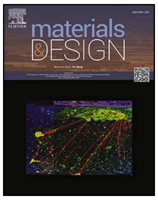
Materials & Design 266 (2026) 116097

Contents lists available at ScienceDirect

Materials & Design

journal homepage: www.elsevier.com/locate/matdes



Inverse design of high-entropy superalloys using machine learning and generative artificial Intelligence

François Rousseau [a,*], Thierry Belmonte [a], Frédéric Sur [c], Alexandre Nominé [a,b]

[a] Institut Jean Lamour, Université de Lorraine, CNRS, Campus Artem, 2 Allée André Guinier, 54011 Nancy, France
[b] Jožef Stefan Institute, Jamova Cesta 39 SI-1000 Ljubljana, Slovenia
[c] Laboratoire Lorrain de Recherche En Informatique Et Ses Applications (LORIA), Université de Lorraine, CNRS, Inria, Campus Scientifique, 615 Rue Du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France

ARTICLE INFO

Keywords:
Materials discovery
Inverse design
High-entropysuperalloys
Multi-objective optimization
Machine learning surrogates
Conditionalvariationalautoencoder
Computational materials

ABSTRACT

We introduce a structure-agnostic inverse-design workflow that turns heterogeneous literature and database evidence into experiment-prioritized shortlists of high-temperature high-entropy superalloy candidate chemistries. Unlike most data-driven high-entropy alloy (HEA) design studies that optimize a small set of proxies, we provide an end-to-end, reproducible decision-support loop that treats high-temperature creep and oxidation as first-class objectives and outputs compact shortlists for downstream validation. From curated multi-source data, we learn structure-agnostic, physics-informed surrogate models for the key high-temperature objectives – creep resistance (Larson–Miller parameter), oxidation kinetics (parabolic rate constant), melting point, density and elastic properties – and map predictions to "desirability" normalized scores. Candidate alloys are screened by a uniform feasibility floor and Pareto non-domination, then compressed by Ward–medoid clustering to yield compact, diversity-preserving shortlists for downstream validation. To explore beyond brute-force enumeration under the same admissibility rules, we couple the screening stage to a constraint-conditioned variational autoencoder, and retain only generated candidates that pass the full surrogate stack. The resulting compressed Pareto sets extend the occupied property envelope of legacy Ni-based superalloys while remaining interpretable through elemental-role analyses and Ashby-style trade-off maps. An external thermodynamic plausibility cross-check further shows that higher microstructure-oriented scores enrich the candidate space in single-phase HEA-compatible chemistries with wider stability windows and more favorable transition classes. Finally, we show how the same pipeline can be restricted to sustainability-compatible element pools, enabling performance-aware exploration under supply-risk and footprint constraints.

1. Introduction

Accelerating the discovery of high-performance materials remains a central challenge across the technologies needed for the "twin" (green & digital) transitions. Classical discovery – hypothesis, synthesis, characterization, iteration – has delivered outstanding alloys but typically at the cost of long cycles, narrow local searches and path dependence on existing chemistries [1,2]. The maturation of large materials databases and learning algorithms now enables inverse design: posing target properties first and then searching composition space for candidates likely to satisfy them [3–5]. This strategy is especially attractive if we can learn robust property surrogates from composition-derived, physics-informed descriptors, thus remaining predictive even where crystal structures are not yet known.

Superalloys are a cornerstone class of structural materials whenever high mechanical loads must be sustained at elevated temperatures for long durations. They enable critical technologies such as aero-engines and land-based gas turbines, where raising turbine inlet temperature remains a primary lever for efficiency, as well as a range of industrial components subjected to creep, thermo-mechanical fatigue, and hot corrosion/oxidation. In these extreme environments, performance is not captured by a single scalar metric: it emerges from a delicate balance between high-temperature strength, microstructural stability, environmental resistance, and, increasingly, constraints related to cost and element availability.

The exceptional high-temperature capability of superalloys is

---

* Corresponding author.
E-mail address: francois.rousseau@univ-lorraine.fr (F. Rousseau).

https://doi.org/10.1016/j.matdes.2026.116097
Received 1 March 2026; Received in revised form 8 April 2026; Accepted 22 April 2026

fundamentally rooted in microstructural engineering. In the archetypal case of Ni-based superalloys, strength is achieved through a γ (FCC) matrix reinforced by a dense dispersion of γ' precipitates (ordered 1_2), which are typically coherent or semi-coherent with the matrix. Such crystallographic coherency lowers interfacial energy, enables a fine and stable precipitate distribution, and makes precipitate–dislocation interactions highly effective, thereby slowing plastic flow and enhancing creep resistance. In addition to intragranular precipitation strengthening, many superalloys also rely on grain-boundary phases (e.g., carbides and borides) to stabilize grain boundaries and mitigate intergranular creep mechanisms, at the cost of additional trade-offs with ductility and damage tolerance.

Discovering new superalloy chemistries is therefore intrinsically challenging, because performance depends on a coupled design space involving composition, processing, heat treatments, and the resulting microstructural transformations (precipitation and partitioning pathways, coarsening/rafting, potential formation of brittle TCP phases, and the evolution of protective oxide scales, among others). In practice, predicting macroscopic proxies such as the Larson–Miller parameter or parabolic oxidation constants requires capturing, at least indirectly, the underlying transformation pathways – an endeavor that typically demands long-duration testing and advanced characterization. This complexity contributes to the relatively limited availability of curated, comparable datasets for high-entropy superalloys (HESA), despite the broad range of possible chemistries.

High-entropy alloys (HEA) offer a natural testbed for inverse design because their multi-principal-element degrees of freedom dramatically expand the compositional search space and can unlock unusual combinations of stiffness, creep resistance, oxidation behavior and melting temperature. Even restricting to equimolar five-component alloys chosen from a palette of 40 elements yields 658,008 candidates; prior predictive screening suggests that only a fraction may form high-temperature solid solutions (e.g., 30,201 equimolar candidates predicted by Chen et al. [6] to form a solid solution, reducing the equimolar space by roughly twenty-fold). However, once non-equimolar stoichiometries and practical compositional tuning are considered – often essential to optimize precipitation, stability, and high-temperature performance – the effective design space expands again by orders of magnitude.

Within HEA, we focus on high-entropy superalloy candidates (HESA) for high-temperature service; throughout this paper, “HESA candidates” denotes alloys shortlisted by structure-agnostic surrogates built primarily from composition-derived descriptors, without claiming superalloy status in the microstructural sense (i.e., without asserting the formation of the required γ/γ' type or equivalent strengthening microstructures). Accordingly, the present workflow should be understood as a composition-first screening stage: it prioritizes candidate chemistries for downstream structure-aware validation, but does not certify γ/γ' formation, phase stability over temperature, TCP avoidance, or the final microstructure resulting from processing and heat treatment. More expensive downstream validation tools – such as structure-aware thermodynamic calculations and targeted experiments – are natural follow-up steps once the composition space has been sufficiently reduced, but they cannot be applied at the same scale as the initial exploration. The practical target is not a single optimum but a multi-constraint compromise balancing creep (Larson–Miller parameter), oxidation kinetics, elastic properties, density and melting point – properties that must be aligned rather than optimized in isolation. Beyond performance, down-selection in realistic deployment scenarios also depends on non-physical dimensions (cost, supply security, environmental footprint). Consequently, a key bottleneck is not generating candidates, but rapidly identifying regions of interest where targeted experiments and microstructure-resolved optimization are most likely to succeed – this is precisely the purpose of the present work.

These bottlenecks are exacerbated in superalloy-grade systems, where performance is microstructure-governed and cannot be reliably inferred from a single proxy. First, despite the frequent framing of “high-temperature” applications, high-temperature creep and oxidation resistance are usually not explicitly addressed, typically falling back on proxies such as melting point or elevated-temperature strength. Second, targeting sound candidates from a large set for a necessary limited number of experiments requires to conceive an adapted methodology. Third, non-physical constraints – criticality, supply risk, environmental footprint – are rarely incorporated.

Against this background, our contribution is an end-to-end, reproducible inverse-design loop built specifically to close these gaps, while remaining structure-agnostic at the screening stage. Concretely, we make four advances:

i. High-temperature objectives as first-class targets: we integrate creep- and oxidation-related metrics directly into the same unified multi-property screening loop, rather than treating them as downstream checks or leaving them implicit.
ii. Robust and transferable surrogates: starting from curated data, we learn property surrogates from physics-informed elemental descriptors selected in an – automated – structured way (through a preliminary benchmark of the importance of each descriptor), enabling prediction across composition space and supporting extrapolation to elements not present in the training chemistry.
iii. Closed-loop constrained generation with validity screening: we use a constraint-conditioned generative model to propose new stoichiometries under 10 simultaneous targets, with tunable constraint weights to explore controlled compromises; generated candidates are then systematically re-validated through the full surrogate stack before being used to enrich the learning set and iterate the design loop.
iv. Decision-grade shortlisting and interpretability: we turn large feasible/Pareto-optimal sets into experimentally actionable, diversity-preserving shortlists using Ward-linkage compression with medoid representatives to reduce the set to a human-manageable shortlist, complemented by clustering to identify families of alloys and associated property profiles. Then follow partial-dependence analyses to reveal element-level roles and structure the design space into interpretable families.

To better ensure systematic exploration of composition space, we additionally generate compositions via a Hamming-distance–based enumeration strategy. Finally, we show how the same pipeline can be restricted to sustainability-compatible element pools, thereby moving from “physics-only optima” toward solutions that remain viable under strategic and environmental constraints.

The remainder of the manuscript is organized as follows. We first summarize the inverse-design workflow and the main screening outcomes, then examine the plausibility and diversity of the shortlisted candidate set through comparisons with literature HESA, an external thermodynamic plausibility check, and compositional clustering. We next analyze model interpretability and elemental roles. Finally, we discuss the main limitations of the framework, including the need for experimental validation and the role of additional non-physical constraints such as supply risk and environmental footprint.

2. Results

We first summarize the workflow and the main screening outputs, then examine the plausibility, diversity, and interpretability of the resulting candidate set through retrospective literature benchmarking, an external thermodynamic plausibility check, and compositional analysis.

2.1. Overview of the inverse-design workflow

Fig. 1 summarizes the end-to-end workflow: (i) data assembly/

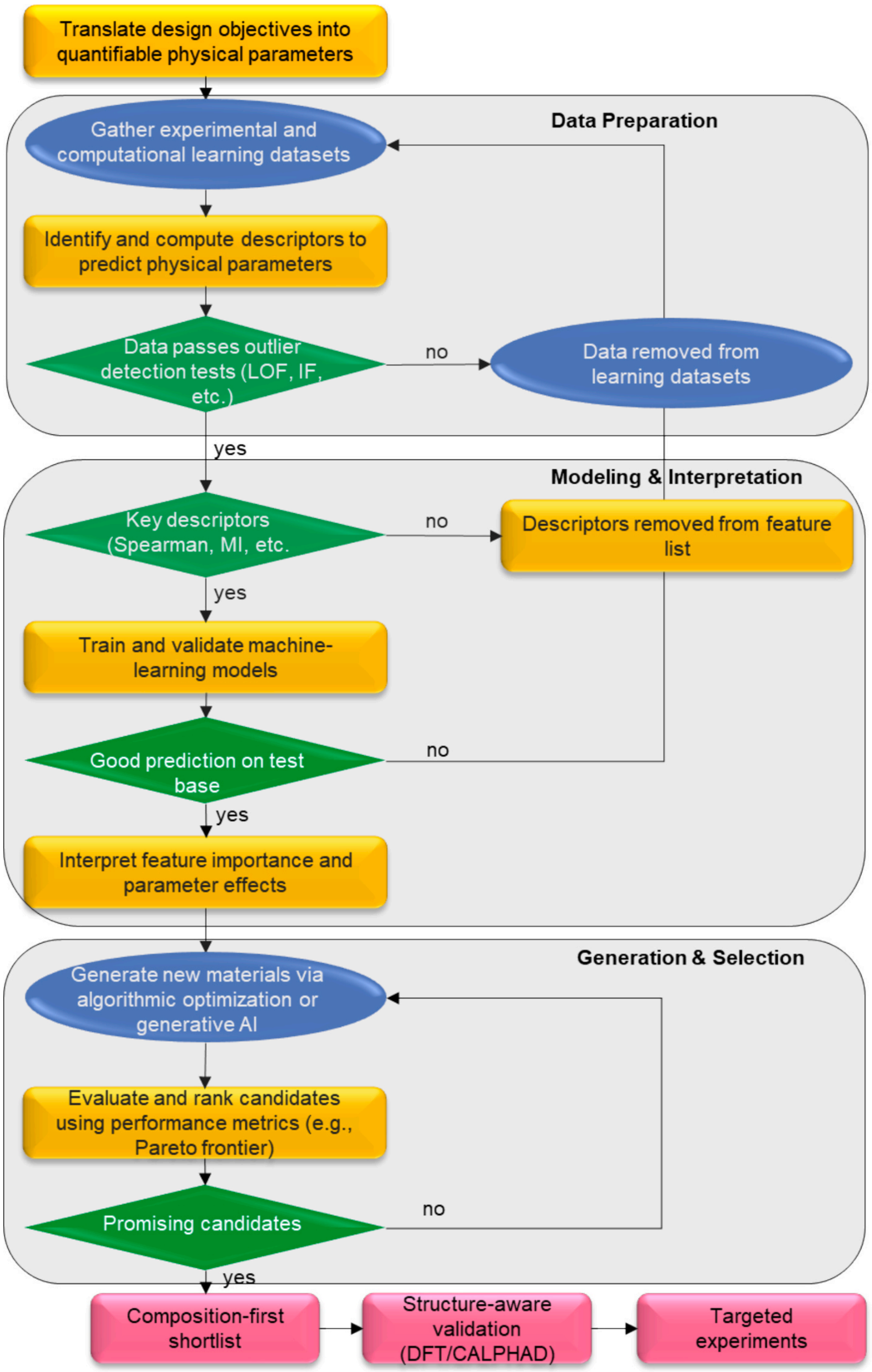


Fig. 1. Workflow for Generative Artificial Intelligence & Machine Learning Driven Materials Discovery Targeted to a Specific Application.

curation and physics-informed descriptor engineering (§1.2–1.4 of Supplementary Information (SI)); (ii) training/validation of structure-agnostic surrogates using a test base (20%) and feature selection (§1.5 of SI); (iii) multi-objective scoring, “Pareto” filtering (non-domination) (§3.2 of SI) and Conditional Variational Auto-Encoder (CVAE)-based exploration (§4.2–4.5 of SI). We use this figure here only as a visual roadmap; implementation details (outlier control, descriptor reduction, model selection) are given in Methods §5.1–5.4, while scoring and Pareto compression are specified in §5.5–5.6. Supervised model performance is reported in Table 3. We seed our exploration from the equimolar quinary single-phase HEA map reported by Chen et al. [6], who screened the equimolar quinary space and identified a subset of candidate single-phase solid solutions; we use the same 40-element pool and their candidate set as a thermodynamically motivated starting point, then generate off-equimolar variants to explore application-relevant trade-offs (see Fig. 21 in SI).

The workflow shown here corresponds to the large-scale composition-first screening stage; downstream structure-aware validation and targeted experiments are considered as subsequent steps once the shortlist has been reduced to a manageable size.

2.2. Main screening outcomes and diversity of Pareto candidates

The screening stage yields a large but structured set of non-dominated candidates, whose overall compositional diversity and performance spread are first summarized before more focused plausibility and interpretability analyses.

Across panels, the compressed Pareto set systematically extends the occupied property envelope beyond legacy Ni-based superalloys with totally new chemistries (see Fig. 14 for illustration and §4.7 in SI for a systematic review). Fig. 2 provides an insight on the element's frequency of occurrence and correlations, revealing more dispersed composition and letting an important place to unusual elements (Sc, Nb, Zr, Ta, etc.).

Each node denotes a chemical element, with its area proportional to the element's mean atomic fraction across all Pareto-optimal alloys (reference circles: 1% and 10% mean content, as shown in the legend). Edges join pairs of elements whose Pearson correlation coefficient satisfies $|r| \geq 0.3$ – and in light grey pairs whose correlation coefficient satisfies $|r| \geq 0.25$; their thickness scales linearly with $|r|$, so that thicker bonds indicate stronger positive or negative correlations. A node bound with at least 3 "strong" edges ($|r| \geq 0.3$) is highlighted with a navy-blue ring.

The Pareto front that is eventually obtained contains over 2,000 candidate alloys and is provided herewith (see **Data Availability** section). The initial stochastic brute-force screening is restricted to quinary (5-element) alloys according to single-phase solid solution candidates from Chen et al. [6]. Downstream enrichment steps (CVAE sampling and gradient-based refinement) may generate up to N elements (here: $N \leq 7$ as implemented), to allow local improvements while keeping complexity bounded. We report elemental cardinality explicitly for all shortlisted candidates and keep the stochastic brute-force and generative stages clearly separated in the scripts outputs. A selection of 7 samples of the Pareto front is provided in Table 1a and illustrates the diversity of compositions & performances.

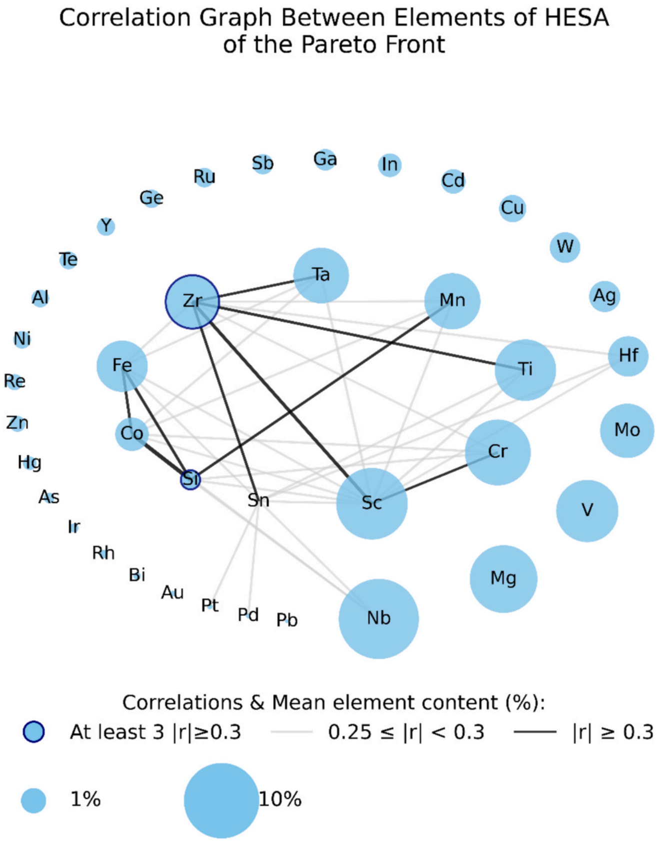


Fig. 2. Correlation Network of Elemental Compositions (CLR) on the HESA Pareto Front.

Table 1a
Sampling of the Pareto front.

Stoichiometric Composition	Density	Bulk Modulus (Gpa)	Melting Point (K)	Creep (LMP)	High-Temperature Oxidation Resistance
Mo: 29% Zr: 29% Sc: 22% Ta: 17% Hf: 3%	8.2	140	2310	31.4	0.4
Cr: 37% Fe: 29% Co: 23% Si: 6% V: 4%	7.6	200	1740	30.3	−1.1
Mo: 32% Sc: 23% V: 22% Ta: 16% Ga: 6%	8.7	190	2140	31.8	0.3
Mo: 55% Ta: 16% Sc: 10% V: 10% Ti: 4% Nb: 3% Zr: 2%	9.3	190	2500	31.2	0.4
Sc: 21% Ta: 21% Zr: 21% Mo: 21% Nb: 16%	8.2	150	2480	31.4	0.4
Cr: 39% Ti: 29% Al: 17% Ta: 12% Te: 2%	6.4	130	1810	31.1	0.3
Sc: 28% V: 24% Ta: 14% Ti: 13% Zr: 8% Nb: 8% Hf: 4%	7.0	140	2120	31.2	0.3

These 7 candidates illustrate the diversity in composition and performances of the Pareto front. The presence of rare elements like Ta or Sc, confirmed in Fig. 2, must be noticed: it is both original, and raises supply risk concerns.

2.3. External plausibility check using literature HESA

As a retrospective external plausibility benchmark (not an experimental validation), we computed the same g-scores (see Table 4) for a subset of HESA compositions reported in the review by Ayaz et al. [7] and compared their score distribution to our screened candidates (Fig. 3). At the aggregate level, the literature HESA occupy an intermediate region of the $\left(\min g_i, \operatorname{mean} g_i\right)$ plane: their minimum & median mean score are provided in Table 1b:

Most literature HESA remain limited by one or a few bottleneck criteria, most often VEC, formation enthalpy, Ω parameter (competition entropy / enthalpy), and high-temperature oxidation. This is consistent with the deliberately conservative intent of the screening stage: we prioritize compositions that simultaneously satisfy all constraints under a structure-agnostic model, while acknowledging that some reported "HESA" may rely on microstructure-dependent mechanisms not captured here. Overall, this benchmark does not constitute experimental validation, but it does support the plausibility of the ranking and makes the main bottlenecks more explicit.

Minimum score $\min_i(g_i)$ versus mean score $\left(\sum_i g_i\right)/10$ for (i) HESA compositions extracted from the Ayaz et al. [7] review (circles) and (ii) screened candidates from this work (diamonds). The vertical dashed line marks the feasibility floor $\min_i(g_i) = 0.3$. Colors indicate which score most frequently acts as the limiting bottleneck among $\{g_4, g_7, g_8, g_{10}\}$; black diamonds satisfy $\min_i(g_i) \geq 0.3$. The figure provides a retrospective benchmark of the ranking layer and highlights the dominant bottleneck criteria, but does not constitute prospective experimental validation.

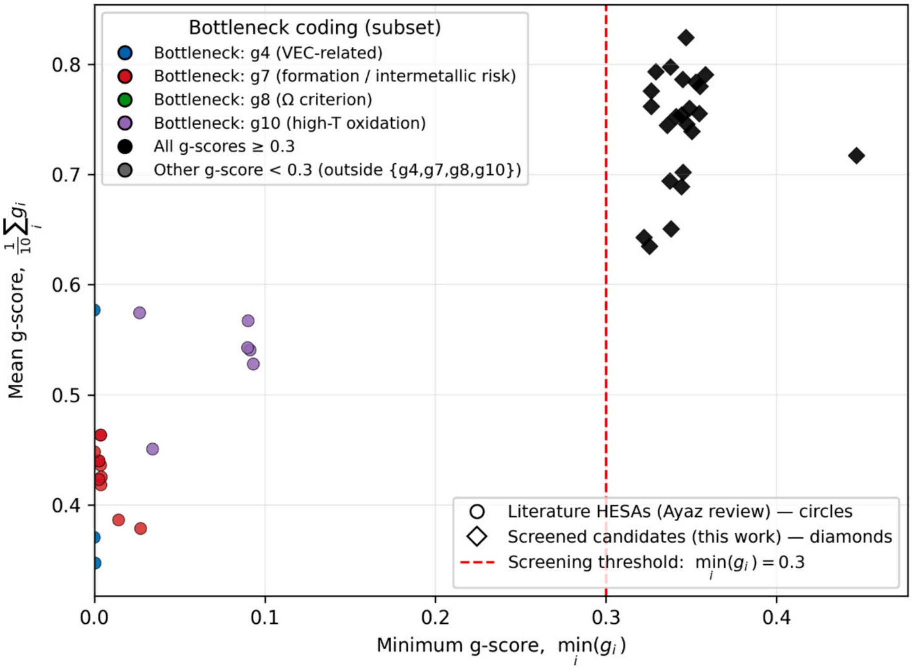


Fig. 3. Retrospective external plausibility benchmark using literature HESA compositions.

Table 1b
Summary of the retrospective literature benchmark ().

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
Min.	0.78	0.52	0.15	0.00	0.05	0.56	0.00	0.02	0.13	0.03
Mean	0.79	0.58	0.44	0.29	0.54	0.68	0.11	0.26	0.73	0.20

Source: Ayaz et al. [7]

2.4. *External thermodynamic plausibility check using a third-party phase-prediction route*

High-entropy superalloy (HESA) design cannot be reduced to the sole identification of alloys forming a single-phase solid solution. A relevant HESA must first exhibit a stable high-temperature solid-solution matrix and then evolve, upon cooling or ageing, toward a precipitation-hardened microstructure. Ideally, this transformation should generate a second phase that remains crystallographically coherent, or at least sufficiently compatible, with the matrix while also providing a meaningful lattice mismatch capable of inducing strengthening elastic strains. However, no simple predictive tool is currently available that can reliably assess, over a broad compositional space, interfacial coherence, effective lattice mismatch, or the actual ability of a given phase transformation to generate a superalloy-type microstructure. At this stage, only experiments can provide the validation datasets ultimately required for fully predictive AI-based models targeting this problem.

Accordingly, the present workflow should be regarded as a first screening stage that enables experimentalists to focus on more promising alloy families within an immense compositional space. In this sense, our approach is intended as a bridge between fully predictive AI-based design–which remains the long-term objective but is not yet achievable because of the lack of sufficiently rich validation data–and either random exploration or research restricted to already known alloy families. To complement the retrospective benchmark against literature HESA compositions, we therefore performed an external thermodynamic plausibility check using the DFT-informed thermodynamic phase-prediction model of Chen et al. [6], queried through its public Python interface and post-processed with an in-house Python scraper. This model is based on a regular-solution thermodynamic formalism with binary interactions obtained from DFT and explicit competition with intermetallic phases up to ternaries, which makes it suitable for large-scale HEA screening while remaining more physically grounded than purely empirical descriptor rules. Importantly, it still remains an intermediate thermodynamic screening layer rather than a full microstructural validation tool.

For this analysis, we focused on the five desirability terms in our decision layer that were initially introduced because of their expected connection to phase selection or microstructural plausibility, namely g_4, g_5, g_7, g_8, and g_9, associated respectively with VEC, atomic-size mismatch, formation enthalpy, Ω, and configurational entropy. We used their arithmetic mean, $\mathrm{mean}(g_4, g_5, g_7, g_8, g_9)$, as a compact microstructure-oriented score and grouped alloys into three categories: (i) random MPEA alloys with very low mean microstructure-oriented score, (ii) Pareto-front alloys with lower mean score, and (iii) Pareto-front alloys with higher mean score. Within the Pareto-derived subsets, we further distinguished candidates originating from the stochastic brute-force branch seeded from the Chen et al. HEA map and candidates produced by the generative CVAE stage, in order to separate thermodynamically safer proposals from more exploratory ones. This distinction is important because the brute-force branch is rooted in a composition space already biased toward single-phase HEA formation, whereas the generative branch is allowed to explore more audacious chemistries.

From the Chen et al. route, we extracted three thermodynamic/phase indicators. The first is the fraction of alloys forming a stable single-phase solid solution, which we interpret here as a necessary prerequisite for the existence of an HEA-type matrix. The second is the width of the corresponding stability window in temperature. The third concerns the class of transformation predicted below the solid-solution stability limit.

We distinguish four classes: *Spinodal*, *Spinodal + IM* (inter-metallics), *IM precipitation*, and *Multiple IM*. In the absence of an explicit model for crystallographic coherence and effective lattice mismatch, the first three classes cannot be considered as validated HESA microstructures, but they remain, at least in principle, compatible with the preservation of a solid-solution matrix together with a secondary transformation pathway. By contrast, the *Multiple IM* class is considered here as clearly unfavorable, since decomposition into several intermetallic compounds is unlikely to produce a coherent or controllable precipitation-hardened architecture. For this reason, rather than using only the fraction of pure spinodal decompositions, we retained as a conservative performance indicator the fraction of cases that do not fall into the *Multiple IM* class. This choice provides a robust thermodynamic plausibility criterion while avoiding overclaiming about the precise nature of the final microstructure.

The figure summarizes three outputs obtained from the DFT-informed thermodynamic phase-prediction model of Chen et al. [6], for alloy subsets defined from the mean microstructure-oriented score $\text{mean}(g_4, g_5, g_7, g_8, g_9)$: (top) fraction of compositions predicted to form a single-phase HEA, (middle) average stability window expressed in units of 10^3K, and (bottom) fraction of alloys whose predicted transformation class below the solid-solution stability limit is not *Multiple IM*. The left column corresponds to random MPEA alloys with very low mean microstructure-oriented score. The two Pareto-front categories correspond to lower and higher score ranges, respectively. Within Pareto-derived sets, each square is split into two triangles: the upper-left triangle reports candidates from the stochastic brute-force branch seeded from the Chen et al. single-phase HEA map [6], whereas the lower-right triangle reports candidates generated by the CVAE branch. Numbers indicate the average metric value within each subgroup together with the subgroup size. The figure is intended as a retrospective thermodynamic plausibility check of the composition-first screening stage, not as an experimental validation of HESA microstructures.

Several conclusions emerge from Fig. 4. First, the contrast between the random low-score MPEA alloys and the Pareto-derived candidates is very strong. In the present sample, the random low-score subset is almost entirely devoid of single-phase HEA formation, exhibits a nearly vanishing stability window, and almost never avoids the clearly unfavorable *Multiple IM* regime. By contrast, Pareto-derived candidates are overwhelmingly HEA-compatible and display substantial stability windows. This indicates that the composition-first pipeline does not merely identify arbitrary chemistries: it enriches the candidate space in alloys that remain thermodynamically plausible from the viewpoint of a downstream phase-based screening route.

Second, the microstructure-oriented score level matters. Moving from the lower-score Pareto subset to the higher-score Pareto subset preserves near-unity HEA compatibility while increasing the average

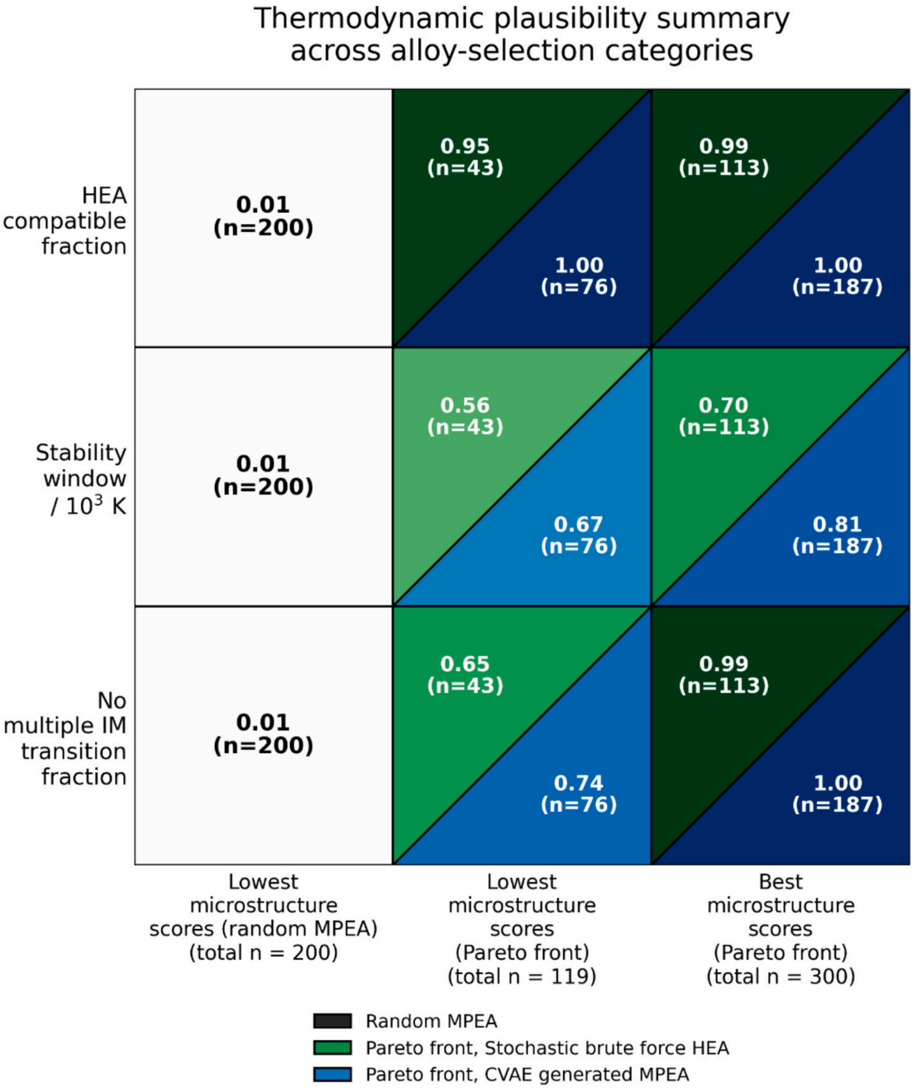


Fig. 4. External thermodynamic plausibility summary across alloy-selection categories.

stability window and, importantly, strongly increasing the fraction of transitions that avoid the clearly unfavorable *Multiple IM* class. This is consistent with the intended role of g_4, g_5, g_7, g_8, and g_9: while they do not guarantee any superalloy microstructure, they appear to steer the search toward composition regions that satisfy more favorable thermodynamic prerequisites for a HESA-like pathway.

Third, the comparison between the two Pareto-generation routes is informative. The stochastic brute-force branch, which starts from a composition space already biased toward HEA formation, unsurprisingly performs very well in terms of HEA compatibility. The generative CVAE branch reaches a comparable HEA-compatibility level and, in the highest-score subset, exhibits similarly large or even slightly larger stability windows, together with an equally strong suppression of clearly unfavorable *Multiple IM* outcomes. This suggests that the generative stage is not merely producing implausible outliers but can recover candidates that remain thermodynamically competitive with those inherited from the brute-force HEA map.

To refine the interpretation of the third indicator, Fig. 5 reports the detailed distribution of the four phase-transition classes within the Pareto-derived candidate subsets. This additional representation is useful because the “no multiple IM” metric of Fig. 4 is intentionally conservative: it tells us whether a candidate avoids a clearly unfavorable outcome, but it does not distinguish between the three remaining transformation families, which are not equally informative from a metallurgical standpoint.

Stacked histogram showing the relative fractions of the four transition classes predicted below the solid-solution stability limit for the Pareto-derived candidate subsets: *Spinodal*, *Spinodal + IM*, *IM precipitation*, and *Multiple IM or no HEA*. The four bars correspond to (i) stochastic brute-force HEA candidates with lower mean microstructure-oriented score, (ii) CVAE-generated candidates with lower mean score, (iii) stochastic brute-force HEA candidates with higher mean score, and (iv) CVAE-generated candidates with higher mean score. Each bar sums to 100%. The class *Multiple IM or no HEA* merges clearly unfavorable multiple-intermetallic decompositions with the absence of a stable high-temperature solid solution, both of which are considered incompatible with the targeted HESA-like architecture. This figure complements Fig. 4 by resolving the internal distribution of phase-transition scenarios within the Pareto-derived sets.

Fig. 5 confirms and sharpens the message of Fig. 4. In the lower-score Pareto subsets, unfavorable outcomes remain common: the merged class *Multiple IM or no HEA* still accounts for about one quarter to one third of the candidates, and no pure *Spinodal* case is observed. By contrast, in the higher-score Pareto subsets, the unfavorable class becomes nearly absent, while the candidate space becomes dominated by *Spinodal + IM* transitions and also starts to exhibit a non-negligible fraction of pure *Spinodal* decompositions. In other words, increasing the mean microstructure-oriented score does not demonstrate the direct formation of a HESA microstructure, but it clearly shifts the transition landscape away from strongly unfavorable decomposition patterns and toward transformation families that are at least not inconsistent with a precipitation-based strengthening route.

Overall, Fig. 4 and Fig. 5 support a more precise and more defensible interpretation of our workflow. The present pipeline should not be viewed as a direct predictor of the final superalloy microstructure, and the Chen et al. phase-prediction route [6] used here should likewise be understood as an intermediate thermodynamic screening layer rather than as a substitute for targeted CALPHAD/DFT analyses or experiments. Rather, the combined evidence supports interpreting the workflow as a composition-first screening stage that significantly enriches the search space in thermodynamically plausible HEA candidates and, among them, reduces the frequency of clearly unfavorable decomposition modes before any downstream structure-aware validation or experimental synthesis.

2.5. *Chemical families and clustering of the Pareto front*

Using a Ward linkage process, these alloys were grouped in four clusters regarding their chemical composition (§4.7.2 of SI). The physical performances of these clusters are represented in Fig. 6, together with the qualitative composition of their centroid – keeping only significant elements, i.e. over 5 at. %.

Physical characterization of the 4 clusters were defined through Ward clustering, by providing the average value and standard deviation within each cluster of the physical quantities mentioned above. This shows that while alloys belonging to the green & red clusters reach the highest melting point, alloys from the blue cluster are promising, regarding their resistance to high-temperature oxidation and bulk & shear moduli, or density. The variations within the cluster do not affect the rankings vs physical properties.

The clusters are chemically roughly described by the composition of their centroid (summarized as the list of elements with a content over 5% in the legend).

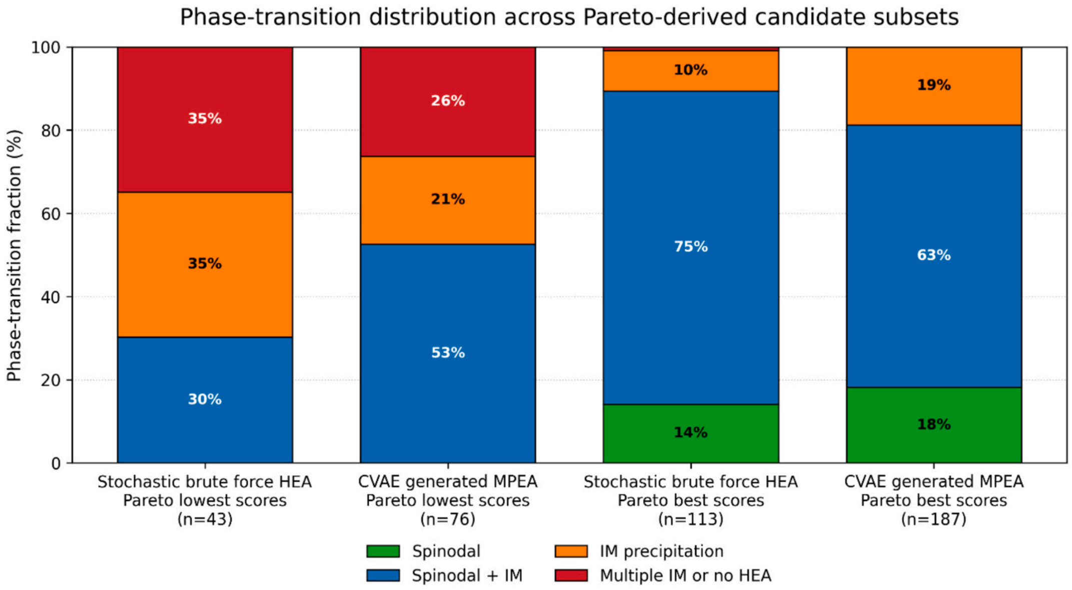


Fig. 5. Phase-transition distribution across Pareto-derived candidate subsets.

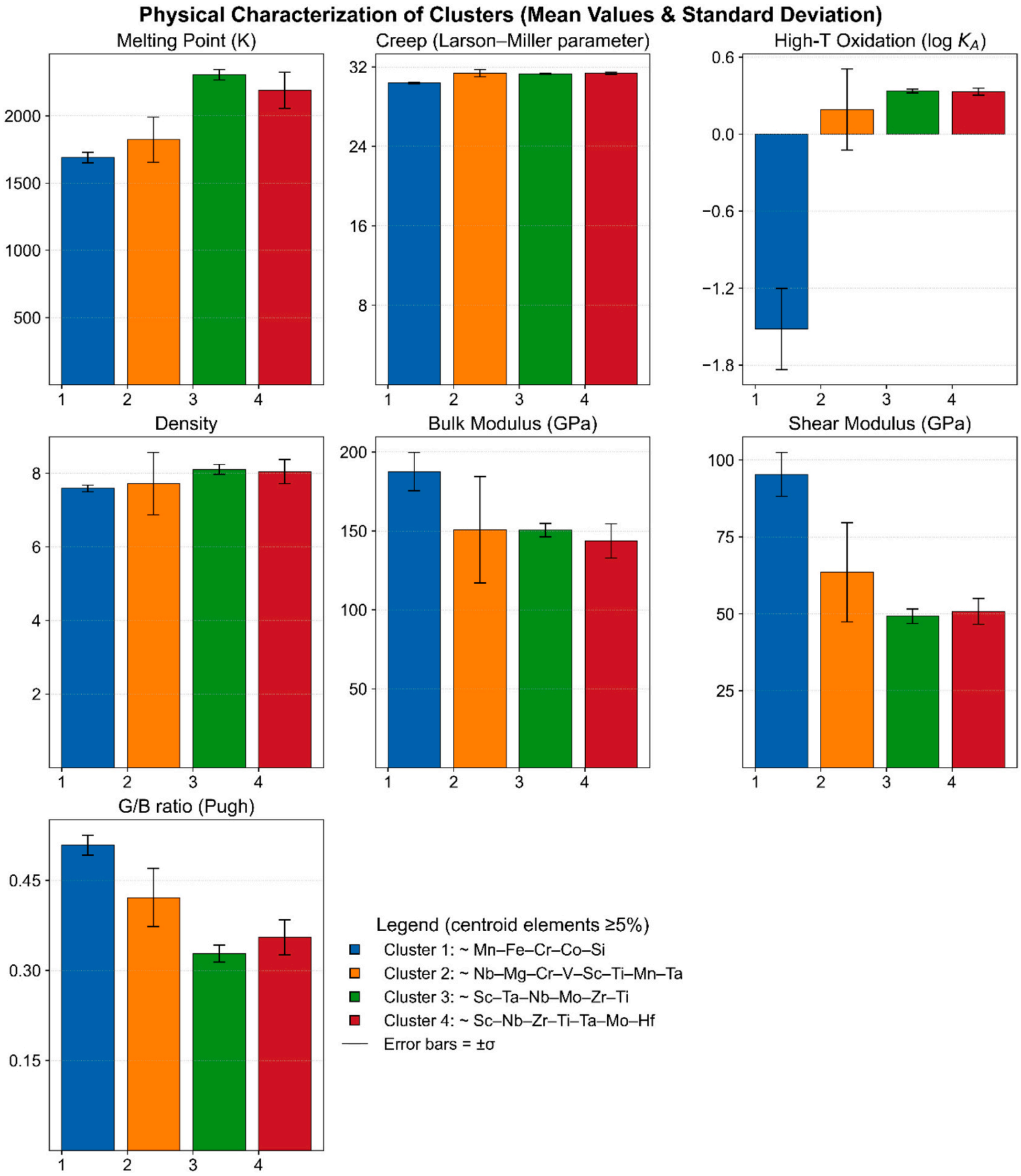


Fig. 6. Physical Characterization of Clusters (Mean Value & Standard Deviation).

2.6. *Elemental roles and model interpretability*

To move beyond aggregate metrics, we analyze drivers of predictions with complementary viewpoints: embedded feature importance, permutation importance, SHAP and partial dependence (§3.4 of SI). Because the atomic fractions sum to one, the data contains artificial collinearity and distorted distances which makes standard machine-learning methods unreliable. To mitigate the issue, we therefore apply a centered log-ratio (CLR) transform.

For the different physical performances predicted, feature-importance highlights the set of elements repeatedly selected by boosted-tree splits; permutation importance detects their out-of-sample influence; SHAP yields directionality and spread, consolidating which high/low elemental contents promote the score. Focusing on consistent trends across these three measures increases robustness and limits artefacts from any single method.

Partial dependence summarizes marginal trends per element. This disambiguates "frequent splitter" effects from genuine monotone contributions and supports chemically interpretable rules that feed into the multi-objective selection downstream (Fig. 7).

Heatmap depicting the significance and directionality of each element's effect (when forced at a fixed proportion of 30%) on the five considered physical properties. Green regions correspond to consistently positive effects (mean relative Δ significantly above zero), red indicates consistently negative effects (mean relative Δ significantly below zero), yellow denotes ambiguous impacts (mean ± standard deviation includes zero), and white reflects neutral or negligible impacts.

Beyond hold-out test performance, we verify that the most significant directional element effects are consistent with well-established metallurgy in the relevant regimes (SI, Table 3). These checks do not replace experiments, but increase confidence in the surrogate models.

3. Discussion

3.1. *Limitations and the need for experimental validation*

Validation operates at two distinct levels in the present study. First, each surrogate property model is quantitatively evaluated on unseen test

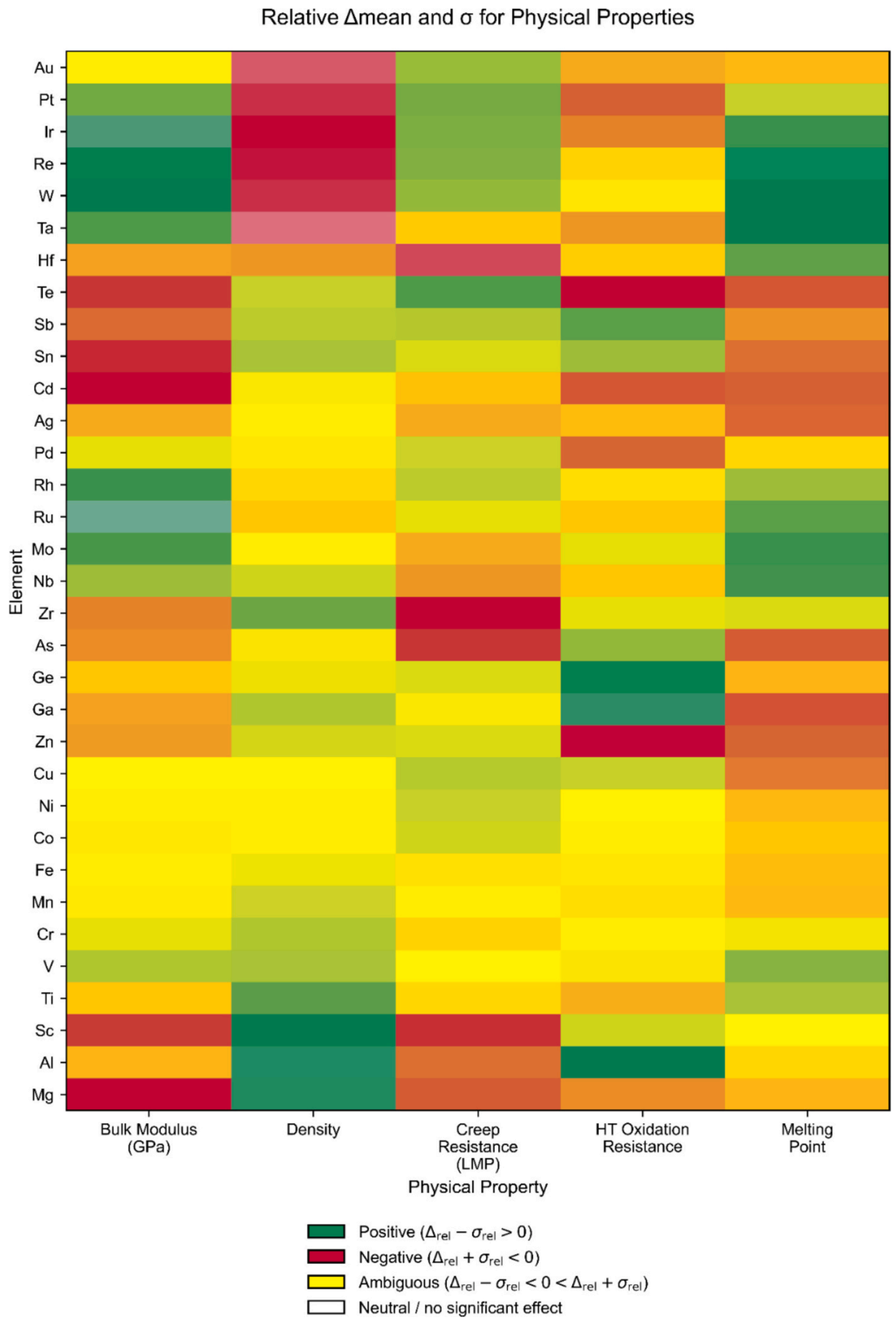


Fig. 7. Mean Shift & Standard Deviation of Partial Dependence of Physical Properties.

data. Second, the multi-objective screening/ranking layer is benchmarked retrospectively against reported literature HESA compositions and legacy reference alloys. By contrast, prospective validation of newly shortlisted candidates would require synthesis, characterization, and long-duration testing, and therefore remains outside the scope of the present computational screening study.

This limitation is not incidental but structural to the present study design. In superalloy development, composition selection and microstructure optimization are tightly coupled but remain distinct stages. The present work addresses the first stage only: reducing a very large chemical space to a physically motivated shortlist of candidate chemistries. Establishing phase constitution, γ/γ' stability, TCP avoidance, grain-size effects, and processing-dependent microstructural optimization requires downstream structure-aware calculations and experimental validation, and therefore lies outside the scope of the present composition-first screening framework.

Our framework relies on structure-agnostic surrogates trained on heterogeneous sources and on generative proposals filtered by multi-objective scores. While the close agreement between predicted and observed values on the held-out test set supports the ability of the surrogates to generalize beyond the training data, several limitations remain. First, targets with scarce and noisy measurements – especially high-temperature oxidation – exhibit wider uncertainty bands and stronger sensitivity to curation choices than the better-populated computational targets. In the present workflow, the oxidation branch should therefore be read as a conservative selector within the screening stack rather than as a standalone certification tool. Second, composition-first descriptors were chosen deliberately to retain predictive coverage in regions where no crystal structure or processed microstructure is yet known. This broadens the accessible search space, but it also means that the screening stage does not resolve the microstructural features that ultimately govern service performance, such as phase constitution, γ/γ' stability, precipitate morphology, grain-size effects, or the possible formation of brittle competing phases. Third, generative sampling proposes chemically plausible mixtures under score constraints, yet feasibility of synthesis, phase stability, and processability (e.g., possibility of growing a unique crystal, segregation, oxidation during processing, etc.) remains at least a technological lock and cannot be certified in silico.

Consequently, experimental synthesis and characterization are indispensable to probe phase constitution and validate property predictions, closing the design–test–retrain loop. Computational screening should precede and inform, not replace, targeted experimentation, especially for complex HESA where microstructure is decisive.

The present workflow should therefore be read as the first stage of a broader alloy-discovery strategy. Its role is to reduce a very large composition space to a physically motivated shortlist. Once this reduction has been achieved, more expensive downstream tools – including structure-aware thermodynamic calculations and targeted synthesis/characterization – become both feasible and meaningful. The absence of such downstream steps in the present manuscript should therefore be read as a scope choice, not as a claim that they are unnecessary.

As retrospective benchmarks in the absence of prospective validation, we complemented the literature-HESA score-space comparison (Fig. 3 and Table 1) with an external thermodynamic plausibility check based on a third-party phase-prediction route (Fig. 4). The latter does not certify γ' formation, TCP avoidance, or the final processing-dependent microstructure, but it does show that higher microstructure-oriented scores enrich the candidate space in single-phase HEA-compatible compositions, with wider stability windows and a lower fraction of unfavorable transitions. These benchmarks do not replace targeted CALPHAD/DFT analyses or experiments, but they strengthen the interpretation of the workflow as a composition-first pre-screening stage rather than as a direct microstructure predictor.

3.2. Beyond physics: Industrial choices integrate cost, security of supply, and environmental footprint

Industrial down-selection rarely optimizes physical performance alone. Cost, supply-security, and environmental burden are co-determinants of viable choices. Considering various parameters: energy footprint (data source: ANSYS Granta. Granta Selector Database. ANSYS, Inc.) – a first-principle decent proxy for prices & overall environmental impact [8,9] –, availability (world annual production [10]), risks induced by by-product exposure or monopolistic situations – informed through the Herfindahl Hirschman Index [10] – we proceed to a filtering of elements based on 6 rules detailed in SI (section 5, and more specifically subsection 5.3 and Table 6) and summarized on Fig. 8, then reassesses candidates on a "sustainable" front. Fig. 9 summarizes the outcome: when supply risk and energy footprint are considered alongside seven physical performance axes, the "sustainable" Pareto set remains competitive – often challenging or outperforming legacy superalloys on moduli, density, Pugh ratio, and high-temperature oxidation resistance – while naturally improving the non-physical indicators. This motivates prioritizing those candidates for experimental validation.

The elements considered "unsuitable" are in red with the regards of the rules suggested in the SI (see Table 6), whereas the elements present in the clean dataset are in green. The elements that were not considered for potential HESA are not colored.

On this diagram, seven "physical performances" are represented, in addition to the sustainability indicators "supply risk [11]" and "energy footprint". In spite of the fact that many elements were filtered out in the "sustainable" alloys set, it is noticeable that this set still achieves very interesting performances for bulk and shear modulus, density, Pugh ratio, and high-temperature oxidation, challenging or even outperforming legacy superalloys, while naturally obtaining the best scores for supply risk and energy footprint. Unless a really high melting point or resistance to creep is required, its physical performances compare with the overall best Pareto front, making the alloys of this set a priority for experimental validation.

3.3. Generality of the inverse-design route

Our inverse-design strategy proceeds from curated, structure-agnostic descriptors to validated surrogate models, then to multi-objective screening and finally to constrained generative exploration. Although developed here for HESA, this workflow naturally extends beyond this specific alloy family. Any materials class where composition

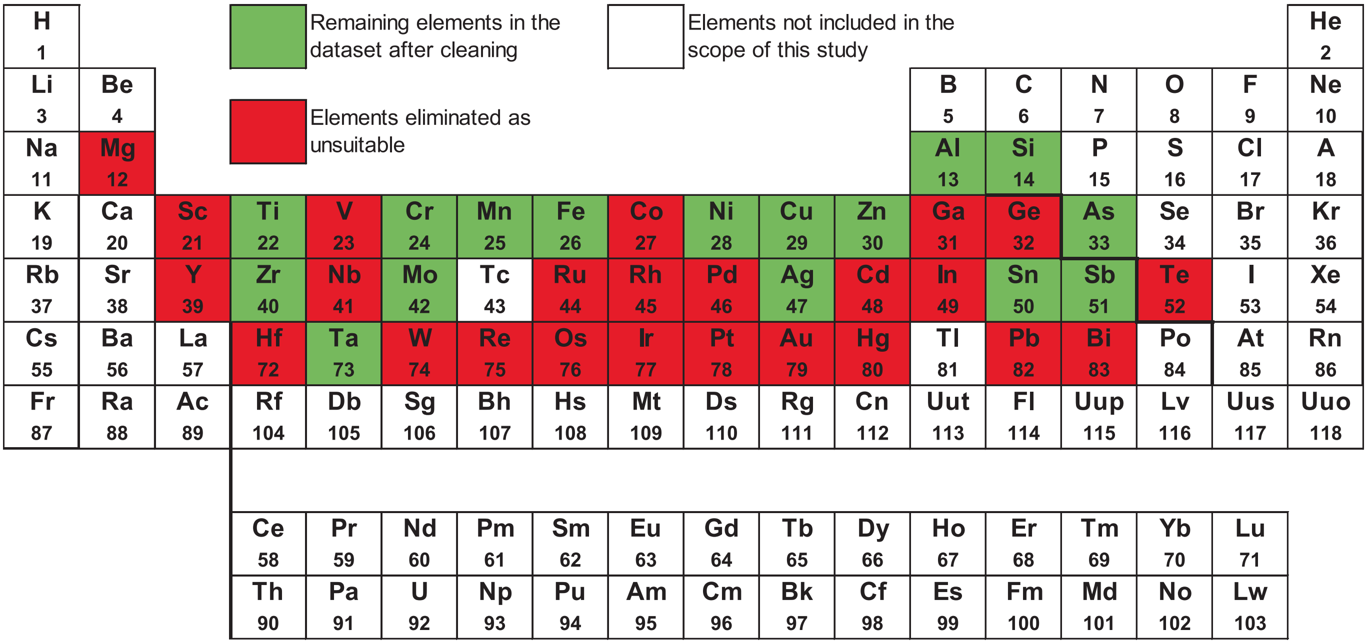


Fig. 8. Remaining elements in the data set after cleaning.

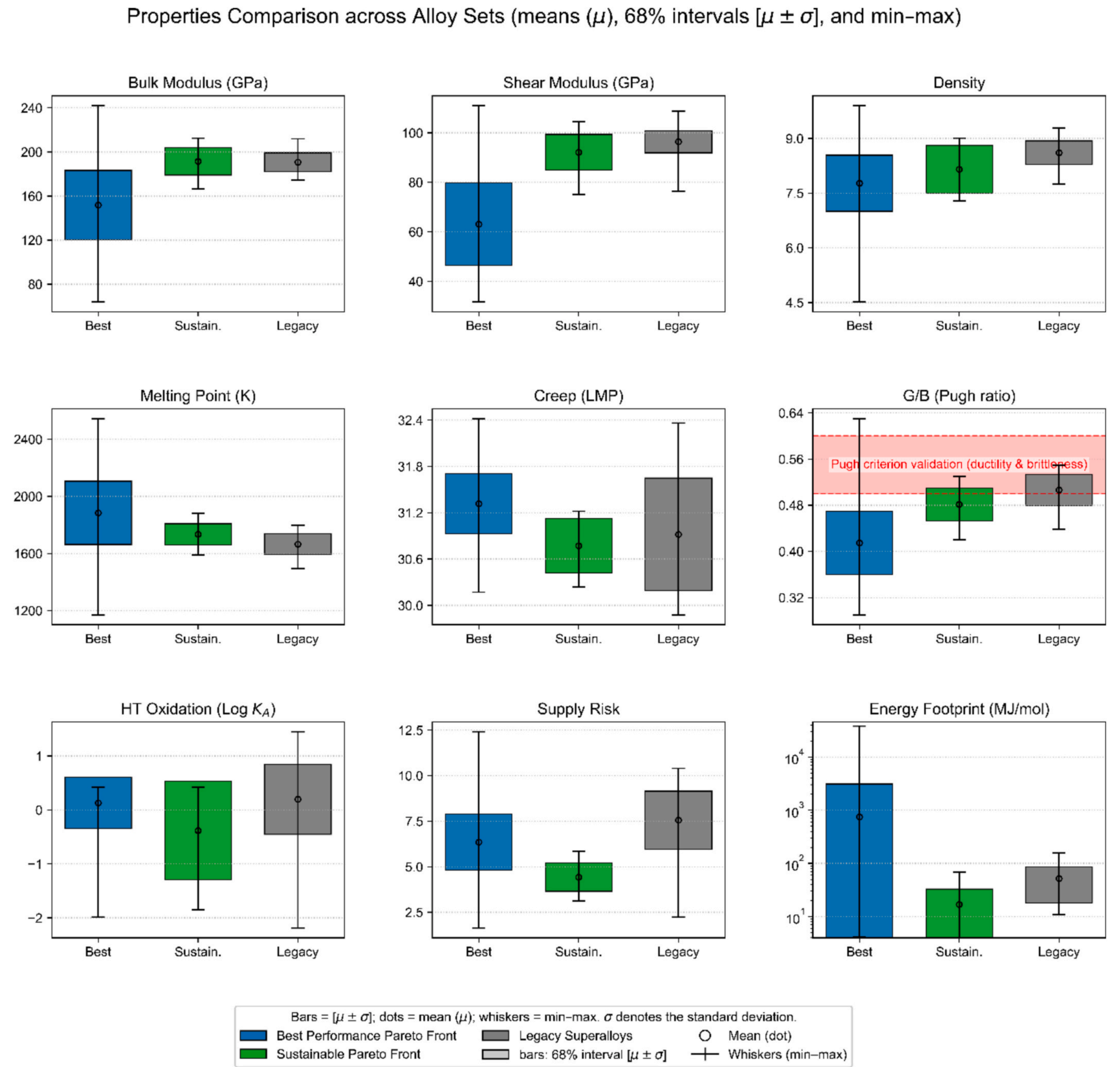


Fig. 9. Comparison of Properties across Alloys Sets: Overall Pareto Front, "Sustainable" Alloys Pareto Front and Legacy Superalloys.

is a primary design lever and target properties can be consistently learned from available data can adopt the same recipe. Examples include cathode/anode chemistries, thermoelectric and magnetic alloys, coatings and oxidation-resistant systems, or high-entropy ceramics, provided that (i) descriptors capture the relevant physics at the stage of interest (stoichiometric, structural), (ii) the data are available in quantity and quality to train the surrogate models, (iii) the score space encodes the right trade-offs (performance, durability, processing), and (iv) a retrainable loop couples predictions to experiments. In all cases, the generative component is model-agnostic with respect to chemistry; once its proposals are passed through the same feasibility floors and non-domination filters, the downstream analysis is unchanged. The main adaptation is domain-specific conditioning (targets and floors) and, where appropriate, the inclusion of non-physical constraints earlier in the loop to steer discovery toward deployable solutions.

4. Conclusion

We presented an end-to-end, structure-agnostic inverse-design workflow that converts heterogeneous literature and database evidence into experiment-prioritized shortlists of high-temperature high-entropy superalloy candidate chemistries. The pipeline learns physics-informed surrogate models built primarily from composition-derived descriptors, complemented where needed by branch-specific test-condition variables, and maps predicted properties to desirability scores for multi-objective screening. A uniform feasibility floor and Pareto non-domination identify feasible trade-offs, while Ward–medoid compression yields compact, diversity-preserving representative sets suitable for experimental planning and Ashby-style visualization. Unlike many HEA design studies that rely on a small set of proxies, our workflow treats high-temperature creep (LMP) and oxidation kinetics as first-class objectives within the same screening loop. A constraint-conditioned VAE expands exploration under the same admissibility rules, with systematic re-validation through the full surrogate stack. Finally, we illustrated how restricting the element pool using explicit sustainability rules can still retain competitive candidates under supply-risk and footprint considerations. Experimental synthesis and downstream structure-aware validation remain essential to confirm phase constitution, exclude brittle competing phases, and validate the service-relevant behavior of the shortlisted candidates.

5. Methods

5.1. Data sources & inclusion/exclusion

We assembled a multi-source alloys dataset spanning thermodynamics, elasticity, melting behavior and service-relevant performance from five curated repositories: DeepMind GNoME [12] (thermodynamic stability, densities for predicted stable crystals), the Materials Project [13] (formation energies, elastic moduli from DFT), MPContribs [14] – Melting Points, the Cambridge Superalloy Database [15–26] (experimental creep summarized by the Larson–Miller parameter, LMP), and NASA cyclic-oxidation studies [27,28,28–44] (parabolic rate constant K_A). To remain predictive for conjectured alloys without known structures, we exclude structure-dependent features and learn primarily on composition-derived, physics-meaningful descriptors, complemented where needed by branch-specific test-condition variables. The corresponding mapping from physical quantity to source is summarized below. Although the overall workflow is multi-source at the study level, each surrogate model is trained independently for a single target property. Experimental and computational labels are therefore not pooled into a single regression task.

5.1.1. Inclusion rules

(i) Keep fully metallic chemistries for thermodynamic/elastic/melting subsets; (ii) retain GNoME-stable entries and Materials Project entries with non-positive energy above hull; (iii) include creep records only when both stress and temperature are reported, so that the Larson–Miller parameter (LMP) is consistently defined; (iv) include oxidation records only when the test temperature is reported, so that temperature-dependent oxidation behavior can be modeled explicitly.

5.1.2. Exclusion rules

(i) Discard compounds with positive energy above convex hull (thermodynamic instability) in DFT-derived subsets; (ii) pre-filter non-fully metallic compositions before descriptor generation for MP/GNoME/MPContribs; (iii) for creep, because data are scarce, tolerate limited non-metallic presence but add fraction of non-metals as an explicit descriptor; (iv) for oxidation, discard records lacking temperature.

A consolidated overview of targets and sources (formation and decomposition energies, bulk modulus B, shear modulus G, density ρ, melting point T_m, LMP at specified stress, and the decimal logarithm of the parabolic constant $\log_{10}K_A$) appears in Table 2.

Table 2
Physical Quantities to be Learnt and Data Sources available for the Process.

Physical Quantity	Source	Citation
Formation Energy	Materials Project	[13]
Decomposition Energy	GNoME	[12]
Melting Point	MP Contribs	[14]
Shear Modulus	Materials Project	[13]
Bulk Modulus	Materials Project	[13]
Density	GNoME	[12]
Creep (LMP)	Cambridge Super Alloy Database and other sources	[15–26]
High-Temperature Oxidation	NASA and various sources	[27,28,28–44]

5.2. Data curation & outlier handling

Heterogeneous provenance, variable sample sizes (notably for creep and high-temperature oxidation [27,28]), and occasional reporting inconsistencies require a target-specific curation pipeline. In practice, each target is handled with its own source family, inclusion/exclusion criteria, outlier policy, feature-selection workflow, and validation setting.

To clear out redundancies, we compute the Pearson correlation r between all pairs of candidate descriptors, convert it to distances $d = 1 - |r|$ and apply average-linkage hierarchical clustering. Cutting the dendrogram at $|r| \geq 0.95$ yields clusters of near-duplicate descriptors (Fig. 10); we retain one representative per cluster to mitigate multicollinearity and stabilize downstream importance estimates.

In principle, these correlations are purely related to the descriptors definition and should not depend on the target. However, for targets with limited data (such as creep or high-temperature oxidation), the smaller sample size can slightly modify the dendrogram and may lead to additional merging of descriptors.

The performance of machine learning highly depends on the quality of the data – The basic assumption of machine learning is that data are distributed according to the same probability law –, and therefore statistical outliers detection is necessary. We flag suspicious records using four complementary detectors (§1.3 of SI): (i) Local Outlier Factor (density anomalies), (ii) Isolation Forest (global isolation), (iii) k-nearest-neighbor residuals against a robust trend (response inconsistencies), and (iv) Huber-regression residuals (high-influence points under a robust linear fit).

To avoid arbitrary thresholds, we evaluate the detector stability using 50 bootstrap resamples, and the treatment of the four-view outlier pipeline is property-aware (strict for creep/oxidation, moderate for well-behaved targets such as density) depending on the quality and the quantity of the raw data. For creep and high-temperature oxidation, we adopted the most conservative curation setting of the study, retaining only records not flagged by any of the four outlier detectors. This target-specific treatment is particularly important for the scarce experimental branches, where the available literature is intrinsically more heterogeneous than the large computational datasets used for the thermodynamic or elastic targets.

All curation steps (filters, cluster cuts, detector hyper-parameters, removals) are logged for reproducibility in the accompanying code archive (see **Data Availability** and **Code Availability** sections).

5.3. Descriptors & feature engineering (§1.4 of SI)

We use composition-derived, physics-informed descriptors to expose models to size, electronic, thermodynamic and periodic-trend drivers while remaining structure-agnostic (predictive for conjectured alloys). The initial panel (42 features) is engineered once from normalized atomic fractions and then specialized per target through ranking and selection. We deliberately avoid structure-dependent feature sets and learn on composition-derived, physics-meaningful descriptors only:

- **Size & packing.** Mean and spread of metallic radii; size-mismatch $\delta \equiv \sqrt{\sum_i x_i (1 - r_i/\bar{r})^2}$ (where x_i stands for the stoichiometric coefficient of element i, r_i for its atomic radius, and $\bar{r}$ is the average atomic radius); extreme radius ratios; and a virtual density proxy $d_{virt} \equiv \frac{10^7}{N_A}\frac{\overline{M}}{\bar{r}^3}$ (where $\overline{M}$ stands for the average atomic mass) capturing packing efficiency – useful to anticipate actual density.
- **Electronic structure.** Pauling/Allen electronegativities (mean, standard deviation, range), first ionization energy, electron affinity, valence-electron count (VEC) statistics; subshell-level counts where

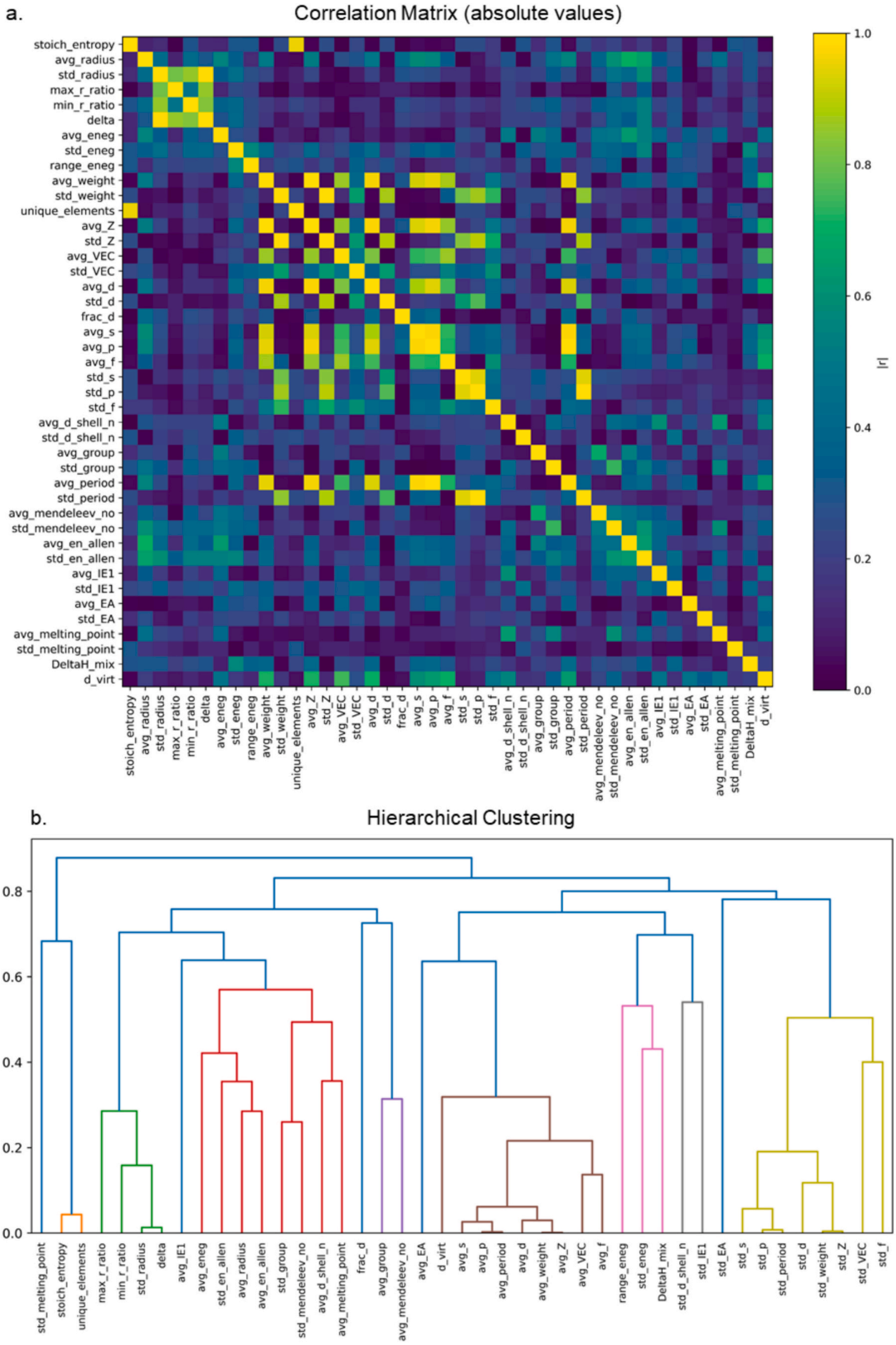


Fig. 10. Analysis of Descriptors Redundancy & Clustering (Shear Modulus) (a) Absolute correlation matrix for the full set of composition-derived descriptors. (b) Average-linkage hierarchical clustering of descriptors based on their pairwise distances; colors highlight groups of strongly correlated descriptors. The dendrogram is cut at a distance of 0.05 to define descriptor clusters used in the subsequent analyses.

available. These correlate with mixing/ordering tendencies and phase preferences.

- **Thermodynamics & entropy**. Miedema-based mixing enthalpy proxies, stoichiometric (configurational) entropy, and derived quantities (e.g., Ω for solid-solution stabilization; G/B for ductility). Predicted targets from surrogates (e.g., E_f, B, G, ρ, T_m, LMP, $\log_{10}K_A$) are computed downstream and used for scoring, not as inputs.

5.3.1. Descriptor reduction & selection process (per target)

For each target property, the descriptor set is processed in four steps (Fig. 11):

i) redundant variables are pruned by correlation-based clustering as in Fig. 10; ii) the remaining descriptors are ranked using four complementary criteria over 50 bootstrap resamples – absolute Spearman correlation (monotonic signal), mutual information (general dependence), XGBoost feature importance (nonlinear context), and RFECV with XGBoost as estimator (Fig. 11); iii) a geometric elbow is applied to each importance curve to define data-driven cut-offs; iv) the four binary selections are combined into a consensus stability score, from which descriptors above the final elbow are retained as a compact subset without imposing an arbitrary feature count.

Focusing on stoichiometric descriptors allows previously unseen chemistries to be featurized and screened without requiring known crystal structures, thereby providing broader transferability than element-identity encoding. Predictions for candidates lying far from the training manifold should nevertheless be interpreted cautiously.

5.4. Supervised ML surrogates

We trained structure-agnostic surrogates on composition-derived descriptors (Sections 5.1–5.3) to predict the target properties used downstream for screening and generative conditioning. Models are XGBoost regressors with early-stopping on a held-out validation fold, stratified shuffles to preserve target distributions, and hyper-parameters optimized on training data only to avoid leakage. Performance is reported on unseen test splits and summarized with the coefficient of determination R^2, Mean Absolute Error (MAE) and calibration diagnostics.

We trained independent surrogates for creep (LMP), high-temperature oxidation, formation enthalpy, bulk and shear moduli, melting point, and density. The creep and oxidation branches should not be interpreted symmetrically. For creep, the supervised target is the Larson–Miller parameter (LMP), i.e. a standardized scalar creep indicator in which temperature is already embedded in the target definition, and the feature-selection/training pipeline was provided with the applied stress as an explicit branch-specific variable. By contrast, the high-temperature oxidation branch remains more heterogeneous and incorporates temperature explicitly through the inverse-temperature variable $1/T$. The creep branch should therefore be interpreted as a predictor of the standardized scalar target LMP, not as a universal direct predictor of rupture life under arbitrary combinations of stress and temperature. Representative parity plots (on test sets) are provided with typical test R^2 in the $[0.8, 1]$ (see Fig. 12 as an illustration) range depending on target data quality (notably lower for oxidation due to sparse and heterogeneous measurements). A compact summary table (size, selected-feature count, outlier-threshold setting, test R^2) is also provided (Table 3). These target-specific validation results should be distinguished from the validation status of the overall inverse-design workflow. In the present study, the surrogate models are quantitatively validated on unseen test data, whereas the downstream ranking/screening layer is only benchmarked retrospectively against literature-reported HESA compositions and legacy reference alloys.

On the test base, this training reached almost $R^2 = 0.94$.

Each row in Table 3 corresponds to an independently trained surrogate associated with its own data regime; the heterogeneity challenge therefore lies in the target-specific literature or repository source, not in a direct mixing of experimental and computational labels within one model.

Among the seven supervised surrogates, the high-temperature oxidation branch is currently the least mature one, reflecting both the limited dataset size (172 entries) and the heterogeneous nature of the underlying literature measurements. Its output should therefore be interpreted as a conservative screening signal rather than as a definitive quantitative certification of oxidation behavior.

5.5. Multi-objective scoring (g-scores) & baseline filter

To rank candidates consistently across heterogeneous targets, we convert each predicted property into a desirability score $g_i \in [0, 1]$ using sigmoid or Gaussian mappings (see §3.2 of SI) anchored to literature-supported thresholds or favorable intervals. These mappings are not intended to make the underlying physical variables “equivalent”, but to express how well each predicted property satisfies an application-driven design requirement. The g_i values should therefore be interpreted as engineering desirability functions rather than absolute scores. They provide a common decision layer for screening across heterogeneous objective types – maximize, minimize, or remain within a target window – while the shortlisted candidates are still interpreted in terms of their underlying predicted physical properties. Each g_i was parameterized from a literature-supported acceptable interval or threshold. For one-sided objectives, sigmoid functions were used, with the midpoint corresponding to the transition around the acceptable limit and the width controlling how progressively the score decays away from the desired regime. For target-window objectives, Gaussian desirability functions were centered on the literature-supported favorable range and scaled so that $g_i = 0.5$ at the range boundaries. The generic construction of these mappings is illustrated in Fig. 23 of the Supplementary Information. Importantly, the workflow does not rely on a weighted-sum formulation assigning equal scalar weights to creep and oxidation. Candidates are first filtered through a uniform feasibility floor and then screened by Pareto non-domination across the full 10-dimensional desirability space. The ten objectives $\{g_1, \cdots, g_{10}\}$ cover creep resistance [46], high T_m, low density [47], empirical BCC-favoring VEC window [48], empirical δ size-mismatch control [49–53], Pugh ratio (G/B) [54], intermetallic avoidance [49], empirical solid-solution tendency (Ω) [49–52], configurational entropy ΔS_{mix} [51], and high-temperature oxidation resistance [51]. Exact thresholds and window shapes are summarized in Table 4.

This table summarizes the literature-supported thresholds or favorable intervals used to parameterize the desirability functions. These thresholds should be understood as literature-supported engineering thresholds for the present screening problem, not as uniquely validated universal constants. These g_i functions are used as a screening layer for admissibility filtering and Pareto analysis; they do not replace the underlying physical interpretation of the predicted properties.

The VEC-, δ-, and Ω-based terms should not be interpreted as universal phase-formation laws. They are used here as empirical, literature-supported heuristics that bias the search toward composition regions historically associated with favorable HEA/HESA behavior. In the present workflow, they function as soft desirability terms within a broader multi-objective screen; none of them is considered individually necessary or sufficient to guarantee the targeted phase constitution.

After computing all g_i, we apply a baseline feasibility filter requiring $g_i \geq 0.3$ for all i, ensuring no single objective catastrophically fails while retaining diversity for Pareto analysis. This rule is enforced before dominance tests and generation loops, and it is identical across brute-force, legacy and generative candidates.

Because the high-temperature oxidation branch is the most data-limited surrogate, we also examined how strongly the feasible set depends on the oxidation-related cutoff. On a 15,700-alloy baseline set,

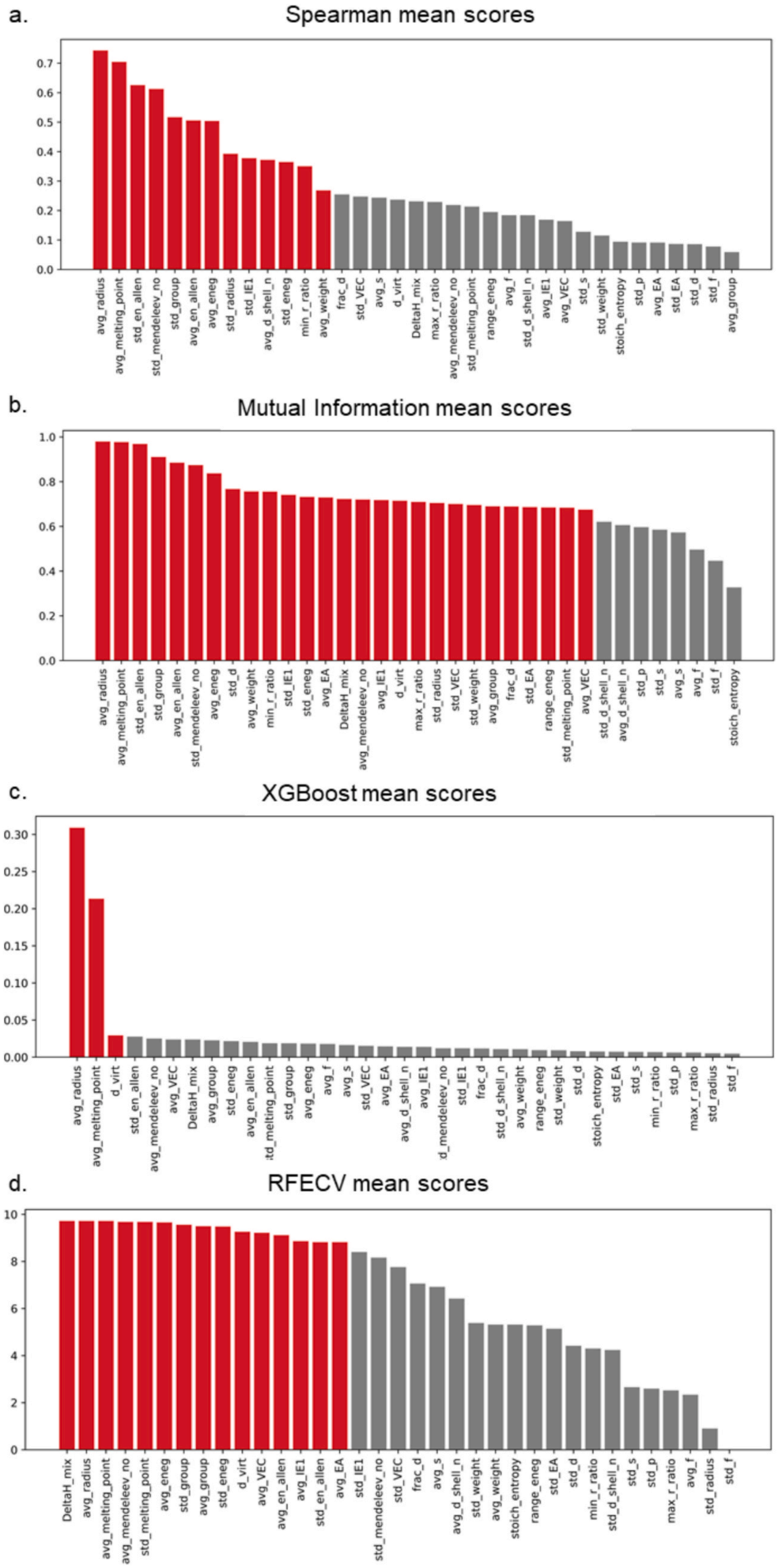


Fig. 11. Representation of the 4 Scores of the Descriptors Selected by the Clustering (Shear modulus) The bar plots represent the scores obtained by the descriptors for each selection method: (a) Spearman score, (b) Mutual Information score, (c) XGB Embedded score and (d) RFECV score. The descriptors selected by the elbow method appear in red, others appear in grey. The selection comprises the bar at which the elbow is identified. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

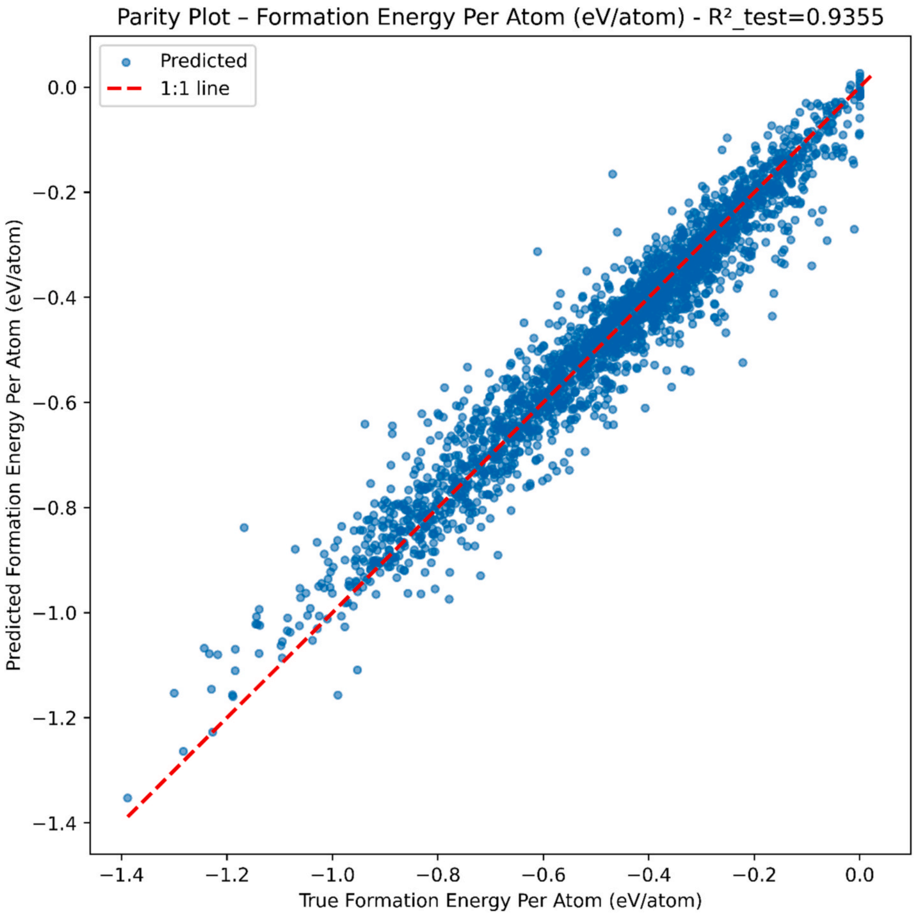


Fig. 12. Predicted Values for the Formation Enthalpy (eV / at.) versus Real Values.

Table 3

Targets Learning Summary.

Target	Source of Data	Size of dataset	Outlier score threshold	Number of selected features	R^2 reached on test dataset
Creep (LMP)	Scientific Literature [15–26] (experimental)	295	0/4	11	0.99
HT Oxidation	Scientific Literature [27,28,28–44] (experimental)	172	0/4	28	0.80
Formation Enthalpy	Materials Project (DFT)	11,061	2/4	23	0.94
Bulk Modulus	Materials Project (DFT)	1898	1/4	30	0.98
Shear Modulus	Materials Project (DFT)	1749	0/4	20	0.90
Melting Point	MPContribs (GNN)	13,880	2/4	23	0.90
Density	Deep Mind (GNN) & Materials Project (DFT)	238,787	4/4	30	1.00

109 candidates satisfy $g_1, \cdots, g_9 \geq 0.3$. Among these, 46, 39, 35, 28, 13, and 3 candidates also satisfy $g_{10} \geq 0.1,\ 0.2,\ 0.25,\ 0.3,\ 0.35,\ 0.4$, respectively. This confirms that the oxidation filter acts as a deliberately stringent selection criterion: relaxing it moderately enlarges the feasible set, whereas tightening it further sharply reduces the number of retained candidates.

Retaining demanding creep and high-temperature oxidation criteria is nevertheless a deliberate design choice in the present superalloy-oriented workflow, because the objective is not to maximize the number of admissible candidates, but to reduce a very large search space to a manageable shortlist for downstream validation.

5.6. Pareto front computation and compression

We construct the Pareto frontier on the 10-dimensional score space by marking a candidate as non-dominated if no other candidate is at least as good in all g_i and strictly better in at least one. Computation is applied to the merged dataset (stochastic brute-force screened HESA, AI generated HESA). The initial frontier can contain thousands of items; we therefore compress it for interpretability.

We represent each alloy by its CLR-encoded composition vector, perform Ward-linkage hierarchical clustering with an empirically chosen Euclidean cut in CLR-space (see Fig. 13), then extract medoids (observed compositions minimizing intra-cluster distance) as representatives. This reduces near-duplicates while preserving chemical interpretability and diversity.

The graph on the left edge shows the Within-Cluster Sum of Squares (WCSS), a classic metric to define the best value for K in K-mean clustering. This graph shows a significant gain until $K = 4$, then the decrease of WCSS slows down: this value is retained to perform the clustering.

The 3D graph on the right edge represents the 4 clusters in the latent space defined by the 3 first principal components PC1, PC2 and PC3.

The (Ward) compression of the Pareto front to a few hundred representatives suitable for Ashby-style visualization and expert review (illustration on Fig. 14).

Ashby diagram depicting density against melting point, critical

Table 4

Summary of the Physical Performance Objectives, the Quantitative Values and the Associated Filters.

SCORES	Physical Scope	Physical Quantity	Target Range	Unit	Filter Type	Sources
g_1	Creep resistance	Larson-Miller Parameter	> 23		sigmoid, $[20, 26]$	[46]
g_2	Operates at high-temperature	Melting Point	> 1500	K	sigmoid $[1200, 1800]$	[47]
g_3	Weight	Density	< 7.5		sigmoid $[6, 9]$	
g_4	Empirical BCC-favoring tendency	Valence electron per atom	$\in [6.87, 8]$	electron per atom	gaussian window	[48]
g_5	Empirical size-mismatch control	δ mismatch	$\in [0.03, 0.066]$		gaussian window	[49–53]
g_6	Mechanical behavior (ductility)	Shear / bulk moduli ratio (Pugh Criteria)	> 0.55		sigmoid $[0.5, 0.6]$	[54]
g_7	No intermetallic formation	Formation enthalpy	$[-0.156, 0.052]$	eV/atom	gaussian window	[49]
g_8	Empirical solid-solution tendency	Ω	> 1.1		sigmoid $[1.0, 1.2]$	[49–52]
g_9	High configurational entropy	Stoichiometric entropy	$[1.31, 2.32]$		gaussian window	[51]
g_{10}	Resistance to high-temperature oxidation	$\log K_A$	> 0		sigmoid $[-0.25, 0.25]$	[30,45]

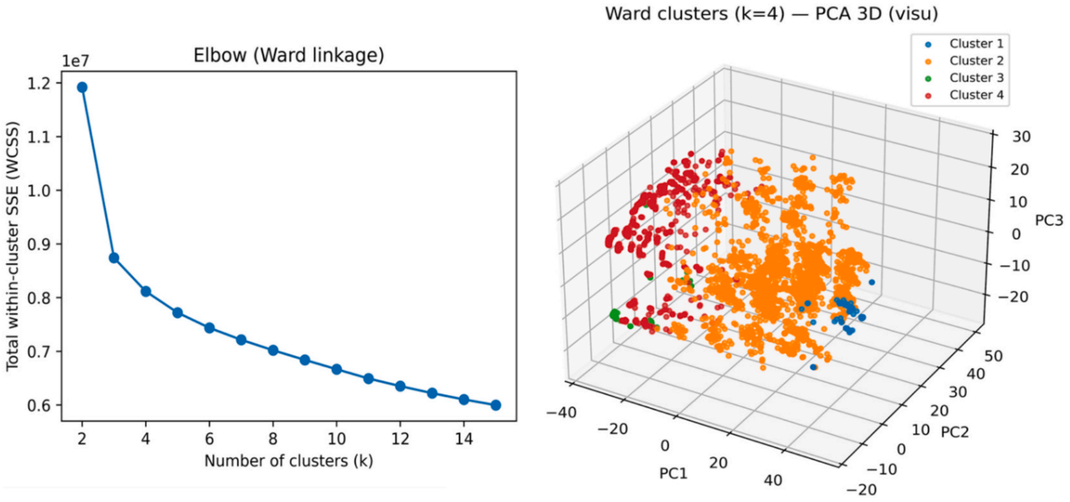


Fig. 13. Choice of K in the Ward Clustering Method.

properties for aerospace and energy applications. Markers reflect the alloy generation methodology. Generative and gradient-optimized alloys (diamonds, crosses, stars) occupy high-performance domains characterized by elevated melting temperatures with moderate-to-low densities. These generative compositions consistently exceed the melting points observed in classical superalloys, indicating novel alloys with enhanced thermostructural capabilities and potentially lighter-weight solutions for demanding applications.

The compressed Pareto set provides a compact, diverse summary of high-performing candidates, enabling clear trade-off inspection and shortlisting for experiments; examples and narratives appear in the Results section.

5.7. Generative model (CVAE) and exploration enrichments

We employ a Conditional Variational Autoencoder (CVAE) to sample plausible HEA/HESA compositions under explicit multi-objective conditions. Compositions are encoded with the centered log-ratio (CLR) transform to place mixtures in an unbiased Euclidean space; the condition vector y aggregates normalized targets (selected g-scores) with a per-score minimum and a minimum total score to avoid single-objective collapse. The network uses two dense layers in the encoder/decoder with a moderate latent size (~16), trained with a schedule that first prioritizes reconstruction, then statistical regularization (KL), and finally introduces a reward term nudging the latent toward regions decoding to high-score compositions while guarding against reconstruction drift (Fig. 15).

Given a desired condition y (e.g., class C_0), sample $z\ p(z)$ and decode $\tilde{x} = p_\varphi(x|z, y)$. The candidate $\tilde{x}$ is then screened by a classifier or rule-based filters to check membership $\tilde{x} \in C_0$ (e.g. good physical performance as defined in Table 4).

At inference, we perform iterative propose–filter–adjust cycles: sample $z\ N(0, I)$, decode under y, then apply two light pre-filters before surrogate evaluation: (i) an entropy filter (reject low configurational entropy) and (ii) a minimal-score check on CVAE-predicted scores. Survivors are evaluated by the supervised surrogates from their composition (see 5.4); retained candidates (meeting all conditions) are deduplicated and accumulated. Batch-level statistics (entropy, mean scores, survival share) steer the reward schedule epoch by epoch.

To systematically diversify the explored composition space and support the CVAE in generating varied candidates, we complement CVAE sampling with:

- Hamming-based discrete augmentation over the presence/absence mask (fixed cardinality, five-element alloys), selecting binary “words” far from previously explored sets; and
- gradient-guided stoichiometric refinement (finite-difference gradients with simplex projection) starting from promising Pareto candidates to climb toward higher multi-objective scores despite non-

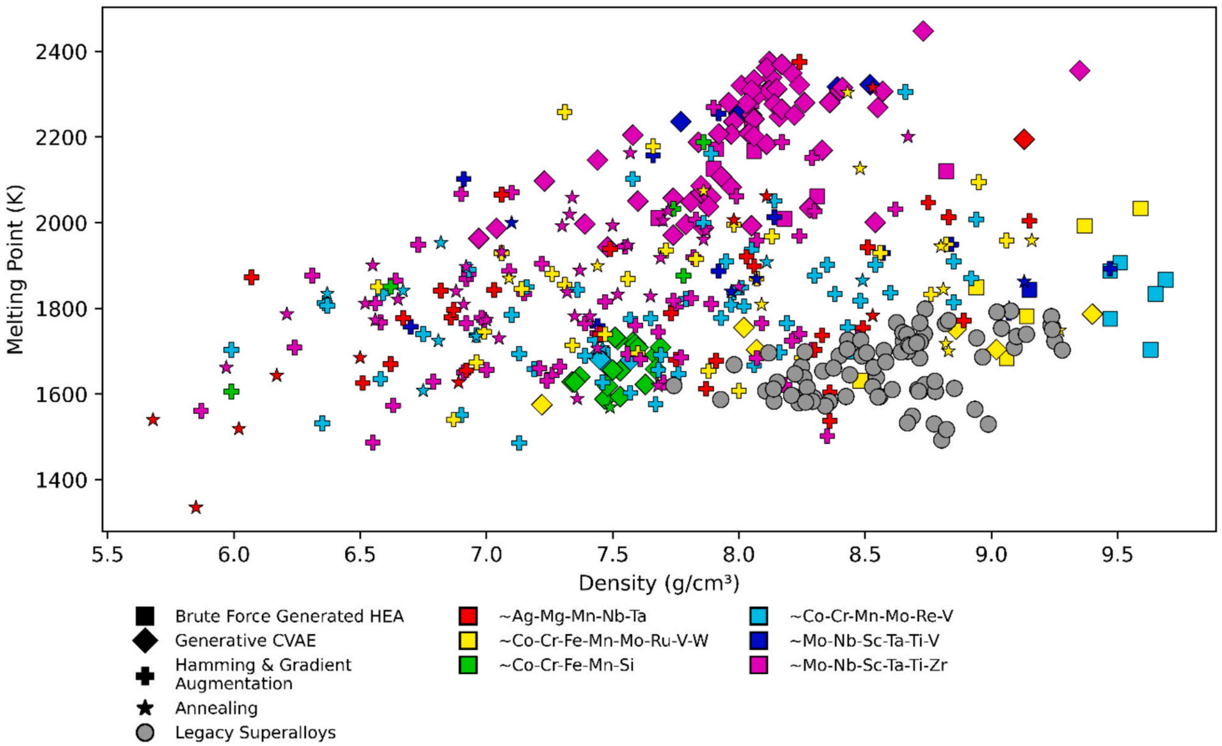


Fig. 14. Density vs. Melting Point.

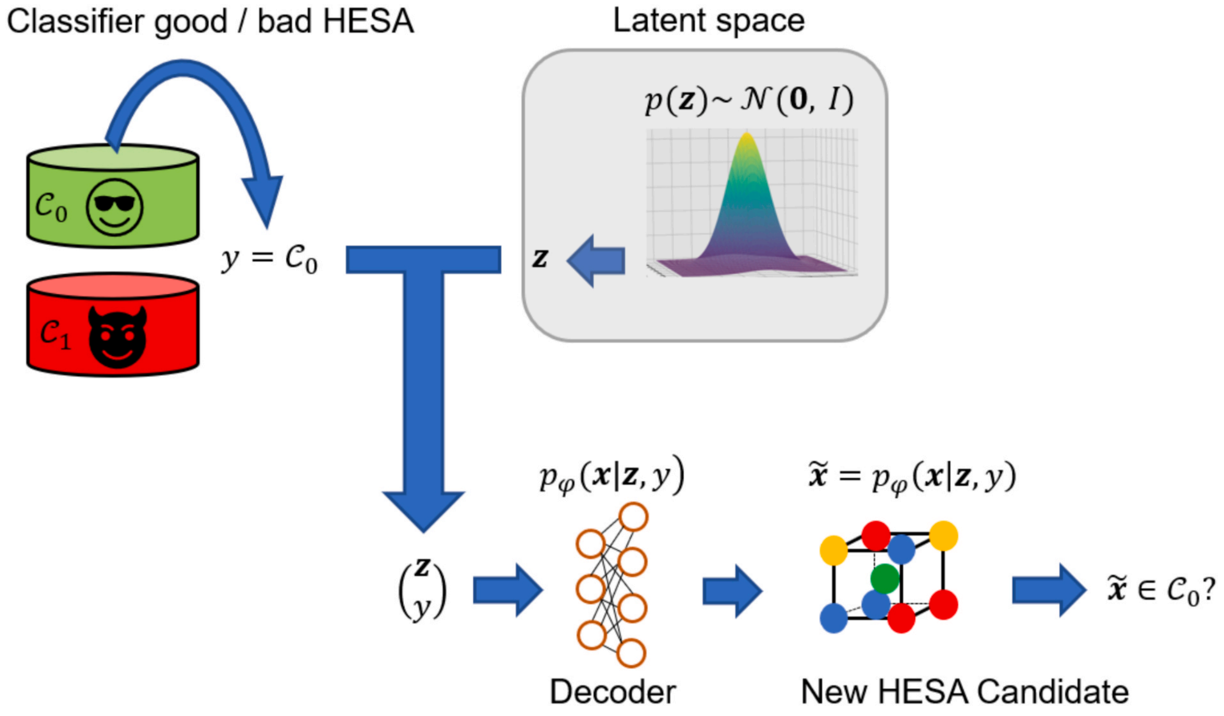


Fig. 15. Conditional VAE at inference (generation).

differentiabilities. These steps produced improved candidates while maintaining feasibility.

The learning database is enriched with the improved candidates, and the generative model is run again. The overall candidates pool is then compressed via Ward linkage and medoid selection before Pareto analysis and visualization.

6. Code availability

The complete code of the inverse design pipeline, including the training scripts and sample data, is publicly available on GitHub under https://github.com/fr55160/materials_inverse_design_pipeline, except for the largest files which are publicly available at https://doi.org/10.5281/zenodo.17566991 and https://doi.org/10.5281/zenodo.175

67098.

7. Declaration of generative AI use

During the preparation of this work, the author(s) used ChatGPT (OpenAI) as an AI-assisted tool in order to (i) survey and compare candidate mathematical methods relevant to the study, (ii) assist with translation, (iii) improve clarity and wording of passages, and (iv) draft and debug analysis scripts. After using this tool, the author(s) reviewed, verified, and edited the content and code as needed, and take full responsibility for the content of the published article.

Funding Declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

François Rousseau: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Thierry Belmonte:** Writing – review & editing, Validation, Supervision. **Frédéric Sur:** Writing – review & editing, Validation, Supervision, Methodology. **Alexandre Nominé:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.matdes.2026.116097.

Data availability

Full training datasets are deposited on Zenodo under https://doi.org/10.5281/zenodo.17567098, except for the largest files which are publicly available at https://doi.org/10.5281/zenodo.17566991 and https://doi.org/10.5281/zenodo.17567098.

References

[1] Shewmon, Paul G., Gill, Charles Burroughs, Charles, James A., Lorig, Clarence H., Metallurgy - Copper Alloy, Zinc Alloy, Brass | Britannica, Encyclopedia Britannica (2024). https://www.britannica.com/science/metallurgy (accessed May 1, 2025).

[2] Materials Selection in Mechanical Design, 2010. https://shop.elsevier.com/books/materials-selection-in-mechanical-design/ashby/978-0-08-095223-9 (accessed May 1, 2025).

[3] A. Valizadeh, R. Sahara, M. Souissi, Alloys innovation through machine learning: a statistical literature review, Sci. Technol. Adv. Mater.: Methods 4 (2024), https://doi.org/10.1080/27660400.2024.2326305.

[4] A.A. Deshmukh, R. Ranganathan, Recent advances in modelling structure-property correlations in high-entropy alloys, J. Mater. Sci. Technol. 204 (2025) 127–151, https://doi.org/10.1016/j.jmst.2024.03.027.

[5] V. Verma, C.H. Belcher, D. Apelian, E.J. Lavernia, Diffusion in high entropy alloy systems – a review, Prog. Mater Sci. 142 (2024) 101245, https://doi.org/10.1016/j.pmatsci.2024.101245.

[6] W. Chen, A. Hilhorst, G. Bokas, S. Gorsse, P.J. Jacques, G. Hautier, A map of single-phase high-entropy alloys, Nat. Commun. 14 (2023) 2856, https://doi.org/10.1038/s41467-023-38423-7.

[7] A. Muhammad, H. Minouei, N. Park, Review of high-entropy superalloys: design, microstructure, and properties, Mater. Chem. Phys. 351 (2026) 131962, https://doi.org/10.1016/j.matchemphys.2025.131962.

[8] B.J. Watson, R.G. Eggert, Understanding relative metal prices and availability: combining physical and economic perspectives, J. Ind. Ecol. 25 (2020) 890–899, https://doi.org/10.1111/jiec.13087.

[9] Z.J.N. Steinmann, A.M. Schipper, M. Hauck, S. Giljum, G. Wernet, M.A. J. Huijbregts, Resource footprints are good proxies of environmental damage, Environ. Sci. Technol. 51 (2017) 6360–6366, https://doi.org/10.1021/acs.est.7b00698.

[10] USGS, Mineral commodity summaries 2024, USGS, Reston, VA, 2024. https://doi.org/10.3133/mcs2024.

[11] E. Commission, E. Directorate-General for Internal Market Industry, SMEs, D. Pennington, E. Tzimas, C. Baranzelli, J. Dewulf, S. Manfredi, P. Nuss, M. Grohol, A. Van Maercke, Y. Kayam, S. Solar, B. Vidal-Legaz, L. Talens Peirò, L. Mancini, C. Ciupagea, L. Godlewska, P. Dias, C. Pavel, D. Blagoeva, G. Blengini, V. Nita, C. Latunussa, C. Torres De Matos, F. Mathieux, A. Marmier, Methodology for establishing the EU list of critical raw materials – Guidelines, Publications Office, 2017. https://doi.org/10.2873/769526.

[12] A. Merchant, S. Batzner, S.S. Schoenholz, M. Aykol, G. Cheon, E.D. Cubuk, Scaling deep learning for materials discovery, Nature 624 (2023) 80–85, https://doi.org/10.1038/s41586-023-06735-9.

[13] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: the Materials Project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (2013) 011002, https://doi.org/10.1063/1.4812323.

[14] Q.-J. Hong, S.V. Ushakov, A. van de Walle, A. Navrotsky, Melting temperature prediction using a graph neural network model: from ancient minerals to new materials, in: Proceedings of the National Academy of Sciences 119, 2022, https://doi.org/10.1073/pnas.2209630119.

[15] D.O.B. Apriandanu, Y. Yulizar, CuO-bentonite-gold nanocomposites: facile green preparation and their characterization, Mater. Lett. 284 (2021) 128911, https://doi.org/10.1016/j.matlet.2020.128911.

[16] Defense Technical Information Center, DTIC ADA065740: Handbook of International alloy Compositions and Designations. Volume II. Superalloys, 1978. http://archive.org/details/DTIC_ADA065740 (accessed May 14, 2025).

[17] S. Gorsse, M.H. Nguyen, O.N. Senkov, D.B. Miracle, Database on the mechanical properties of high entropy alloys and complex concentrated alloys, Data Brief 21 (2018) 2664–2678, https://doi.org/10.1016/j.dib.2018.11.111.

[18] S. Li, P.J. Withers, S. Kabra, K. Yan, The behaviour and deformation mechanisms for 316L stainless steel deformed at cryogenic temperatures, Mater. Sci. Eng. A 880 (2023) 145279, https://doi.org/10.1016/j.msea.2023.145279.

[19] J. Liu, Y. Li, H. Ju, Y. Wang, C. Yan, L. Zhang, Z. Liu, S. Li, Microstructures and mechanical properties of additively manufactured T91 steel and T91/316H bimetallic components by laser powder bed fusion, Mater. Sci. Eng. A 892 (2024) 145277, https://doi.org/10.1016/j.msea.2023.145277.

[20] X. Liu, Z. Chen, Y. Chen, S. Yang, Y. Pan, Y. Lu, S. Qu, Y. Li, Y. Yang, C. Wang, Multicomponent Co-Ti-based superalloy with high solvus temperature and low lattice misfit, Mater. Lett. 284 (2021) 128910, https://doi.org/10.1016/j.matlet.2020.128910.

[21] K. Matsugi, Y. Murata, M. Morinaga, N. Yukawa, Realistic advancement for nickel-based single crystal superalloys by the d-electrons concept, Superalloys (1992) 307–316.

[22] Y. Ogawa, O. Takakuwa, K. Tsuzaki, Solid-solution hardening by hydrogen in Fe–Cr–Ni-based austenitic steel: temperature and strain rate effects, Mater. Sci. Eng. A 879 (2023) 145281, https://doi.org/10.1016/j.msea.2023.145281.

[23] B. Ohl, L. Owen, H. Stone, D.C. Dunand, Microstructure and mechanical properties of L12-strengthened Co–Ni–Fe-based superalloys, Mater. Sci. Eng. A 884 (2023) 145276, https://doi.org/10.1016/j.msea.2023.145276.

[24] D. Pu, X. Chen, J. Wang, J. Tan, J. Li, H. Yang, B. Feng, K. Zheng, F. Pan, Effect of Ti particles on the microstructure and mechanical properties of as-extruded Titanium-reinforced Mg-9Gd-4Y–1Zn–1Mn matrix composites, Mater. Sci. Eng. A 879 (2023) 145278, https://doi.org/10.1016/j.msea.2023.145278.

[25] P. Taylor, Machine learning approach to model the microstructure and strength of nickel superalloys, PhD Thesis, Apollo - University of Cambridge Repository, 2022. https://doi.org/10.17863/CAM.101577.

[26] M. Yang, C. Huang, Z. Yao, Z. Yang, Z. Li, C. Wang, H. Wu, T. Yang, Y. Zhao, S. Wang, X. Liu, Development of a high-strength Fe–12Mn maraging steel via designing lath interfacial and intragranular nanostructures, Mater. Sci. Eng. A 886 (2023) 145280, https://doi.org/10.1016/j.msea.2023.145280.

[27] C.A. Barrett, A statistical analysis of elevated temperature gravimetric cyclic oxidation data of 36 Ni- and Co-base superalloys based on an oxidation attack parameter, NASA, Lewis Research Center, Cleveland, Ohio, 1992.

[28] C.D. Taylor, B.M. Tossey, High temperature oxidation of corrosion resistant alloys from machine learning, npj Mater. Degrad. 5 (2021) 38, https://doi.org/10.1038/s41529-021-00184-3.

[29] K.A. Rozman, M. Detrois, P.D. Jablonski, M.C. Gao, J.A. Hawk, Effect of Mn in CrCoFeNiMn high entropy alloy on creep performance, (2020). https://doi.org/10.1007/s11665-020-05103-2.

[30] Y. Sun, J. Ni, Machine learning advances in high-entropy alloys: a mini-review, Entropy 26 (2024), https://doi.org/10.3390/e26121119.

[31] S. Guo, C. Ng, J. Lu, C.T. Liu, Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, J. Appl. Phys. 109 (2011) 103505, https://doi.org/10.1063/1.3587228.

[32] L. Tan, K. Ali, P. Ghosh, A. Arya, Y. Zhou, R. Smith, P. Goddard, D. Patel, H. Shahmir, A. Gandy, Design Principles of low-activation High Entropy Alloys, J. Alloy. Compd. 907 (2022) 164526, https://doi.org/10.1016/j.jallcom.2022.164526.

[33] P. Chauhan, S. Chopra, S. Thangaraju, Inter-dependency relationships in high-entropy alloys: phase stability criteria, Adv. Eng. Mater. 21 (2019) 1900251, https://doi.org/10.1002/adem.201900251.

[34] R.K. Nutor, Q. Cao, X. Wang, D. Zhang, Y. Fang, Y. Zhang, J.-Z. Jiang, Phase selection, lattice distortions, and mechanical properties in high-entropy alloys, Adv. Eng. Mater. 22 (2020) 2000466, https://doi.org/10.1002/adem.202000466.

[35] X. Yang, Y. Zhang, Prediction of high-entropy stabilized solid-solution in multi-component alloys, Mater. Chem. Phys. 132 (2012) 233–238, https://doi.org/10.1016/j.matchemphys.2011.11.021.

[36] Y. Zhang, W. jie Peng, Microstructural control and properties optimization of high-entropy alloys, Procedia Engineering 27 (2012) 1169–1178. https://doi.org/10.1016/j.proeng.2011.12.568.

[37] S. Kamran, K. Chen, L. Chen, Ab initio examination of ductility features of fcc metals, Phys. Rev. B 79 (2009) 024106, https://doi.org/10.1103/PhysRevB.79.024106.

[38] Z. Grzesik, Oxidation of pure metals, (2018).

[39] Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition., (n.d.). https://hastie.su.domains/ElemStatLearn/ (accessed February 2, 2025).

[40] N. Babu, R. Balasubramaniam, A. Ghosh, High-temperature oxidation of Fe3Al-based iron aluminides in oxygen, Corros. Sci. 43 (2001) 2239–2254, https://doi.org/10.1016/S0010-938X(01)00035-X.

[41] C.A. Barrett, C.E. Lowell, Resistance of Ni-Cr-Al alloys to cyclic oxidation at 1100 and 1200°C, Oxid. Met. 11 (1977) 199–223, https://doi.org/10.1007/BF00606544.

[42] K.A. Christofidou, N.G. Jones, M.C. Hardy, H.J. Stone, The Oxidation Behaviour of Alloys based on the Ni–Co–Al–Ti–Cr System, Oxid. Met. 85 (2016) 443–458, https://doi.org/10.1007/s11085-015-9606-8.

[43] S. Cruchley, Sun ,J. F., Taylor ,M. P., Evans ,H. E., Bowen ,P., Sumner ,J., Nicholls , J. R., Simms ,N. J., Shollock ,B. A., Chater ,R. J., Foss ,B. J., Hardy ,M. C., S. and Stekovic, Cautionary note on use of focused ion beam sectioning as technique for characterising oxidation damage in Ni based superalloys, Materials at High Temperatures 31 (2014) 27–33. https://doi.org/10.1179/0960340913Z.0000000004.

[44] Y. Gui, Z. Liang, H. Shao, Q. Zhao, Corrosion behavior and lifetime prediction of VM12, Sanicro 25 and Inconel 617 in supercritical carbon dioxide at 600 °C, Corros. Sci. 175 (2020) 108870, https://doi.org/10.1016/j.corsci.2020.108870.

[45] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67, https://doi.org/10.1080/00401706.1970.10488634.

[46] P.Y. Hou, Impurity effects on alumina scale growth, J. Am. Ceram. Soc. 86 (2003) 660–668, https://doi.org/10.1111/j.1151-2916.2003.tb03355.x.

[47] P. Huczkowski, D.J. Young, T. Olszewski, A. Chyrkin, W.J. Quadakkers, Effect of sulphur on the oxidation behaviour of possible construction materials for heat exchangers in oxyfuel plants in the temperature range 550–700 °C, Oxid. Met. 89 (2018) 651–681, https://doi.org/10.1007/s11085-017-9809-2.

[48] R.P. Oleksak, J.H. Tylczak, C.S. Carney, G.R. Holcomb, Ö.N. Doğan, High-temperature oxidation of commercial alloys in supercritical CO2 and related power cycle environments, JOM 70 (2018) 1527–1534, https://doi.org/10.1007/s11837-018-2952-7.

[49] B.A. Pint, More ,K. L., I.G. and Wright, The use of two reactive elements to optimize oxidation performance of alumina-forming alloys, Materials at High Temperatures 20 (2003) 375–386. https://doi.org/10.1179/mht.2003.044.

[50] B.A. Pint, Unocic ,K. A., Brese ,R. G., J.R. and Keiser, Characterization of chromia scales formed in supercritical carbon dioxide, Materials at High Temperatures 35 (2018) 39–49. https://doi.org/10.1080/09603409.2017.1389371.

[51] A. Sato, Y.-L. Chiu, R.C. Reed, Oxidation of nickel-based single-crystal superalloys for industrial gas turbine applications, Acta Mater. 59 (2011) 225–240, https://doi.org/10.1016/j.actamat.2010.09.027.

[52] S.-L. Shang, Y. Wang, B. Gleeson, Z.-K. Liu, Understanding slow-growing alumina scale mediated by reactive elements: perspective via local metal-oxygen bonding strength, Scr. Mater. 150 (2018) 139–142, https://doi.org/10.1016/j.scriptamat.2018.03.002.

[53] J.L. Smialek, A. Garg, T.P. Gabb, R.A. MacKay, Cyclic oxidation of high Mo reduced density superalloys, , Metals 5 (2015) 2165–2185, https://doi.org/10.3390/met5042165.

[54] C. Yu, J. Zhang, D.J. Young, High temperature corrosion of Fe-Cr-(Mn/Si) alloys in CO2-H2O-SO2 gases, Corros. Sci. 112 (2016) 214–225, https://doi.org/10.1016/j.corsci.2016.07.026.