble; genetic algorithm; SHAP; multi-objective optimization; BCC alloys; high-entropy alloys

Check for updates

Academic Editor: Flavio Deflorian

Received: 2 June 2026
Revised: 22 June 2026
Accepted: 28 June 2026
Published: 2 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Solid-state hydrogen storage in metallic alloys remains one of the most promising routes toward safe, compact, and reversible energy carriers for mobile and stationary applications [1,2]. Among the various alloy families, body-centered cubic (BCC) V- and V–Ti-based solid-solution alloys have attracted sustained attention because of their high theoretical hydrogen capacity (∼3.8 wt%), tunable thermodynamics, and reasonable cycling kinetics [3–7]. Early Ti–Cr–V studies and recent reviews show that alloying with Ti, Cr, Fe, Mn, Al, and related elements modifies the electronic structure, lattice distortion, and phase stability, creating a vast composition space that is difficult to explore through conventional one-alloy-at-a-time experiments [8–11]. The recent emergence of multi-principal-element or "high-entropy" alloys (HEAs) [12,13] has further expanded this design space; Sahlberg et al. [14] demonstrated that the TiVZrNbHf HEA stores nearly twice as much hydrogen as its constituent binary hydrides, and Karlsson et al. [15] confirmed the BCC-to-BCT structural transition during hydrogenation.

Machine learning (ML) offers a data-efficient alternative. Supervised regression models trained on experimental databases can map composition or calculated descriptors to target properties, identify the most influential features, and even propose untested compositions. Gradient-boosted trees [16], random forests [17], and neural networks [18,19] have been applied to predict hydrogen weight percent and plateau-pressure-related descriptors [20,21], V-based alloy capacities [9,22], and hydride thermodynamic properties or inverse-design candidates [23–25]. Despite these advances, most published studies focus on a single target property and a single model architecture, making it hard to compare the relative merits of composition-based versus property-based descriptors, or to conduct multi-property optimization. Specifically, no prior study on BCC V-based alloys has systematically compared composition-only (C), physics-derived property-only (P), and composition–property fusion (CP) descriptor strategies across multiple target properties within a unified ensemble framework, nor coupled such a comparison with multi-objective inverse design and first-principles verification. Table 1 summarizes representative recent ML studies on hydrogen storage alloys and highlights the gap addressed by this work.

Table 1. Comparison of representative ML studies on hydrogen storage alloys. Dataset size, target properties, algorithm family, and inverse-design availability are summarized. This work compares C/P/CP descriptors across seven targets and couples GA-stacking with NSGA-II and DFT verification.

Study	Alloy System	Dataset Size	Target Properties	Algorithm	Inverse Design
Rahnama et al. [20]	Multiple hydrides	∼300	H wt% (single)	RF, NN	No
Witman et al. [21]	Intermetallic hydrides	<100	Thermodynamic (single)	Gradient boosting	No
Lu et al. [22]	V–Ti–Cr–Fe	∼200	Capacity (single)	Ensemble	No
Suwarno et al. [9]	AB_2 hydrides	∼400	Multiple (2)	RF, XGB	No
Beltrame et al. [23]	Various hydrides	Variable	Thermodynamic (single)	RF + augmentation	Partial
Lu et al. [24]	Various H-storage	Large	Multiple (3+)	Multiple	Yes
This work	BCC V-based	95–901	Seven targets (C/P/CP)	GA-stacking	NSGA-II + DFT

Feature engineering is another open question. Raw elemental fractions capture stoichiometry directly, but they ignore the physics that governs hydrogen–metal interactions—valence electron concentration (VEC), atomic size mismatch (δ), electronegativity difference ($\Delta\chi$), and mixing enthalpy (ΔH_{mix}) [26–28]. Property-based descriptors encode this physics at the cost of requiring auxiliary calculations. Whether the extra effort is worthwhile, and whether

fusing both descriptor sets yields further improvement, has not been systematically tested for BCC hydrogen storage alloys.

Hyperparameter optimization is equally critical. Grid or random search scales poorly with the number of learners in an ensemble. Genetic algorithms (GAs) offer a population-based metaheuristic that can explore high-dimensional discrete-continuous search spaces more efficiently [29,30]. Coupling GA-optimized stacking ensembles with post-hoc interpretability tools such as SHapley Additive exPlanations (SHAP) [31] enables both accurate prediction and physical insight.

Beyond prediction, the practical goal is to design new alloys that balance multiple—often competing—objectives: high hydrogen capacity, appropriate plateau pressure, low desorption enthalpy, long cycling life, and affordable raw-material cost. For practical hydrogen storage applications, relevant targets typically include a desorption plateau pressure near ambient (∼0.1–1 MPa at room temperature), an effective capacity exceeding 1.5 wt%, a cyclic retention rate above 90% over at least 100 cycles, and a desorption enthalpy in the range −20 to −40 kJ mol^{-1} H_2 [1,2]. Multi-objective evolutionary algorithms such as NSGA-II [32] can navigate these trade-offs and produce a Pareto front of non-dominated solutions, provided the surrogate models are reliable.

In this work, we build a curated experimental database of 901 BCC V-based alloy records covering 17 elements and seven hydrogen storage properties, then use target-specific subsets to train GA-optimized stacking ensembles under three descriptor strategies. The resulting 21 models are compared against six baseline algorithms, interpreted with SHAP and LASSO analyses, and further used as surrogate models for NSGA-II ranking of 3000 final candidate compositions across six alloy families. Four representative alloys from the optimized set are finally examined by density functional theory (DFT) calculations to test whether the data-driven predictions remain consistent with phase stability and hydrogen-related charge redistribution.

2. Materials and Methods

2.1. Dataset Construction

The experimental dataset was compiled from 88 published studies reporting hydrogen absorption/desorption measurements on BCC-type V-based alloys at near-ambient conditions (temperatures in the range 273–373 K and pressures in the range 0.0001–10 MPa). Records were included if they reported (i) a BCC-type or BCC-dominated alloy with V as a principal element (>10 at.%), (ii) at least one of the seven target properties measured under the stated near-ambient conditions, and (iii) the full alloy composition in at.% summing to 100%. Records were excluded if the primary crystal structure was non-BCC (e.g., Laves-phase AB_2, FCC, or amorphous), if all properties were measured outside the stated temperature–pressure window, or if no quantitative composition data were provided. Duplicate identification was performed by comparing elemental fractions across all 17 elements: entries with all fractions matching within ±0.5 at.% and the same reported preparation route (arc-melting or sintering) were treated as duplicates and removed unless inter-laboratory property differences exceeded one measurement standard deviation, in which case both records were retained to capture genuine inter-laboratory variability. After removing duplicate entries and records without the target property required for a given task, the target-specific subsets contain 95–901 samples; the count varies because not all properties were measured in every study. Each record stores the at.% fractions of 17 elements (Ti, V, Cr, Fe, Mg, Nb, Hf, Mo, Ta, Zr, La, Mn, Y, Ce, Al, Si, and Sc), seven hydrogen storage targets, and a set of calculated physical–chemical descriptors. The target properties are desorption plateau pressure (P_d), absorption plateau pressure (P_a), effective hydrogen capacity (C_{eff}), maximum hydrogen capacity (C_{max}), des-

orption enthalpy (ΔH_{des}), absorption enthalpy (ΔH_{abs}), and cyclic capacity retention rate (η). The descriptor set includes VEC, atomic radius mismatch (δ), Pauling electronegativity difference ($\Delta\chi$), ΔH_{mix}, ΔS_{mix}, the thermodynamic stability parameter Ω, and an alloy cost descriptor. The alloy cost descriptor is a mass-fraction-weighted raw-material price, $C_{\mathrm{alloy}} = \sum_i w_i p_i$ (USD kg^{-1}), where p_i is the 2022–2024 average spot price of element i listed in Supplementary Table S2 and $w_i = x_i M_i / \sum_j x_j M_j$ is the mass fraction converted from atomic fraction x_i and molar mass M_i. This formulation gives a transparent estimate of the relative raw-material price per kilogram of alloy. It should not be interpreted as a complete economic cost, because processing costs, alloying premiums, purity specifications, and economies of scale are not included. The numerical values reported for candidate alloys (e.g., 19.5 USD kg^{-1} for the V–Ti–Cr–Fe–Al system) can therefore be compared across candidates only as approximate raw-material cost indicators.

Plateau pressures span roughly four orders of magnitude (0.0001–10 MPa), so a $\log_{1+x}$ transform was applied to P_{d} and P_{a} before training; the inverse transform was used when evaluating predictions. The $\log(1 + x)$ form was chosen rather than $\log_{10}(x)$ or $\ln(x)$ because it is defined at zero and maps the smallest recorded pressures ($\approx$0.0001 MPa) to a non-negative value, thereby avoiding undefined or strongly negative outputs for near-zero samples that appear in several publications; for values $x \gg 1$ the transform is effectively equivalent to the natural logarithm.

2.2. Descriptor Strategies

Three complementary feature sets were evaluated. The composition-only (C) setting uses the 17 elemental fractions directly, which makes it simple to apply and easy to interpret without auxiliary calculations. The property-only (P) setting replaces raw compositions with six physically derived descriptors: VEC, δ, $\Delta\chi$, ΔH_{mix}, ΔS_{mix}, and Ω. The composition–property fusion (CP) setting combines both sources of information by joining all elemental fractions with all calculated descriptors.

Prior to training, each feature was standardized to zero mean and unit variance. Highly collinear features (Pearson $|r| > 0.95$) were flagged but retained, since tree-based learners are robust to collinearity; the impact was assessed through LASSO coefficient paths (Section 2.7). It is acknowledged that retaining collinear features does not eliminate interpretation difficulties: even though tree ensembles are prediction-robust, SHAP values and permutation importances may distribute importance arbitrarily among correlated features, so the rankings reported here reflect model-level associations rather than independent physical contributions.

Regarding information redundancy in the CP descriptor set: since the physical descriptors (VEC, δ, $\Delta\chi$, ΔH_{mix}, ΔS_{mix}, Ω) are derived algebraically from the elemental fractions, the CP set contains structurally correlated inputs. Tree-based ensembles tolerate such redundancy because their hierarchical splitting is not sensitive to multicollinearity in the way that linear models are; however, the CP strategy should be interpreted as providing the model with both explicit compositional and pre-computed physics-informed views of the same alloy, rather than as two truly independent information sources.

Regarding the impact of feature collinearity on genetic algorithm efficiency: the GA optimizes over the continuous and discrete hyperparameter space of the base learners (tree depth, number of estimators, learning rate, subsampling ratios) rather than over the feature space itself. Consequently, retaining collinear features does not directly expand the GA search space. Across the 21 prediction tasks, convergence was usually reached before the 100-generation limit, with earlier stopping for the pressure models and longer searches for the smaller-capacity subsets. This behavior suggests that the retained correlated features did not noticeably slow GA convergence.

2.3. Stacking Ensemble Architecture

A two-level stacking ensemble was adopted [33,34]. All models were implemented in Python 3.11 using scikit-learn 1.4 [35] and XGBoost 2.0 [16]. The base layer uses three complementary tree-based regressors: XGBoost, Random Forest, and Extra Trees for C-type models, and XGBoost, Gradient Boosting, and Extra Trees for P-type and CP-type models. Random Forest was selected for the C-type base layer (replacing Gradient Boosting) because preliminary tests showed that Random Forest and XGBoost are already sufficiently diverse for the lower-dimensional C descriptor set (17 features), whereas Gradient Boosting offers greater complementarity when the descriptor space is richer (P and CP sets, up to 23 features) where incremental residual fitting provides additional variance reduction. The meta-learner combines base-layer out-of-fold predictions into a final output. Linear regression was used for C-type tasks; Ridge regression with cross-validated regularization ($\alpha \in \{0.1, 1.0, 10.0\}$) was used for P-type and CP-type tasks, which have more features and a higher risk of overfitting. The difference in meta-learner choice (Linear vs. Ridge) reflects the same logic: the C tasks have fewer meta-features (3 base-learner predictions) and a lower risk of meta-level overfitting, while the P and CP tasks benefit from the regularization provided by Ridge.

Each base learner has 5–8 hyperparameters (tree depth, learning rate, number of estimators, subsampling ratio, etc.). Rather than tuning them independently, we wrapped the entire stacking pipeline into a GA optimization loop.

2.4. Genetic Algorithm Hyperparameter Optimization

A real-coded GA with tournament selection, simulated binary crossover, and polynomial mutation was used to search the joint hyperparameter space of all three base learners simultaneously [36]. The fitness function was the negative five-fold cross-validation mean squared error. Population size was 50; the GA ran for up to 100 generations with early stopping (patience = 5–15 generations depending on the target). Parallelization across 10 CPU threads reduced wall-clock time to ∼20 min per task on a desktop workstation (Intel i5-14600KF, 64 GB RAM).

A fixed random seed of 42 was used for all 21 prediction tasks to ensure full reproducibility. Early-stopping patience was assigned proportionally to the target-specific subset size: patience = 5 generations for the two largest subsets (P_d:$n \approx 680$; P_a:$n \approx 520$; Tasks 1, 3, 4, 6, 7, and 9), because the larger sample volume provides more stable per-generation fitness estimates that allow reliable early termination sooner; patience = 10 for the medium-sized subsets (C_{eff}:$n \approx 430$; ΔH_{des}:$n \approx 280$; ΔH_{abs}:$n \approx 210$; η:$n \approx 310$; Tasks 2, 5, 8, 10–12, and 16–21); and patience = 15 for the smallest subset (C_{max}:$n = 95$; Tasks 13–15), where greater stochastic variation in fold estimates necessitates a more conservative stopping criterion to avoid premature convergence.

MSE was selected as the fitness function rather than normalized metrics such as R^2 or relative RMSE for the following reasons: (i) after per-target standardization to zero mean and unit variance, MSE is effectively scale-invariant across all 21 tasks without requiring a task-specific normalization denominator; (ii) the gradient-based components of the tree-based base learners internally minimize an MSE-equivalent criterion, maintaining alignment between the GA search direction and each learner's own optimization objective; and (iii) MSE penalizes large prediction errors quadratically, which is appropriate for alloy property datasets where a small number of compositionally extreme samples can produce disproportionately large errors that normalized metrics would partially mask.

2.5. Model Evaluation

Models were evaluated by five-fold cross-validation. Four metrics are reported: coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). Six single-algorithm baselines—standalone XGBoost, Random Forest, Gradient Boosting, AdaBoost, SVR, and Ridge regression—were trained on the same folds for comparison. All baseline hyperparameters were set to scikit-learn defaults rather than being subject to the same GA optimization as the stacking pipeline; this means the performance comparison reflects the combined benefit of the stacking architecture and the GA-tuned hyperparameters, and the contribution of each factor cannot be separated from the reported metrics. No independent external test set was held out; all reported R^2, RMSE, and MAE values represent cross-validation estimates and may overestimate generalization performance on truly unseen compositions.

2.6. SHAP Interpretability

TreeSHAP [37] was used to compute exact Shapley values for every prediction. SHAP was applied to the final stacking ensemble retrained on the full dataset after cross-validation hyperparameter tuning, providing global feature importance rankings. Because SHAP is computed in-sample on the full training set, the reported importances may slightly overestimate the contribution of features that correlate with training-set-specific noise. Feature importance was ranked by mean absolute SHAP value; beeswarm and dependence plots were generated to visualize the direction and magnitude of each feature's contribution to the predicted target.

2.7. Feature Selection Analysis

LASSO regression was applied over a logarithmically spaced grid of regularization strengths (α from 10^{-4} to 10^{1}). The coefficient path reveals when each feature enters or exits the model, and the α that minimizes five-fold MAE identifies the most parsimonious feature subset. Because LASSO is a linear method, its coefficient paths identify features that are linearly associated with the target; this linear-model perspective may not fully capture the non-linear importance structure exploited by the tree ensembles. LASSO paths are therefore presented as a complementary regularization-based view rather than a direct surrogate for tree-ensemble importance, and agreement between LASSO and SHAP rankings constitutes internal consistency within the same dataset rather than independent validation.

2.8. Multi-Objective Alloy Optimization

NSGA-II [32] was used to search for optimal alloy compositions in six alloy families: V–Cr–Fe–Mn–Al, V–Ti–Cr, V–Ti–Cr–Fe, V–Ti–Cr–Fe–Al, V–Ti–Cr–Fe–Mn, and V–Ti–Cr–Fe–Mn–Al. The trained retention-rate CP model (Task 16, $R^2 = 0.934$) served as the main surrogate for cyclic capacity retention. The CP model was preferred over the P-type retention model ($R^2 = 0.965$) for this purpose because the NSGA-II loop independently predicts cycle count and BCC phase ratio for each candidate composition using their own surrogate models; by routing these predictions as additional features into the CP model, the optimizer can exploit the correlations between retention and these co-optimized properties in a self-consistent multi-surrogate framework. In a purely prospective setting where neither cycle count nor BCC phase ratio is available, the P-type retention model would be the operationally appropriate choice. The optimization simultaneously sought high predicted cyclic capacity retention, high predicted cycle count, high predicted BCC phase ratio, and a low cost proxy learned from the dataset cost descriptor. It should be noted that cycle count and BCC phase ratio are not direct experimental inputs to the NSGA-II search; they are additional properties predicted by separate trained surrogate models,

so the optimization objectives themselves carry prediction uncertainty. The resulting Pareto candidates therefore represent compositions that are simultaneously predicted to be favorable by multiple surrogate models, not experimentally verified optima.

The NSGA-II population comprised 500 individuals evolved over 180 generations for each alloy family. Composition constraints ensured elemental fractions summed to 100 at.% and remained within the search bounds defined for each family; the element-wise lower and upper bounds for each of the six alloy families are provided in Supplementary Table S3. The final population contains 3000 candidate compositions in total, and the resulting Pareto front was stratified into six layers based on non-domination rank. Representative candidates were then selected for DFT validation.

The applicability domain (AD) of the surrogate models was assessed using a bounding-box criterion in the 17-dimensional elemental composition space: a candidate composition is considered within the primary AD if every elemental fraction falls within the minimum–maximum range observed for that element in the training data. In addition, candidates within 10% of an elemental upper boundary are marked as near-boundary cases because their predictions are more sensitive to local data sparsity. This affects several Mn-rich candidates with Mn contents above 45 at.%, close to the training-set maximum of 50 at.% in Supplementary Table S1. These compositions are not outside the bounding-box AD, but their predicted properties should be treated with elevated uncertainty and are identified by the △ flag in the representative-candidate table in the Results section. A more rigorous AD assessment based on leverage statistics or convex-hull membership is recommended in future work where computational resources permit.

2.9. First-Principles Calculation Details

First-principles calculations were performed using the Vienna ab initio Simulation Package (VASP 6.4) [38,39] with the projector augmented-wave (PAW) method [40] and the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional [41]. Special quasi-random structures (SQS) were constructed for four representative alloys—V–Ti–Cr–Fe, V–Ti–Cr–Fe–Al, V–Ti–Cr–Fe–Mn, and V–Ti–Cr–Fe–Ce—in both BCC and FCC lattices using 54-atom supercells. The first three compositions were selected from the NSGA-II Pareto families; V–Ti–Cr–Fe–Ce was included as a rare-earth-bearing reference to probe the generality of the BCC stability and charge-transfer trends beyond the six optimized families, given that Ce appears in a small fraction (≈3%) of the training dataset. A plane-wave cutoff of 400 eV and a Γ-centered $3 \times 3 \times 3$ k-mesh were adopted. Ionic positions and cell shape were relaxed until forces converged below 0.01 eV Å^{-1}. Convergence tests confirmed that the 400 eV cutoff and the $3 \times 3 \times 3$ k-mesh yield total energies converged to about 1 meV/atom; a $4 \times 4 \times 4$ mesh changes $\Delta E_{\mathrm{BCC-FCC}}$ by <2 meV/atom for the V–Ti–Cr–Fe reference composition. An additional 128-atom SQS supercell was compared with the 54-atom reference for V–Ti–Cr–Fe: $\Delta E_{\mathrm{BCC-FCC}}$ differs by only 0.002 eV/atom between the two supercell sizes, supporting the use of the 54-atom cell for the relative phase-stability comparison. SQS supercells were generated using the ATAT 3.50 `mcsqs` code [42]; the maximum correlation function error for pair and triplet clusters up to the second-neighbor shell was <0.02 for all four alloys, satisfying the standard SQS quality criterion of <0.05. Hydrogen absorption energies were evaluated by placing H atoms at tetrahedral and octahedral interstitial sites; the absorption energy per H atom is defined as

$$E_{\mathrm{abs}} = \frac{1}{N_{\mathrm{H}}}\left[E(\mathrm{alloy} + N_{\mathrm{H}}\,\mathrm{H}) - E(\mathrm{alloy}) - \frac{N_{\mathrm{H}}}{2}\,E(\mathrm{H_2})\right], \quad (1)$$

where $E(\mathrm{alloy} + N_{\mathrm{H}}\,\mathrm{H})$ is the DFT energy of the alloy supercell with N_{H} hydrogen atoms at interstitial sites, $E(\mathrm{alloy})$ is the energy of the empty supercell, and $E(\mathrm{H_2})$ is the DFT energy

of an isolated H_2 molecule; a negative E_{abs} indicates exothermic absorption. Bader charge analysis [43] was used to quantify charge transfer, and the electron localization function (ELF) was computed to visualize bonding character.

3. Results

3.1. Data Overview

Figure 1 shows the distribution of desorption plateau pressure (P_d) across the indexed compounds, with the grey horizontal line marking the median P_d of the plotted samples. The element-wise composition statistics indicate that V appears in virtually every sample at 10–90 at.%, reflecting its role as the primary BCC-stabilizer. Ti (0–30 at.%) and Cr (0–30 at.%) are the next most common alloying additions, followed by Fe, Mn, and Al. Several elements (Mg, Hf, Ta, Y, Ce, Si, Sc) occur in fewer than 5% of samples, which limits model accuracy for compositions rich in these elements. Leave-group-out analysis shows that the mean absolute error for compositions containing these rare elements is 15–30% larger than the overall MAE, indicating that model predictions for rare-element-rich alloys carry elevated uncertainty and should be treated with additional caution. Supplementary Table S1 provides a complete statistical reference for all 17 elements, listing the minimum, median, and maximum atomic percentages observed across the full 901-record dataset; this table also serves as a practical reference for readers wishing to assess whether a target composition falls within the training data envelope (applicability domain).

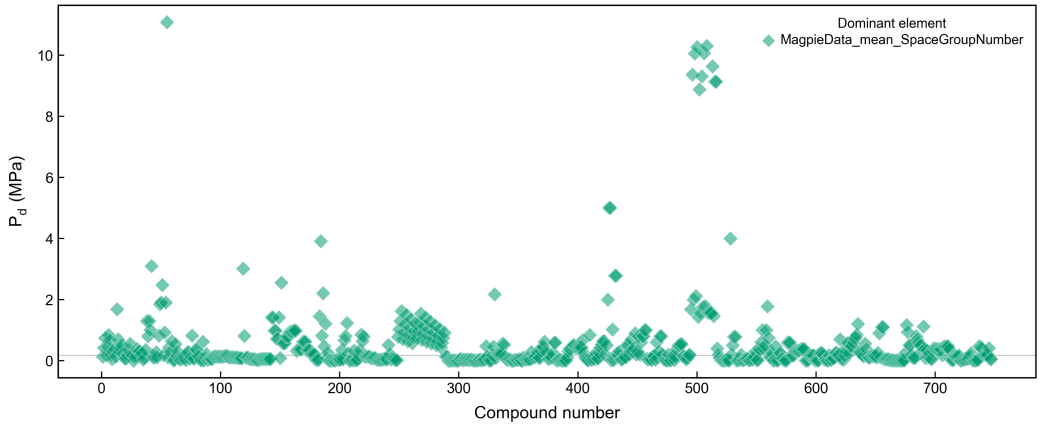


Figure 1. Distribution of desorption plateau pressure (P_d) across the indexed compounds in the dataset. Each marker represents one compound, and the marker style follows the category shown in the legend. The grey horizontal solid line indicates the median P_d of the plotted samples, serving as a visual reference for comparing low- and high-pressure records.

3.2. Correlation Analysis

The correlation matrix (Figure 2) reveals physically expected relationships. Ti content positively correlates with lattice parameter (larger atomic radius), while Cr increases VEC—both well-established trends in BCC alloy metallurgy [3,6]. Moderate inter-feature correlations justify the use of tree-based learners, which partition features hierarchically and are less affected by multicollinearity than linear models.

3.3. Model Performance Comparison

Table 2 summarizes the R^2 values of all 21 stacking models and six baselines. Across all seven target properties, the GA-optimized stacking ensemble achieves the highest R^2 in 19 out of 21 tasks. The two exceptions are Task 13 (C_{max}, C-type), where standalone XGBoost performs best, and Task 15 (C_{max}, CP-type), where Ridge regression is marginally higher. Both tasks use the smallest target-specific subset (95 samples), which reduces the benefit of

model combination and makes the ranking more sensitive to fold-level variation. Fold-level standard deviations of R^2 are reported in Table 3. These results should be interpreted as exploratory evidence of ensemble superiority rather than confirmatory proof, given the small training datasets and the absence of an independent external test set.

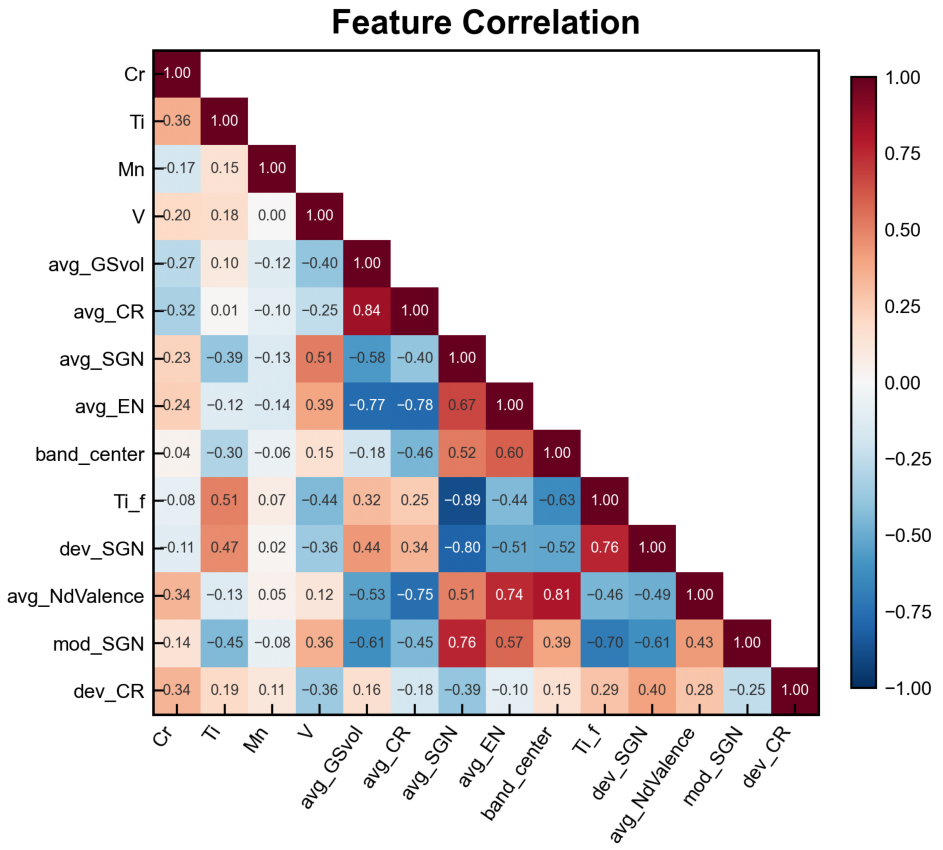


Figure 2. Pearson correlation heatmap for the CP descriptor set (Task 4, desorption plateau pressure). The lower-triangular matrix shows pairwise Pearson correlations among the selected descriptor variables. Strong positive correlations are observed for several descriptor pairs, such as average ground-state volume and average covalent radius, whereas strong negative correlations appear between some size/electronic descriptors and structural descriptors.

Table 3 gives the corresponding MSE, RMSE, and MAE values, so the summary comparison and detailed error metrics can be read together in the main results section.

Several patterns stand out. First, P-type models tend to match or exceed C-type models for P_d (0.981 vs. 0.911), ΔH_{des} (0.910 vs. 0.691), and η (0.965 vs. 0.866), suggesting that physically derived descriptors may capture the underlying hydrogen–metal interactions more efficiently than raw compositions for these targets; the advantage is substantial for P_d and ΔH_{des} but marginal for P_a and C_{max}, where raw compositions perform comparably. Second, the CP fusion strategy provides marginal improvement over the better of C and P for most targets, except for ΔH_{abs} where CP ($R^2 = 0.911$) notably exceeds both C (0.898) and P (0.874). Third, desorption enthalpy is the hardest property to predict from compositions alone ($R^2 = 0.691$), reflecting the complex interplay of lattice strain, electronic structure, and surface kinetics that raw elemental fractions cannot capture.

Figure 3 compares the selected model-performance panels used to visualize the trends summarized in Tables 2 and 3.

Table 2. Summary of R^2 values for the GA-optimized stacking ensemble and six baseline models across 21 prediction tasks. Bold indicates the best model per task. C: composition-only; P: property-only; CP: composition–property fusion.

Target	Type	Stack	XGB	RF	GB	Ada	SVR	Ridge
P_d	C	**0.911**	0.890	0.874	0.860	0.631	0.589	0.399
P_d	CP	**0.973**	0.939	0.891	0.941	0.836	0.803	0.755
P_d	P	**0.981**	0.947	0.926	0.967	0.853	0.831	0.659
C_{eff}	C	**0.810**	0.777	0.789	0.735	0.369	0.593	0.253
C_{eff}	CP	**0.791**	0.775	0.751	0.705	0.481	0.416	−0.089
C_{eff}	P	**0.865**	0.809	0.757	0.733	0.369	0.362	0.201
P_a	C	**0.949**	0.946	0.943	0.942	0.726	0.860	0.429
P_a	CP	**0.947**	0.924	0.917	0.924	0.770	0.840	0.682
P_a	P	**0.952**	0.920	0.931	0.938	0.818	0.820	0.723
ΔH_{des}	C	**0.691**	0.497	0.412	0.385	0.200	−0.069	0.038
ΔH_{des}	P	**0.910**	0.791	0.797	0.790	0.502	0.143	−0.513
ΔH_{des}	CP	**0.921**	0.835	0.838	0.875	0.348	0.314	0.258
C_{max}	C	0.863	**0.886**	0.819	0.843	0.718	0.599	0.018
C_{max}	P	**0.866**	0.854	0.863	0.865	0.764	0.794	0.848
C_{max}	CP	0.864	0.845	0.864	0.833	0.740	0.766	**0.869**
ΔH_{abs}	C	**0.898**	0.858	0.804	0.632	0.440	0.141	0.102
ΔH_{abs}	P	**0.874**	0.851	0.799	0.743	0.524	0.483	0.175
ΔH_{abs}	CP	**0.911**	0.834	0.845	0.826	0.324	0.253	0.194
η	C	**0.866**	0.464	0.839	0.674	−0.056	−0.266	0.419
η	P	**0.965**	0.960	0.919	0.932	0.788	0.660	0.743
η	CP	**0.934**	0.858	0.891	0.914	0.718	0.713	0.137

Table 3. Detailed evaluation metrics of the GA-optimized stacking ensemble for all 21 prediction tasks. $\sigma(R^2)$ denotes the standard deviation of R^2 across the five cross-validation folds; larger values indicate higher fold-to-fold variability, particularly for small-subset tasks (Tasks 13–15, $n = 95$).

Task	Target	Type	R^2	σ (R^2)	MSE	RMSE	MAE
1	P_d	C	0.911	0.021	0.00193	0.0439	0.0299
2	C_{eff}	C	0.810	0.028	0.0563	0.237	0.168
3	P_a	C	0.949	0.014	0.00951	0.0975	0.0688
4	P_d	CP	0.973	0.009	0.000999	0.0316	0.0192
5	C_{eff}	CP	0.791	0.034	0.0589	0.243	0.167
6	P_a	CP	0.947	0.012	0.0102	0.101	0.0653
7	P_d	P	0.981	0.007	0.000836	0.0289	0.0173
8	C_{eff}	P	0.865	0.023	0.0300	0.173	0.119
9	P_a	P	0.952	0.013	0.00919	0.0959	0.0667
10	ΔH_{des}	C	0.691	0.047	31.91	5.649	2.947
11	ΔH_{des}	P	0.910	0.021	10.25	3.202	1.379
12	ΔH_{des}	CP	0.921	0.018	9.463	3.076	1.306
13	C_{max}	C	0.863	0.052	0.113	0.336	0.287
14	C_{max}	P	0.866	0.041	0.121	0.348	0.235
15	C_{max}	CP	0.864	0.047	0.122	0.350	0.238
16	η	CP	0.934	0.017	0.00107	0.0328	0.0130
17	ΔH_{abs}	C	0.898	0.026	8.790	2.965	1.524
18	ΔH_{abs}	P	0.874	0.031	13.10	3.619	1.186
19	ΔH_{abs}	CP	0.911	0.019	7.575	2.752	1.126
20	η	C	0.866	0.029	0.000610	0.0247	0.0139
21	η	P	0.965	0.011	0.000581	0.0241	0.00936

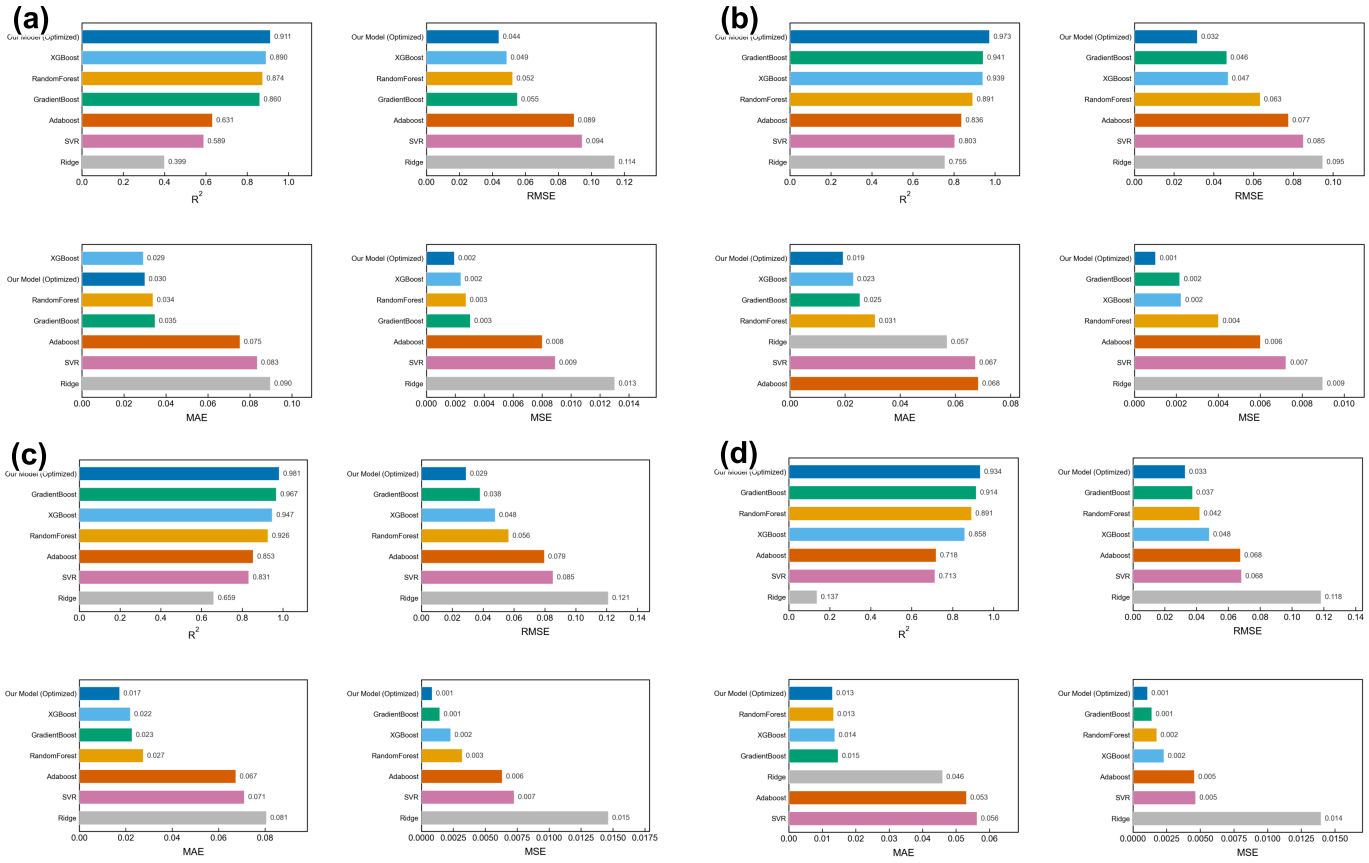


Figure 3. Model comparison panels for selected prediction tasks. Bar charts compare R^2, RMSE, and MAE across seven algorithms. (**a**) Desorption plateau pressure with composition descriptors. (**b**) Same target with CP fusion descriptors. (**c**) Same target with property descriptors—the highest $R^2 = 0.981$. (**d**) Cyclic retention rate with CP fusion, $R^2 = 0.934$.

3.4. Prediction Accuracy

Figure 4 displays parity plots for four representative high-performing models. The P_d P-model (Figure 4a) shows tight clustering around the diagonal across three orders of magnitude, with a test RMSE of 0.029 (log-transformed scale). The CP retention model used for optimization (Figure 4d) achieves an RMSE of 0.033, corresponding to ∼3.3 percentage points in the original retention scale—a precision level useful for screening candidate alloys.

Residual scatter plots (Pred−True vs. True, Supplementary Figure S1) and residual distribution histograms (Supplementary Figure S2) confirm near-zero prediction bias for all four representative models (bias < 0.2 in all target units), with no systematic directional trend across the observed value range. Departures from strict residual normality (Shapiro–Wilk $p < 0.05$ for P_d-C, ΔH_{des}-CP, and η-CP) are attributable to identifiable domain properties: the inverse-log-transform (right-skew in MPa scale), the bounded [0, 1] retention domain (ceiling effect), and low-enthalpy outliers, respectively; the C_{max}-C model ($n_{test} = 15$) is consistent with normality ($p = 0.46$). The RMSE of 3.3 percentage points for the η-CP model is sufficient for relative candidate ranking, but should not be interpreted as a guaranteed absolute accuracy; candidates predicted at $\eta = 100\%$ carry an uncertainty band of ±3.3 pp spanning a practically significant range.

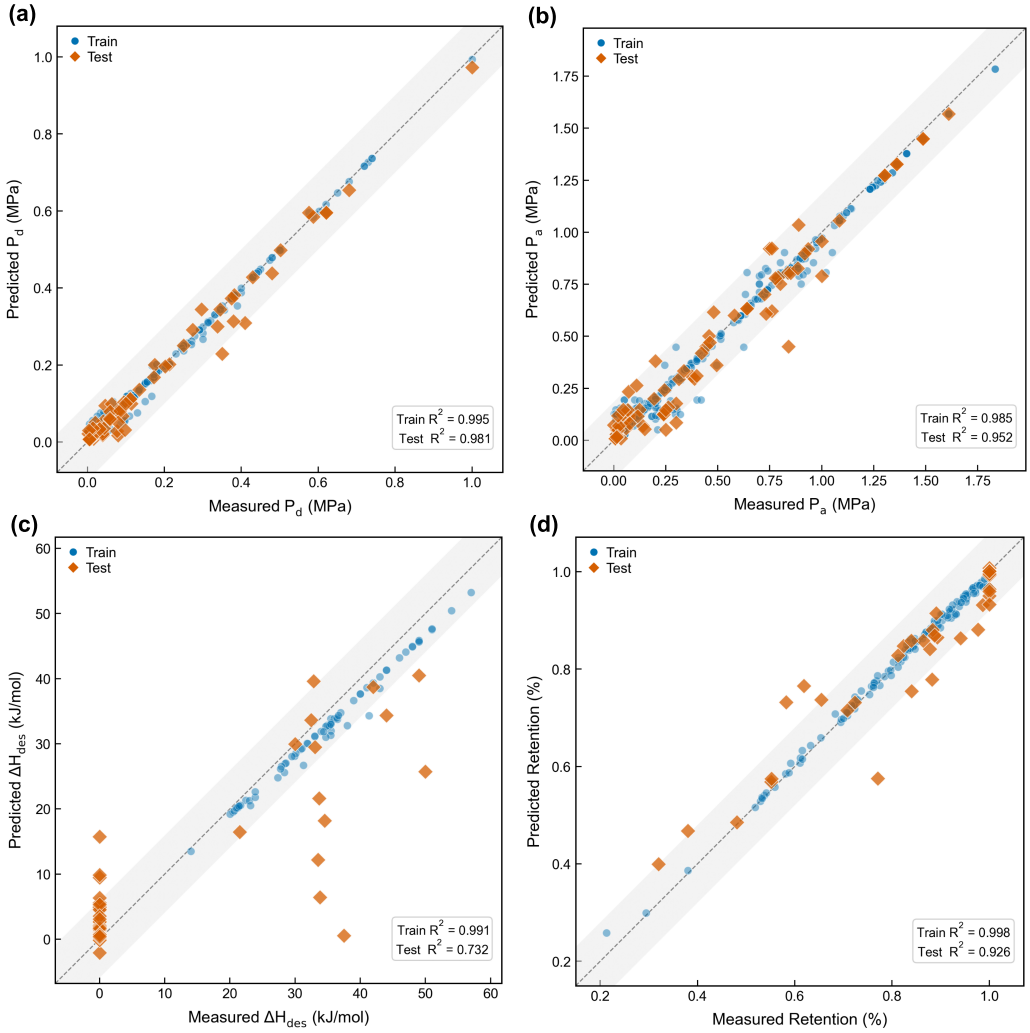


Figure 4. Predicted versus experimental values for four representative high-performing models. Dashed lines indicate perfect prediction. (**a**) Desorption plateau pressure—property model. (**b**) Absorption plateau pressure—property model. (**c**) Desorption enthalpy—fusion model. (**d**) Cyclic retention rate—fusion model used for subsequent alloy optimization. Training and test sets are distinguished by color.

3.5. Feature Importance and SHAP Analysis

SHAP analysis (Figure 5) exposes the physical underpinnings of each model. It should be emphasized that SHAP values quantify the contribution of each feature to the model's output; they reflect the learned statistical associations captured by the ensemble and do not establish causal mechanisms. The following physical interpretations are offered as consistent qualitative explanations, not as model-independent causal claims. In the P-type model for P_d (Figure 5b), VEC is the most influential descriptor: higher VEC pushes desorption pressure upward, consistent with reports that electronic structure and lattice expansion jointly tune hydride stability in BCC and high-entropy alloy hydrides [10,44,45]. Atomic size mismatch (δ) ranks second; larger δ (indicative of severe lattice distortion) tends to lower plateau pressure, as the expanded interstitial sites can accommodate hydrogen at lower chemical potential.

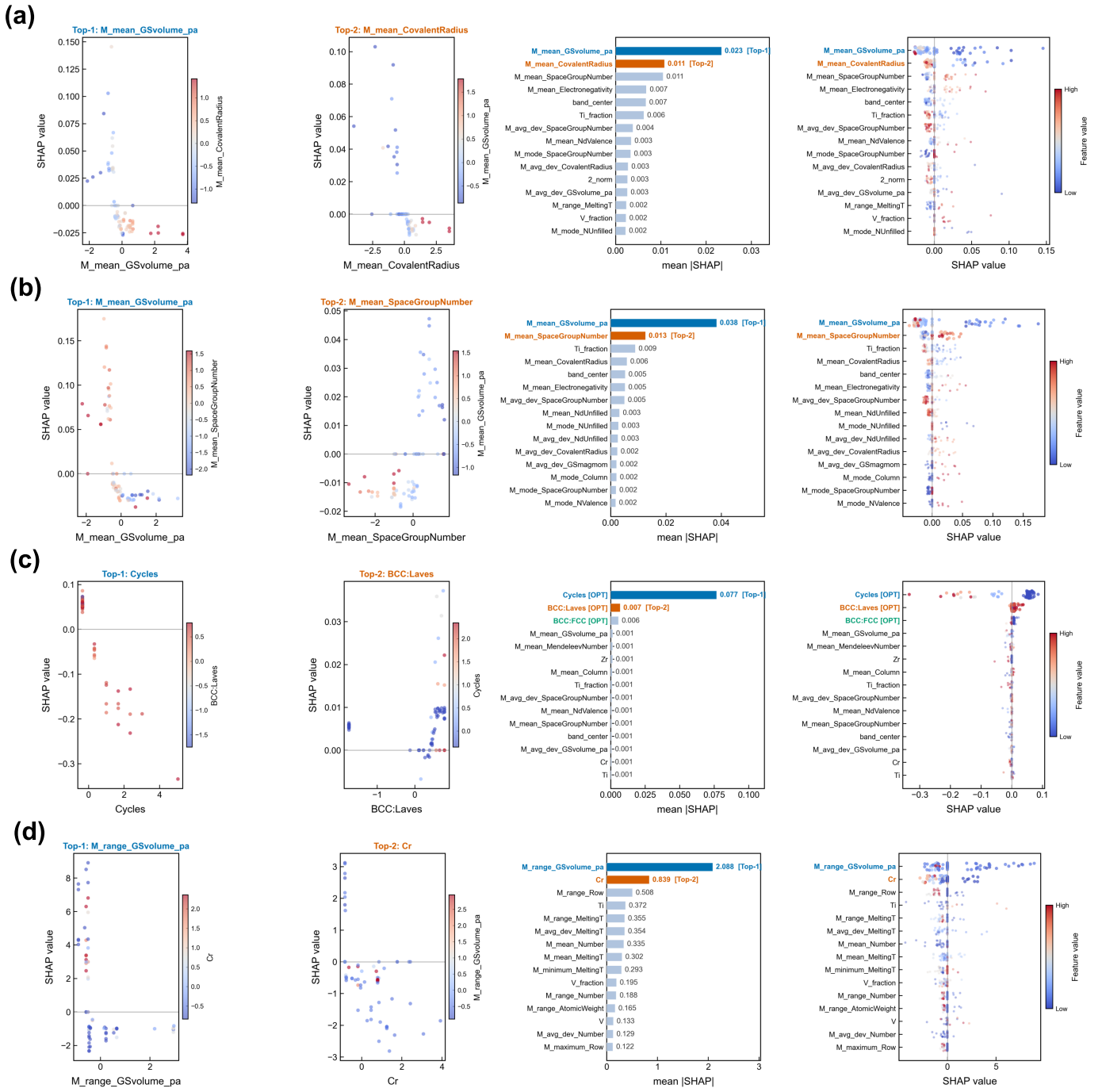


Figure 5. SHAP summary panels for four representative models. In each row, the first two panels are SHAP dependence plots for the two leading features, the third panel is the mean |SHAP| bar ranking, and the fourth panel is the beeswarm summary plot. The gray reference line marks SHAP = 0: it is horizontal in the dependence plots and vertical in the beeswarm plots. Point color encodes the corresponding feature value. (**a**) P_d fusion model—Ti, VEC, and δ dominate. (**b**) P_d property model—VEC alone accounts for >40% of importance. (**c**) Retention model—cycle count and BCC phase ratio are the top features. (**d**) Absorption enthalpy fusion model—Ti and VEC are dominant.

For the retention model (Figure 5c), cycle count is the dominant feature—an expected but nonetheless reassuring result, since retention degrades with cycling. BCC phase ratio ranks second: alloys with a higher BCC fraction retain capacity better, likely because secondary Laves phases act as irreversible hydrogen traps [3,6]. From a practical prediction standpoint, the dominance of cycle count as a feature is methodologically important: the model predicts retention most accurately when cycle count is known, but cycle count is

itself an outcome of the cycling test one wishes to avoid. In a true prospective screening scenario where neither cycle count nor BCC phase ratio has been measured, one would rely on a composition-only (C-type) or property-only (P-type) retention model, which achieves $R^2 = 0.866$ and 0.965, respectively; the CP model ($R^2 = 0.934$) used for NSGA-II optimization thus represents an upper bound on predictive accuracy under the most favorable informational conditions.

In C-type models, Ti content is consistently the top feature for P_d and P_a. The SHAP dependence trends reveal a monotonic negative relationship: increasing Ti from 0 to 25 at.% reduces predicted P_d by approximately one order of magnitude, aligning with experimental PCI trends and recent reviews of V–Ti–Cr-based alloys [3,4,8].

Figure 6 summarizes the permutation-importance distributions for the same representative tasks.

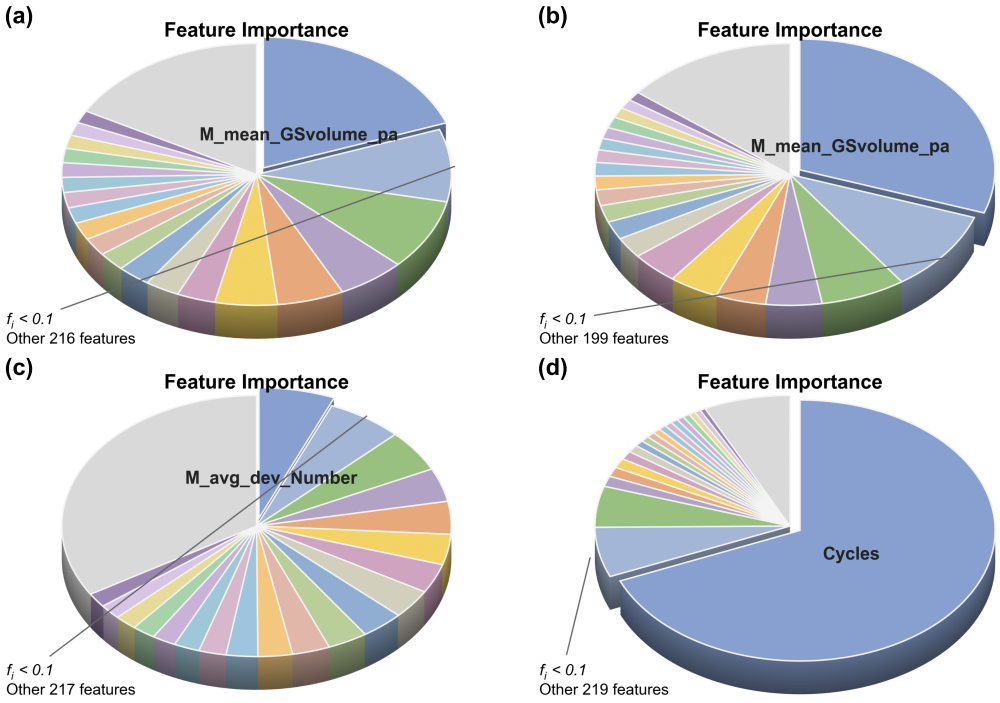


Figure 6. Feature importance pie charts derived from permutation importance of the stacking ensemble. Different colors distinguish individual features, and the grey grouped sector labeled $f_i < 0.1$ collects low-importance features whose individual importance values are below 0.1; the annotation reports the number of aggregated features in this group. The labeled exploded sector highlights the dominant feature in each representative model. (**a**) P_d CP model. (**b**) P_d P model. (**c**) C_{eff} P model. (**d**) η CP model.

3.6. LASSO Feature Selection

LASSO paths (Figure 7) provide a complementary view: features that survive the heaviest penalization carry the most predictive signal. For the P_d CP model, Ti content and VEC are the last two features to be driven to zero, consistent with SHAP rankings. This convergence across two unrelated interpretability methods strengthens confidence in the physical relevance of these descriptors. However, it should be noted that both SHAP and LASSO are applied to the same dataset, so their agreement constitutes internal consistency rather than independent validation. Furthermore, LASSO is a linear method; its coefficient paths identify features that are linearly associated with the target and may not fully capture the non-linear importance structure exploited by the tree ensembles.

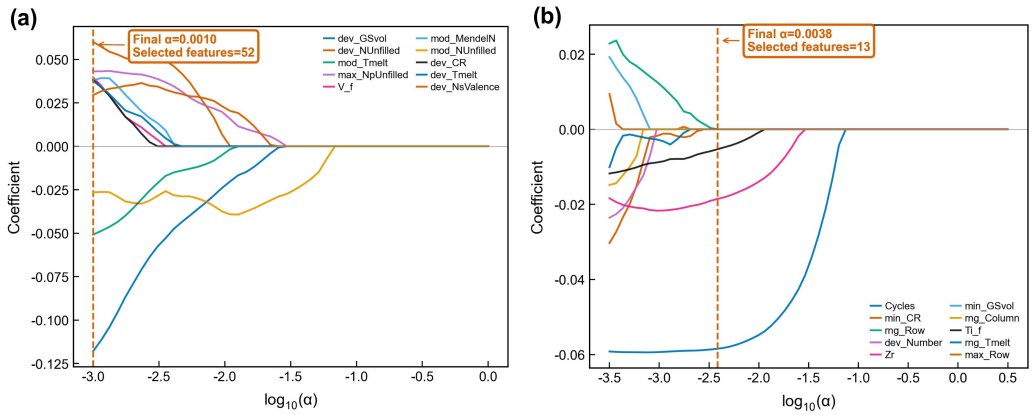


Figure 7. LASSO coefficient paths as a function of regularization strength α. (**a**) For the P_d CP model, Ti and VEC are the last features to shrink to zero, confirming their dominance. (**b**) For the η CP model, cycle count and BCC phase ratio persist at the highest α values.

3.7. Feature–Target Relationships

Figure 8 shows how the dominant features vary with the corresponding target properties before the detailed captioned figure is presented.

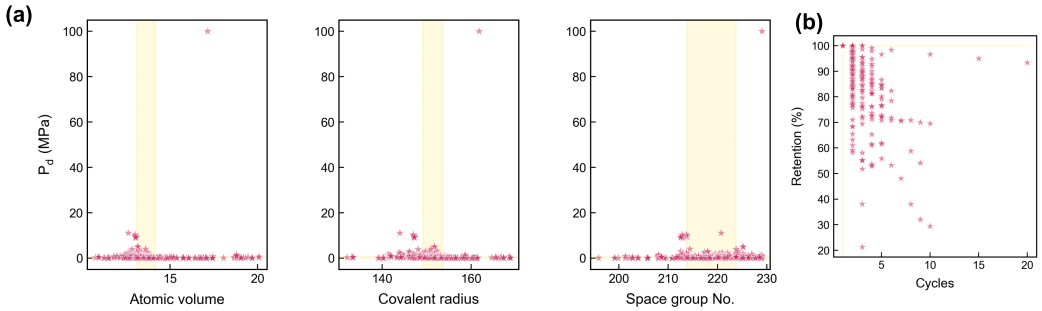


Figure 8. Feature–target scatter plots for (**a**) the top features of the P_d fusion model and (**b**) the dominant feature of the retention model. Yellow shading marks the interquartile range (25th–75th percentile) of the corresponding feature or target values, highlighting the central distribution of the plotted samples. Non-linear trends are visible in both cases, justifying the use of ensemble tree models over linear regression.

3.8. Ternary Composition Prediction Maps

Figure 9 presents ternary composition prediction maps for four alloy families. In the V–Ti–Cr system (Figure 9a), the highest predicted capacity occupies the V-rich corner with moderate Ti and low Cr, while the highest retention appears at balanced V–Ti compositions. The V–Ti–Cr–Fe–Mn system (Figure 9c) shows a broad high-retention region at Ti $\approx$ 5–15 at.%, V $\approx$ 20–35 at.%, suggesting considerable compositional flexibility for achieving >95% cyclic retention.

3.9. Multi-Objective Optimization

Table 4 summarizes representative candidates selected from the Pareto-ranked population. The V–Ti–Cr–Fe–Mn system produces the largest number of solutions with retention > 99%, including several candidates with very low values of the model-estimated cost proxy. The Ti-free V–Cr–Fe–Mn–Al system reaches similar predicted retention in low-cost regions of the search space, though its predicted BCC phase ratio is sometimes lower. The V–Ti–Cr–Fe–Al system yields a candidate with retention rounded to 100.0% ($Ti_{9.2}V_{48.2}Cr_{26.9}Fe_{0.1}Al_{15.6}$), but its cost proxy is higher than that of the Mn-rich candidates. Predicted retention values of 100.0% represent the upper saturation of the model's output range and should be interpreted with caution. All predicted η values carry a cross-

validation uncertainty of ±6.6 percentage points (±2 × $RMSE_{\eta\text{-}CP}$, Task 16); the lower 95% bound for all 100.0% predictions is therefore ≈93.4%. The applicability-domain (AD) flag in Table 4 uses a bounding-box criterion: compositions where Mn exceeds 45 at.% are flagged (△) as near-boundary cases within 10% of the training-data upper boundary (max Mn = 50 at.% in Supplementary Table S1), and their predicted properties carry elevated uncertainty. All listed candidates remain within the min–max elemental bounds of the training data. The estimated raw-material costs (USD kg^{-1}) are computed from 2022–2024 average spot prices after converting atomic fractions to mass fractions and are intended as relative raw-material cost indicators; they do not include processing, synthesis, or market-availability factors.

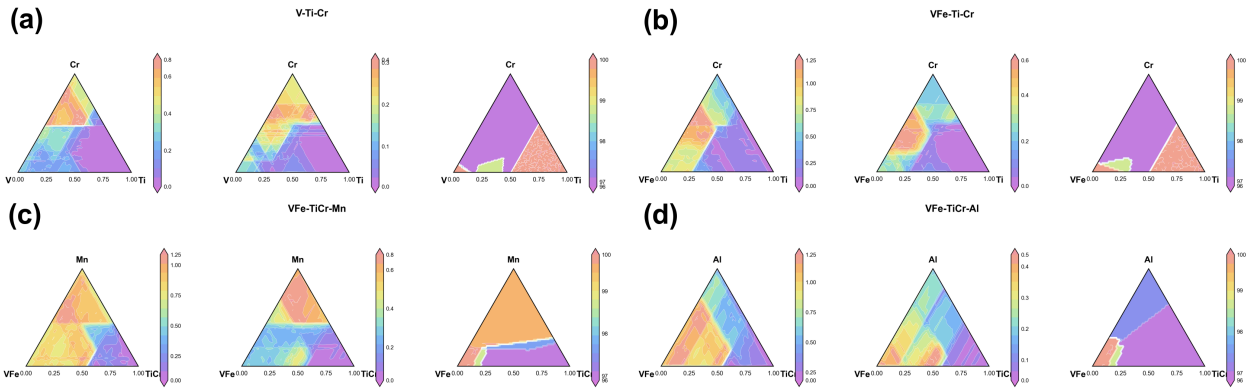


Figure 9. Ternary composition maps of predicted hydrogen storage properties for four alloy families. Each panel projects the multi-element composition onto a pseudo-ternary diagram by fixing the minority-element contents at their median values in the training data and varying only the three principal elements subject to the normalization constraint; predicted property values therefore represent alloys with those fixed minor-element contents, and the variation in property with minor-element composition is not captured in the projection. Color encodes the predicted target value. These maps enable rapid identification of promising composition regions without additional experiments. (**a**) V–Ti–Cr system. (**b**) V–Ti–Cr–Fe system (Fe fixed at median). (**c**) V–Ti–Cr–Fe–Mn system (Fe and Mn fixed at median). (**d**) V–Ti–Cr–Fe–Al system (Fe and Al fixed at median).

Figure 10 visualizes the NSGA-II optimization landscape and the final Pareto-layer ranking.

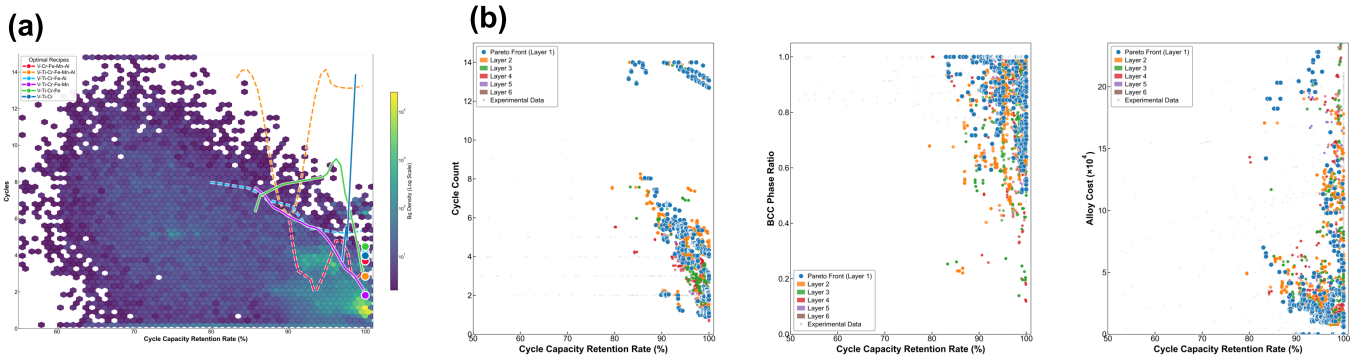


Figure 10. NSGA-II multi-objective optimization results. (**a**) Optimization landscape in retention–cycle-count space. The color bar denotes the logarithmic background density of candidate solutions, and the "Optimal Recipes" legend identifies the alloy-system-specific Pareto curves and highlighted best points. (**b**) Final Pareto-layer projections ranked by non-domination in retention–cycle-count, retention–BCC phase ratio, and retention–alloy-cost spaces. The legend labels Layer 1 as the Pareto front, Layers 2–6 as successive dominated layers, and the grey points as experimental data used as the reference background.

3.10. DFT Validation

Table 5 compiles the DFT results. All four alloys are energetically more stable in BCC than in FCC, with $\Delta E_{BCC-FCC} = -0.084$ to -0.132 eV/atom. This agrees with experimental reports that V- and V–Ti-based solid-solution alloys commonly crystallize in the A2 (BCC) structure [3,6]. The BCC SQS supercell lattice parameters fall in the range 8.94–8.99 Å; after conversion to the corresponding two-atom BCC cell scale, they are consistent with reported V-based BCC lattice dimensions.

Table 4. Representative optimized alloy compositions from NSGA-II screening. Compositions are in at.%. η: predicted retention rate (%); n: predicted cycle count (cycles); BCC: predicted BCC phase ratio; Cost: mass-fraction-weighted average raw-material price $C_{alloy} = \sum_i w_i p_i$ (USD kg^{-1}) using the 2022–2024 spot prices from Supplementary Table S2. All predicted η values carry a cross-validation uncertainty of ±6.6 percentage points (95% interval: ±2 × RMSE, Task 16). AD: applicability-domain flag (✓: all elemental fractions within training-data min–max range; △: one or more elements within 10% of the training-data boundary).

System	Composition	η (%)	n	BCC	Cost (USD kg^{-1})	AD
V–Cr–Fe–Mn–Al	$V_{33.5}Cr_{11.5}Fe_{7.0}Mn_{46.7}Al_{1.2}$	100.0	108	0.91	11.5	△
V–Cr–Fe–Mn–Al	$V_{50.1}Cr_{27.6}Fe_{2.1}Mn_{16.8}Al_{3.4}$	100.0	399	0.91	17.8	✓
V–Ti–Cr–Fe–Mn	$Ti_{10.0}V_{27.3}Cr_{10.4}Fe_{6.8}Mn_{45.4}$	100.0	93	0.76	10.6	△
V–Ti–Cr–Fe	$Ti_{0.4}V_{71.8}Cr_{26.9}Fe_{1.0}$	99.8	425	0.83	23.7	✓
V–Ti–Cr–Fe–Al	$Ti_{9.2}V_{48.2}Cr_{26.9}Fe_{0.1}Al_{15.6}$	100.0	418	0.64	19.5	✓
V–Ti–Cr–Fe–Mn–Al	$Ti_{0.5}V_{62.9}Cr_{12.8}Fe_{2.1}Mn_{16.6}Al_{5.1}$	100.0	285	1.00	20.4	✓

Table 5. DFT-calculated properties of four representative alloys. E/atom: total energy per atom; a: SQS supercell lattice parameter; μ/atom: magnetic moment per atom. BCC is energetically preferred in all cases.

Alloy	Phase	E/Atom (eV)	a (Å)	μ/Atom (μ_B)	ΔE BCC–FCC
V–Ti–Cr–Fe	BCC	−8.911	8.944	0.000	−0.118
	FCC	−8.793	9.041	0.000	
V–Ti–Cr–Fe–Al	BCC	−8.824	8.961	0.000	−0.096
	FCC	−8.728	9.067	0.000	
V–Ti–Cr–Fe–Mn	BCC	−8.927	8.937	0.000	−0.132
	FCC	−8.795	9.052	0.000	
V–Ti–Cr–Fe–Ce	BCC	−8.836	8.988	0.000	−0.084
	FCC	−8.752	9.086	0.000	

Bader charge analysis (Table 6) reveals that Ti consistently donates electrons (+0.685 to +0.762 e/atom in the BCC structures), while Fe and Cr act as electron acceptors. The Fe charge transfer is about −0.97 to −1.05 e/atom; the corresponding Cr range is −0.45 to −0.51 e/atom. Separate Bader analysis of the hydrogenated FCC structures gives an average H charge gain of about 0.63 e, indicating partially ionic M–H bonding. The similar charge-transfer pattern across the four alloys suggests that the H–metal interaction is governed mainly by the transition-metal framework, with alloying elements tuning rather than replacing this bonding mechanism.

Table 6. Bader charge transfer (electrons) for selected elements in BCC structures. Positive values indicate electron donation.

Element	V–Ti–Cr–Fe	V–Ti–Cr–Fe–Al	V–Ti–Cr–Fe–Mn	V–Ti–Cr–Fe–Ce
Ti	+0.735	+0.685	+0.755	+0.762
V	+0.014	+0.022	+0.032	+0.022
Cr	−0.447	−0.503	−0.476	−0.508
Fe	−1.046	−0.966	−1.003	−0.998

The ELF maps (Figure 11b) show low electron localization (ELF < 0.3) in the interstitial regions of BCC lattices, indicating metallic bonding with delocalized electrons. The charge density difference maps (Figure 11c) for hydrogenated structures reveal pronounced charge accumulation around H atoms and depletion around neighboring metal atoms, confirming that hydrogen incorporation involves significant charge redistribution.

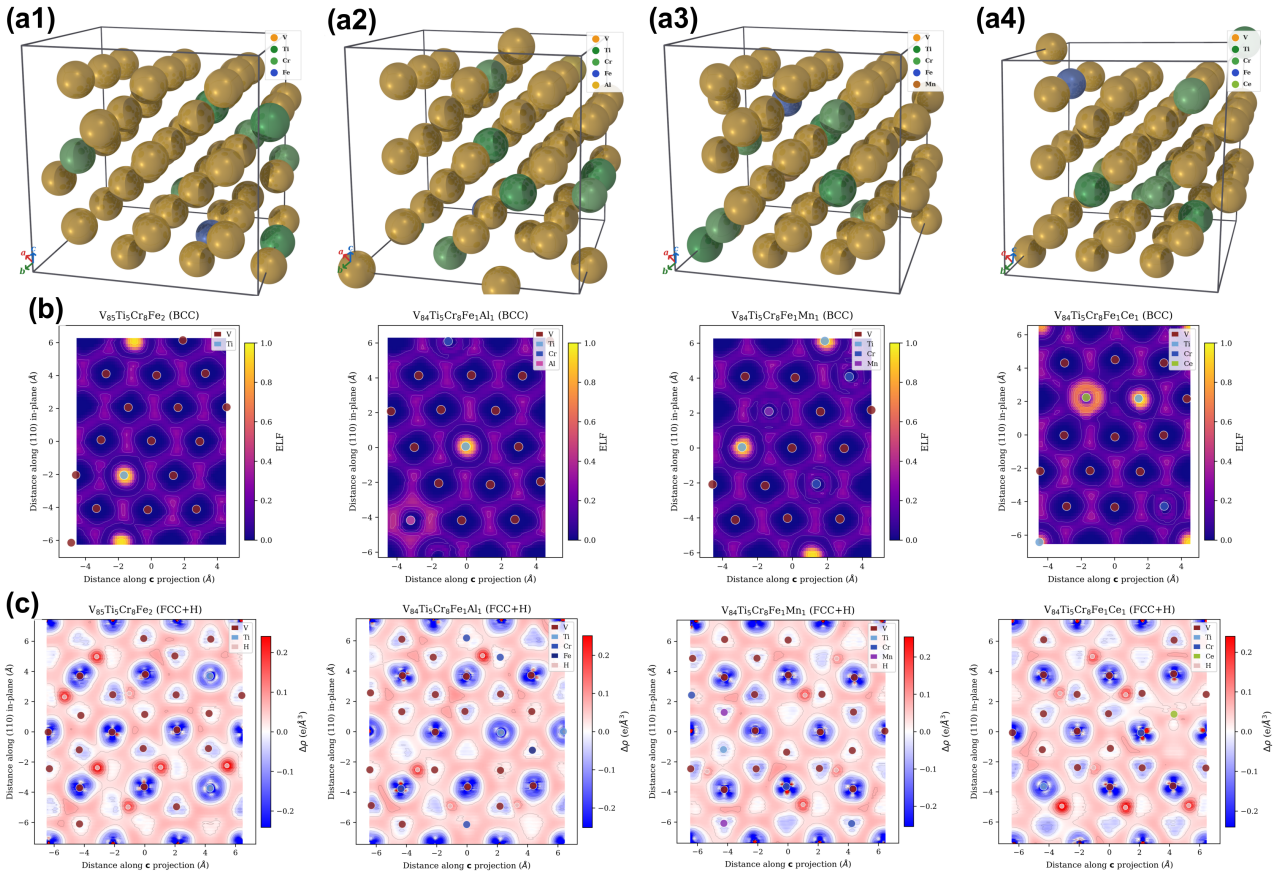


Figure 11. DFT results for four representative alloys: V–Ti–Cr–Fe, V–Ti–Cr–Fe–Al, V–Ti–Cr–Fe–Mn, and V–Ti–Cr–Fe–Ce (left to right). (**a1**–**a4**) Relaxed BCC supercell structures for the four alloys listed from left to right. (**b**) Electron localization function (ELF) maps on the (110) plane of BCC lattices. (**c**) Charge density difference maps for FCC structures after H absorption, showing charge accumulation (red) around H sites.

Although the NSGA-II optimization primarily targeted macroscopic cyclic performance metrics—cyclic capacity retention (η), cycle count (n), BCC phase ratio, and cost proxy—the DFT results provide indirect but physically coherent support for these macroscopic targets through two mechanistic links. First, the calculated BCC–FCC energy difference ($\Delta E_{\mathrm{BCC-FCC}} \approx -0.08$ to -0.13 eV/atom) indicates that all four representative alloys prefer the BCC structure under the present 0 K calculation settings. This is directly relevant to the optimization objective of maximizing BCC phase ratio: a high BCC fraction is associ-

ated experimentally with suppressed formation of irreversible Laves-phase hydrogen traps, which is one primary mechanism by which BCC-dominant alloys maintain high cyclic retention [3,6]. The DFT-confirmed BCC preference thus provides a first-principles rationale for why the model associates high BCC ratio with high retention. Second, the charge accumulation around H interstitial sites (average Bader gain $\approx$ 0.63 e per H) indicates moderately ionic M–H bonding, which is characteristic of reversible hydride formation rather than irreversible phase decomposition. Reversible H uptake and release is a prerequisite for maintaining capacity over repeated absorption–desorption cycles. In this sense, the charge-redistribution pattern observed by DFT is qualitatively consistent with the expectation of good cyclic stability for these alloys. These connections are offered as physically motivated interpretations; a quantitative link between DFT energetics and experimentally measured retention rates would require explicit computation of hydrogen diffusion barriers, finite-temperature thermodynamics, and surface kinetics, which are beyond the scope of the present study and are identified as priorities for future work.

4. Discussion

4.1. Why Property Descriptors Outperform Compositions

The general tendency of P-type models to match or outperform C-type models across most targets has a physical explanation. Raw compositions treat each element as an independent variable, ignoring the collective electronic and structural effects of multi-element mixing. Property descriptors like VEC, δ, and $\Delta\chi$ compress this high-dimensional composition space into a low-dimensional physics-informed representation. VEC, for instance, is a weighted average of all constituent elements' valence electrons, directly related to the d-band filling that governs metal–hydrogen bond strength [44,45]. The atomic size mismatch δ encodes the lattice strain energy associated with chemical disorder—a key factor controlling interstitial site volumes and, by extension, hydrogen solubility.

This finding is consistent with reports from Witman et al. [21], who found that physics-based descriptors improved ML predictions for intermetallic hydrides, and with recent hydride ML studies that expanded composition-based thermodynamic datasets and inverse-design workflows [23,24].

However, C-type models still have value. They require no auxiliary calculations, making them applicable to arbitrary compositions without knowledge of the underlying physics. For absorption plateau pressure (P_a), the C-type model already reaches $R^2 = 0.949$—nearly as good as the P-type model (0.952)—suggesting that the composition–property mapping for P_a is relatively direct and does not benefit much from explicit physics encoding.

4.2. The Role of Stacking

The stacking ensemble outperforms every individual learner in 19 of 21 tasks. The improvement varies: for hard targets like ΔH_{des} (C-type), the stack lifts R^2 from 0.497 (XGBoost alone) to 0.691—a 39% relative improvement. For already-accurate targets like P_a (C-type), the gain is only 0.3% (R^2 from 0.946 to 0.949). The stacking meta-learner effectively downweights the least reliable base learner for each sample, reducing variance without increasing bias.

The two exceptions deserve comment. For C_{max} (Task 13 and 15), the dataset has a relatively narrow dynamic range (0–3.9 wt.%) and only 95 usable records in the current target-specific subset. In this regime, the overhead of model combination may slightly overfit, and a single well-tuned learner—or even Ridge regression—can be sufficient. This result is a reminder that model complexity should match data volume and target variability.

4.3. Alloy Design Implications

The NSGA-II results point to the V–Cr–Fe–Mn–Al system (without Ti) as an underexplored design space. The optimized compositions, e.g., $V_{33.5}Cr_{11.5}Fe_{7.0}Mn_{46.7}Al_{1.2}$, combine near-100% predicted retention with a low estimated raw-material cost of approximately 11.5 USD kg^{-1}. High Mn content appears repeatedly in these candidates, consistent with the low-cost search objective ($p_{Mn} = 1.7$ USD kg^{-1}, among the lowest of all elements in the dataset) and the positive contribution of Mn-containing compositions in the retention model. These predictions should be treated strictly as model-based screening hypotheses. The proposed compositions have not been experimentally synthesized or tested; their actual phase composition, PCI curves, activation behavior, and long-term cyclic stability remain unknown. Several Mn-rich candidates (Mn $>$ 45 at.%) are flagged with an elevated applicability-domain warning (△ in Table 4), as they lie within 10% of the training-data upper boundary for Mn, and their predicted properties carry elevated near-boundary uncertainty.

For applications requiring both high capacity and long cycle life, such as Ni/MH rechargeable batteries [46], the V–Ti–Cr–Fe–Mn system offers the most balanced region in the current optimization. NSGA-II identified multiple Pareto-optimal compositions with retention $>$ 99%, predicted cycle counts on the order of 100–500 cycles, and relatively low estimated raw-material costs (approximately 11–18 USD kg^{-1}). The DFT results for V–Ti–Cr–Fe–Mn confirm BCC stability ($\Delta E_{BCC-FCC} = -0.132$ eV/atom) and a consistent Ti-to-Fe charge-transfer trend.

4.4. Limitations and Uncertainties

Several caveats apply. First, the dataset was compiled from 88 publications with varying experimental conditions (temperature range, pressure measurement technique, sample preparation method). Measurement noise and systematic biases between laboratories are absorbed into model error but are not explicitly modeled. The cross-validation metrics reported in this work therefore represent a lower bound on the contribution of inter-laboratory systematic error; the true generalization uncertainty on independently synthesized alloys from a different research group may be larger than the reported RMSE values suggest. Second, the property descriptors assume ideal solid solutions; actual alloys may contain secondary phases (Laves, C14, C15) that alter hydrogen storage behavior. The retention model partly compensates by including the BCC phase ratio as a feature, but this ratio is itself a predicted quantity subject to error. Third, the NSGA-II compositions lie outside the training data envelope for some elements (e.g., Mn $>$ 46 at.%), and extrapolation predictions carry higher uncertainty. Fourth, no confidence intervals or prediction uncertainty bounds are provided for individual NSGA-II candidates; future work incorporating conformal prediction or bootstrap uncertainty quantification would enable more reliable extrapolation risk assessment. Fifth, the performance comparison between the stacking ensemble and the six baseline algorithms carries two compounding caveats that should be considered jointly: (a) no independent external test set was used—all metrics are cross-validation estimates that may overestimate generalization; and (b) the baselines used default scikit-learn hyperparameters whereas the stacking pipeline was GA-tuned. The reported performance gap therefore reflects the combined benefit of the ensemble architecture and optimized hyperparameters; the isolated contribution of the stacking architecture over a comparably tuned single learner, evaluated on genuinely unseen data, remains to be established. Experimental synthesis and testing of the top candidates are needed to close the loop.

5. Conclusions

This study developed and validated a GA-optimized stacking ensemble framework for multi-property prediction of BCC-type V-based hydrogen storage alloys. Three descriptor strategies—composition-only (C), property-only (P), and composition–property fusion (CP)—were compared across seven target properties and 21 prediction tasks. Property-based descriptors such as VEC, δ, $\Delta\chi$, and ΔH_{mix} outperform raw compositions for targets governed by electronic structure; the desorption plateau pressure P-model reaches $R^2 = 0.981$, compared with 0.911 for the C-model. The GA-optimized stacking ensemble achieves the highest R^2 in 19 of 21 tasks, with the clearest benefit for targets with broad dynamic ranges and non-linear composition dependence.

The interpretation results are consistent with the metallurgical picture. SHAP analysis identifies VEC and Ti content as the dominant features for plateau pressure, while cycle count and BCC phase ratio control retention. These SHAP-based interpretations reflect learned statistical associations within the ensemble models and should be understood as model-level explanations rather than experimentally confirmed causal relationships. In the design stage, NSGA-II ranks 3000 final candidate compositions across six alloy families and points to two useful directions: the Ti-free V–Cr–Fe–Mn–Al system, which reaches predicted retention above 99% with low cost-proxy values, and the V–Ti–Cr–Fe–Mn system, which offers a better balance between retention and cycling endurance. DFT calculations further support these candidates by confirming the energetic preference for BCC structures ($\Delta E_{\mathrm{BCC-FCC}} \approx -0.08$ to -0.13 eV/atom), consistent Ti→Fe charge transfer, and charge accumulation around H interstitial sites. It must be emphasized that all proposed alloy compositions are model-based screening candidates: no experimental synthesis, phase characterization, PCI measurement, or cyclic stability testing has been performed on any of the NSGA-II outputs, and the DFT validation covers only four representative alloys. The reported R^2 values are cross-validation estimates; without an independent external test set, the true generalization accuracy on new experimental alloys cannot be fully established. These caveats do not diminish the methodological contribution—a unified multi-property, multi-descriptor ensemble framework for BCC V-based alloys—but they indicate that experimental follow-up is essential before any candidate can be recommended for application.

Future work will focus on experimental synthesis of the top NSGA-II candidates, extension to high-entropy alloys with five or more principal elements, and incorporation of temperature- and pressure-dependent thermodynamic models to enable full PCI curve prediction.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/coatings16070794/s1. Figure S1: residual scatter plots for representative models; Figure S2: residual distribution histograms; Table S1: composition statistics for the dataset; Table S2: elemental price factors and molar masses; Table S3: compositional bounds for NSGA-II optimization.

Author Contributions: Conceptualization, Y.W. and X.C.; methodology, Y.W.; software, Y.W.; validation, Y.W. and X.C.; formal analysis, Y.W.; investigation, Y.W.; data curation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, X.C.; visualization, Y.W.; supervision, X.C.; project administration, X.C.; funding acquisition, X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The compiled experimental dataset, trained model configurations, NSGA-II optimization scripts, and figure generation code are openly available in the GitHub repository at https://github.com/koopman-mpc-lab/bcc-alloy-hydrogen-storage-ml (accessed on 27 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BCC	Body-centered cubic
CP	Composition–property fusion
DFT	Density functional theory
ELF	Electron localization function
GA	Genetic algorithm
GB	Gradient boosting
HEA	High-entropy alloy
MAE	Mean absolute error
ML	Machine learning
MSE	Mean squared error
NSGA-II	Non-dominated Sorting Genetic Algorithm II
PAW	Projector augmented wave
PBE	Perdew–Burke–Ernzerhof
PCI	Pressure–composition isotherm
RF	Random forest
RMSE	Root mean squared error
SHAP	SHapley Additive exPlanations
SQS	Special quasi-random structure
SVR	Support vector regression
VASP	Vienna Ab initio Simulation Package
VEC	Valence electron concentration
XGB	Extreme gradient boosting

References

1. Schlapbach, L.; Züttel, A. Hydrogen-Storage Materials for Mobile Applications. *Nature* **2001**, *414*, 353–358. [CrossRef] [PubMed]
2. Sakintuna, B.; Lamari-Darkrim, F.; Hirscher, M. Metal Hydride Materials for Solid Hydrogen Storage: A Review. *Int. J. Hydrogen Energy* **2007**, *32*, 1121–1140. [CrossRef]
3. Shen, S.; Li, Y.; Ouyang, L.; Zhang, L.; Zhu, M.; Liu, Z. V–Ti-Based Solid Solution Alloys for Solid-State Hydrogen Storage. *Nano-Micro Lett.* **2025**, *17*, 175. [CrossRef] [PubMed]
4. Huang, Z.; Li, C.; Chu, Y.; Gu, J.; Li, W.; Xie, J.; Gao, G.; Wang, H.; Fan, M.; Yao, Z. Potential and Challenges for V-Based Solid Solution Hydrogen Storage Alloys. *Energy* **2025**, *316*, 134574. [CrossRef]
5. Akiba, E.; Iba, H. Hydrogen Absorption by Laves Phase Related BCC Solid Solution. *Intermetallics* **1998**, *6*, 461–470. [CrossRef]
6. Hu, H.Z.; Zhang, X.X.; Li, S.S.; Yi, L.C.; Chen, Q.J. A Review of Body-Centered Cubic-Structured Alloys for Hydrogen Storage: Composition, Structure, and Properties. *Rare Met.* **2025**, *44*, 1497–1521. [CrossRef]
7. Kumar, S.; Jain, A.; Ichikawa, T.; Kojima, Y.; Dey, G.K. Development of Vanadium Based Hydrogen Storage Material: A Review. *Renew. Sustain. Energy Rev.* **2017**, *72*, 791–800. [CrossRef]
8. Cho, S.W.; Han, C.S.; Park, C.N.; Akiba, E. The Hydrogen Storage Characteristics of Ti–Cr–V Alloys. *J. Alloys Compd.* **1999**, *288*, 294–298.
9. Suwarno, S.; Dicky, G.; Suyuthi, A.; Effendi, M.; Witantyo, W.; Noerochim, L.; Ismail, M. Machine Learning Analysis of Alloying Element Effects on Hydrogen Storage Properties of AB_2 Metal Hydrides. *Int. J. Hydrogen Energy* **2022**, *47*, 11938–11947. [CrossRef]
10. Nygård, M.M.; Øygarden, V.; Sørby, M.H.; Hauback, B.C. Counting Electrons—A New Approach to Tailor the Hydrogen Sorption Properties of High-Entropy Alloys. *Acta Mater.* **2019**, *175*, 121–129.
11. Ulmer, U.; Dieterich, M.; Pohl, A.; Dittmeyer, R.; Linder, M.; Fichtner, M. Study of the Structural, Thermodynamic and Cyclic Effects of Vanadium and Titanium Substitution in Laves-Phase AB_2 Hydrogen Storage Alloys. *Int. J. Hydrogen Energy* **2017**, *42*, 20103–20110. [CrossRef]

12. Yeh, J.W.; Chen, S.K.; Lin, S.J.; Gan, J.Y.; Chin, T.S.; Shun, T.T.; Tsau, C.H.; Chang, S.Y. Nanostructured High-Entropy Alloys with Multiple Principal Elements: Novel Alloy Design Concepts and Outcomes. *Adv. Eng. Mater.* **2004**, *6*, 299–303. [CrossRef]
13. Cantor, B.; Chang, I.T.H.; Knight, P.; Vincent, A.J.B. Microstructural Development in Equiatomic Multicomponent Alloys. *Mater. Sci. Eng. A* **2004**, *375–377*, 213–218. [CrossRef]
14. Sahlberg, M.; Karlsson, D.; Zlotea, C.; Jansson, U. Superior Hydrogen Storage in High Entropy Alloys. *Sci. Rep.* **2016**, *6*, 36770. [CrossRef] [PubMed]
15. Karlsson, D.; Ek, G.; Cedervall, J.; Zlotea, C.; Møller, K.T.; Hansen, T.C.; Bednarcik, J.; Paskevicius, M.; Sørby, M.H.; Jensen, T.R.; et al. Structure and Hydrogenation Properties of a HfNbTiVZr High-Entropy Alloy. *Inorg. Chem.* **2018**, *57*, 2103–2110. [PubMed]
16. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794.
17. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
18. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555. [CrossRef] [PubMed]
19. Agrawal, A.; Choudhary, A. Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Mater.* **2016**, *4*, 053208. [CrossRef]
20. Rahnama, A.; Zepon, G.; Sridhar, S. Machine Learning Based Prediction of Metal Hydrides for Hydrogen Storage, Part I: Prediction of Hydrogen Weight Percent. *Int. J. Hydrogen Energy* **2019**, *44*, 7337–7344. [CrossRef]
21. Witman, M.; Ling, S.; Grant, D.M.; Walker, G.S.; Agarwal, S.; Stavila, V.; Allendorf, M.D. Extracting an Empirical Intermetallic Hydride Design Principle from Limited Data via Interpretable Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 40–47. [PubMed]
22. Lu, Z.; Wang, J.; Wu, Y.; Guo, X.; Xiao, W. Predicting Hydrogen Storage Capacity of V–Ti–Cr–Fe Alloy via Ensemble Machine Learning. *Int. J. Hydrogen Energy* **2022**, *47*, 34583–34593. [CrossRef]
23. Beltrame, G.; Dematteis, E.M.; Stavila, V.; Rizzi, P.; Baricco, M.; Palumbo, M. Application of Machine Learning and Data Augmentation Algorithms in the Discovery of Metal Hydrides for Hydrogen Storage. *Metals* **2025**, *15*, 1221. [CrossRef]
24. Lu, X.; Luo, S.; Li, J.; Chen, M.; Yao, T.; Xu, Z.; Yan, Y.; Li, J.; Shao, X.; Gao, Z.; et al. FIND: A Forward–Inverse Navigation and Discovery Platform for Hydrogen Storage Alloys Powered by Data-Driven Machine Learning. *J. Mater. Inf.* **2025**, *5*, 48. [CrossRef]
25. Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, 54. [CrossRef]
26. Zhang, Y.; Zhou, Y.J.; Lin, J.P.; Chen, G.L.; Liaw, P.K. Solid-Solution Phase Formation Rules for Multi-Component Alloys. *Adv. Eng. Mater.* **2008**, *10*, 534–538. [CrossRef]
27. Guo, S.; Ng, C.; Lu, J.; Liu, C.T. Effect of Valence Electron Concentration on Stability of FCC or BCC Phase in High Entropy Alloys. *J. Appl. Phys.* **2011**, *109*, 103505. [CrossRef]
28. Yang, X.; Zhang, Y. Prediction of High-Entropy Stabilized Solid-Solution in Multi-Component Alloys. *Mater. Chem. Phys.* **2012**, *132*, 233–238.
29. Holland, J.H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; MIT Press: Cambridge, MA, USA, 1992.
30. Goldberg, D.E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, USA, 1989.
31. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2017; pp. 4765–4774.
32. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
33. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
34. Breiman, L. Stacked Regressions. *Mach. Learn.* **1996**, *24*, 49–64. [CrossRef]
35. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
36. Deb, K. *Multi-Objective Optimization Using Evolutionary Algorithms*; John Wiley & Sons: Chichester, UK, 2001.
37. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
38. Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169–11186.
39. Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15–50.
40. Blöchl, P.E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953–17979. [CrossRef]
41. Perdew, J.P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868. [CrossRef] [PubMed]

42. van de Walle, A.; Tiwary, P.; de Jong, M.; Olmsted, D.L.; Asta, M.; Dick, A.; Shin, D.; Wang, Y.; Chen, L.Q.; Liu, Z.K. Efficient Stochastic Generation of Special Quasirandom Structures. *Calphad* **2013**, *42*, 13–18. [CrossRef]
43. Henkelman, G.; Arnaldsson, A.; Jónsson, H. A Fast and Robust Algorithm for Bader Decomposition of Charge Density. *Comput. Mater. Sci.* **2006**, *36*, 354–360. [CrossRef]
44. Griessen, R.; Driessen, A. Heat of Formation and Band Structure of Binary and Ternary Metal Hydrides. *Phys. Rev. B* **1984**, *30*, 4372–4381. [CrossRef]
45. Smithson, H.; Marianetti, C.A.; Morgan, D.; Van der Ven, A.; Predith, A.; Ceder, G. First-Principles Study of the Stability and Electronic Structure of Metal Hydrides. *Phys. Rev. B* **2002**, *66*, 144107. [CrossRef]
46. Liu, Y.; Pan, H.; Gao, M.; Wang, Q. Advanced Hydrogen Storage Alloys for Ni/MH Rechargeable Batteries. *J. Mater. Chem.* **2011**, *21*, 4743–4755. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.