

Machine learning prediction of young's modulus in multi component titanium based biomedical alloys using extended thermodynamic descriptors

Hassan Ahmad^a, Muhammad Haider^a, Zafar Iqbal^a, Muhammad Zarif^a, Syed Mujtaba ul Hassan^{a*}

^a*Department of Metallurgy and Materials Engineering, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad 45650, Pakistan.*

*Corresponding author. Email: mujtabanaqvi29@gmail.com

Abstract

The development of low-modulus titanium alloys for biomedical implants is frequently constrained by the resource-intensive nature of experimental discovery and the limited size of available datasets. To address this, this study presents a machine learning framework trained on a comprehensive dataset of 689 alloy compositions, extending beyond conventional systems to include multicomponent alloys described using thermodynamic descriptors borrowed from the high-entropy alloy literature. By integrating physicochemical and thermodynamic descriptors, an optimized XGBoost model was developed to predict Young's modulus. The model achieved a test R^2 of 0.69 and a test MAE of 9.67 GPa. However, predictive accuracy is comparatively lower in the low-modulus regime (below 50 GPa), which set it for the primary target range for implant applications. External validation against 24 independent alloys was performed to assess model performance across diverse chemical spaces. Feature importance analysis revealed that configurational mixing entropy and molybdenum equivalence are critical determinants of stiffness and phase stability. These results demonstrate that integrating thermodynamic descriptors into composition-based models improves predictive capability and may support preliminary, coarse-grained screening of candidate low-modulus biomedical alloys prior to experimental synthesis.

KEYWORDS: Titanium alloys; Machine learning; Biomedical alloys; Young's modulus prediction

Impact Statement

This study presents a machine learning model trained on a large dataset of Ti based biomedical alloys to predict Young's modulus. External validation shows that the approach can support faster screening of new alloy compositions.

1. Introduction

Titanium alloys have emerged as the predominant choice for orthopedic implant applications due to their exceptional combination of high specific strength, superior corrosion resistance, and favourable biocompatibility [1], [2], [3]. Despite these advantageous properties, the conventional Ti-6Al-4V alloy, which has served as the gold standard for implants and prostheses for decades, presents a critical limitation: its elevated Young's modulus of 110 GPa [4], [5]. This value substantially exceeds that of human cortical bone, which typically exhibits a Young's modulus below 35 GPa [5], [6]. The pronounced elastic modulus mismatch between the implant material and the host bone tissue creates mechanical incompatibility, ultimately manifesting as the "stress shielding effect" a phenomenon that precipitates abnormal remodeling and subsequent degradation of the surrounding bone tissue [3], [5], [6].

Compounding these mechanical concerns, the constituent elements of Ti-6Al-4V, particularly vanadium (V) and aluminum (Al), have well-documented cytotoxic effects [3], [4], [5], and both have been linked to neurodegenerative pathology, including Alzheimer's disease. This has made the development of cleaner, lower modulus alloys a practical as well as mechanical priority. The challenge is partly economic: the most effective β -stabilizers, namely niobium (Nb), molybdenum (Mo), and tantalum (Ta), are expensive, which limits how freely they can be used in alloy design [4], [7].

Meeting these requirements simultaneously, that is, reduced modulus, acceptable biocompatibility, and reasonable cost, calls for more systematic approaches to alloy discovery. Machine learning has emerged as a practical route here [7], [8], [9]. Conventional alloy development relies heavily on iterative experimentation, which is slow and resource-intensive [10], [11], [12], data-driven methods offer a way to compress that process by learning property trends directly from existing experimental records [13], [14].

The persistent bottleneck for ML models in this space is dataset size. Published studies on Ti-based biomedical alloys have consistently worked with small training sets: Wu *et al.* [15] used 164 compositions for Young's modulus prediction and 112 for martensitic start temperature; an XGBoost-based study of Ti-Mo-Nb-Zr-Sn alloys trained on just 82 entries [16]; and most other efforts have used between 96 and 342 alloys [17], [18]. Even the most carefully curated compilation available covers only 282 multicomponent Ti alloys [19], especially for the complex multicomponent compositions that are increasingly of interest [18], [20], [21].

It is worth being explicit about the scope of the present work. The dataset assembled here is intentionally broad, encompassing 689 alloy compositions that span not only established biomedical systems but also alloys containing elements such as Al, V, Cr, Cu, Mn, Fe, W, and Si, which carry cytotoxicity concerns in clinical contexts. These alloys were included to improve model coverage across chemical space, not as recommendations for implant use. The framework described here is a property-prediction tool; questions of biocompatibility are a separate filter to be applied downstream. The primary novelty of this work is threefold: construction of the large dataset for ML modeling of Ti-based alloys in this application domain, integration of thermodynamic HEA-inspired descriptors into the feature space, and systematic external validation against 24 independently sourced alloys drawn from

chemically diverse families, including refractory multicomponent systems absent from prior training sets.

2. Materials and Methodology

2.1. Dataset Composition and Feature Engineering

The dataset consists of 689 entries, compiled and expanded from experimental literature on titanium-based alloys. Each entry corresponds to a unique alloy composition, containing up to 15 alloying elements. The elemental compositions include Ti, Nb, Zr, Fe, Sn, Ta, Mo, Al, V, Cr, Cu, Hf, Mn, W, Sc, and Si.

The distribution of Young's modulus across the compiled dataset is presented in Figure 1. The majority of alloys are concentrated in the 50 to 100 GPa range with 74.90 GPa Median, 0.74 skewness, with only 51 alloys (7.6% of entries) falling below 50 GPa. This apparent inclusion of low-modulus alloys reflects the current state of the experimental literature rather than a sampling bias, as compositions in this regime including well-characterized systems such as Ti-Nb-Zr-Sn are already experimentally established [6], [15]. The inclusion of intermediate to high modulus alloys provides the model with the contrast necessary to learn the compositional determinants of low-modulus behavior, enabling screening of novel compositions rather than mere interpolation within already characterized systems.

A known limitation of datasets compiled in the literature is heterogeneity in thermomechanical processing conditions. Alloys with nominally identical compositions can exhibit different Young's modulus values depending on whether they were solution-treated, cold-worked, or aged, because YM is sensitive to crystallographic texture developed during processing. To partially address this, alloy phase type (α , $\alpha+\beta$, β /metastable β) was included as a categorical feature, since phase constitution reflects, at least in part, the thermal history

of an alloy. Processing variability is nonetheless expected to contribute to prediction uncertainty for YM, and this limitation is shared across published ML studies on titanium alloys that rely on composition only frameworks [15], [17], [18], [20]. It should be noted that no two entries in the dataset share identical elemental compositions, confirmed by exact matching across all element columns, so the dataset does not contain duplicate records with differing processing histories for the same alloy. Consequently, composition only descriptors cannot fully capture processing dependent variation in Young's modulus, and thermomechanical features represents an irreducible source of prediction uncertainty in the present framework.

For feature selection, we included a comprehensive set of physicochemical, empirical, and thermodynamic descriptors derived from the alloy composition: atomic radius (r), Pauling and Allred-Rochow electronegativity (χ^P , χ^A), valence-electron concentration (VEC), electron-to-atom ratio (e/a), and various mismatch parameters (δr , $\delta\chi_P$, $\delta\chi_A$, $\Delta\chi_P$, $\Delta\chi_A$, ΔVEC). We also included thermodynamic features that capture chemical complexity and phase stability: configurational mixing entropy (ΔS_m), mixing enthalpy (ΔH_{ic}), and solid-solution entropies for bcc and fcc phases (S_{xs}^{bcc} , S_{xs}^{fcc}), as well as lattice-distortion parameters (γ , ϕ_{bcc} , ϕ_{fcc}). A complete list and mathematical definitions of all descriptors are summarized in Table 1.

Although titanium alloys are predominantly bcc (β phase) or hcp (α phase), the fcc excess entropy (S_{xs}^{fcc}) was retained as a comparative descriptor. It is not included on the assumption that Ti alloys form fcc structures; rather, it is borrowed from the high-entropy alloy literature, where it is used alongside its bcc counterpart to quantify the relative thermodynamic affinity for different close-packed structures. Used together give the model a contrast signal for phase-selection tendencies across the compositional space, consistent with established

practice in HEA descriptor development [22]. The dataset contains 421 metastable- β -type (61%), 158 $\alpha / \alpha+\beta$ (23%), and 110 stable- β -type (16%) alloys, with no fcc entries, which confirms that the inclusion of S_{xs}^{fcc} is purely as a discriminating descriptor rather than a structural assumption.

We also computed the molybdenum-equivalent (Mo_{eq}) for each composition as a compositional proxy for β -phase stability in Ti alloys. Mo_{eq} converts the contents of common β -stabilizers (e.g., V, Nb, Ta, Mo, Fe, Cr) to an equivalent Mo content via element-specific weighting factors and has long been used in titanium alloy design to estimate β stability and β -transus trends [23]. Modern formulations further refine Mo_{eq} by fitting the $\beta/(\alpha+\beta)$ phase-boundary slopes, improving correlation with phase stability [10]. Because β stability and phase constitution strongly affect mechanical properties, we included Mo_{eq} and the categorical phase label as inputs to support phase-aware learning.

To understand the relationships between the selected features and the target mechanical properties, we performed a Pearson correlation analysis to identify the most influential input variables. A correlation matrix was generated to visualize the linear relationships between all descriptors and the Young's modulus. A Pearson correlation heatmap across all descriptors is provided in the Supplementary Material (Figure S1) to identify multicollinearity patterns and redundant features. Feature importance was further assessed using SHAP (SHapley Additive exPlanations) analysis, which quantitatively ranks descriptors by their contribution to model predictions, consistent with established practice in materials informatics [24].

2.2. Machine Learning Model Development

The dataset was split into training and testing subsets at a 75:25 ratio, with approximately 517 alloy systems allocated for training. All features were scaled to the [0, 1] range using MinMax normalization to prevent high-magnitude variables from dominating the learning

process. A fixed random seed of 42 was used throughout to ensure reproducibility [25].

2.3. Hyperparameter Optimization

Default hyperparameters are rarely adequate for capturing complex structure property relationships in domain-specific datasets, so a systematic tuning protocol was applied. Grid search and randomized search were combined with k-fold cross-validation to explore the hyperparameter space. This step was especially important for the tree-based models (Random Forest, Gradient Boosting, and XGBoost), where the balance between complexity and generality is sensitive to parameter choice. The resulting configurations, selected to minimize the loss function while controlling overfitting [26], are listed in Table 2.

2.4. Visualization of Results

To examine the relationships between the selected descriptors and Young's modulus, a Pearson correlation heatmap was constructed to identify linear associations between all input variables and the target property. Shapley additive explanations analysis was then applied to move beyond linear associations. A mean absolute SHAP bar chart ranks descriptor by their overall contribution to model predictions, while the SHAP summary plot reveals the directionality of each feature's influence.

3. Results and Discussion

3.1. Model Optimization

Several regression models were trained and evaluated for Young's modulus prediction, spanning linear methods (Linear Regression, Ridge, Lasso, ElasticNet), a nearest neighbor approach (KNN), and tree-based ensembles including Random Forest, Gradient Boosting, and XGBoost.

The performance metrics for all models are reported in Table 3. XGBoost achieved the best test performance, with a test MAE of 9.67 GPa and a test R^2 of 0.69. Gradient Boosting followed closely (test MAE = 9.88 GPa, R^2 = 0.68), and Random Forest also performed competitively (test MAE = 10.57 GPa, R^2 = 0.64). The gap between training and test scores for K-Nearest Neighbors (train R^2 = 1.00, test R^2 = 0.58) points to substantial overfitting, making it unreliable despite its low training error. Linear models performed poorly across the board, with test R^2 values around 0.37–0.38, consistent with their known inability to capture the nonlinear composition–property relationships in multicomponent alloy systems.

The advantage of tree-based ensembles over linear approaches is consistent with prior work in materials informatics, where XGBoost and gradient boosted methods have shown strong performance on structure property modeling tasks [27]. Their iterative boosting mechanism reduces both bias and variance while naturally handling nonlinearities and correlated features, a useful property given the collinearity among the thermodynamic descriptors used here [28].

We also tested a stacked ensemble combining the top three models. Stacking produced only marginal gains over XGBoost alone, consistent with observations in lattice structure mining [29], suggesting that added complexity is not warranted when a single well-tuned model already generalizes adequately [27].

Figure 2 shows predicted versus observed Young’s modulus values for the top-performing models. Training points cluster tightly along the ideal line, while test points show somewhat wider scatter, consistent with the test R^2 of 0.69. Deviations are spread across the full modulus range without any directional trend, indicating the model does not systematically over- or underestimate in any regime. The line plot in Figure 3 traces predicted and actual values of sample alloys from the database, and the two curves follow each other closely throughout the range, with larger disagreements concentrated at extreme data values.

3.2. Feature Importance

Pearson analysis in Figure 4 reveals a more nuanced picture than conventional descriptor rankings might suggest. Iron (Fe) emerges as the single strongest positive correlate with Young's modulus, which, while perhaps counterintuitive at first glance, is physically consistent with Fe acting as a potent β -stabilizer: its incorporation tends to shift alloy compositions toward higher-VEC, denser electronic configurations that, in the multicomponent systems represented here, are associated with elevated stiffness rather than suppressed modulus. Niobium, by contrast, shows one of the strongest negative correlations $r \approx 0.34$, in full agreement with its well-documented role as a β -stabilizer that reliably suppresses Young's modulus in Ti-based alloys [15], [30].

The spread in valence electron concentration ΔVEC carries a moderate positive correlation $r \approx 0.31$, suggesting that compositional heterogeneity in electron count tends to accompany higher modulus values across the assembled alloy space. Configurational mixing entropy (ΔS_m), defined as,

$$\Delta S_m = -R \sum_i x_i \ln x_i$$

where R is the universal gas constant and x_i is the mole fraction of element i [20], shows a negative correlation with Young's modulus $r \approx 0.27$. Alloys with greater chemical complexity and higher ΔS_m values more frequently belong to multicomponent β -Ti families, where the combined stabilizing contributions of multiple alloying elements collectively suppress stiffness. This negative sign underscores the importance of treating ΔS_m as a composite phase-selection signal rather than a simple thermodynamic scalar.

The inter correlation pattern between the most important descriptors is presented in Figure 5. Fe and $(\text{Mo})_{\text{eq}}$ have high co-correlation ($r = 0.67$) due to their both relying on the level of β -

stabilizers, whereas Nb shows negative correlation with both ($r = -0.18$ to -0.34). ΔS_m and $\delta\chi^p$ are least correlated with other descriptors, supporting their role as independent composite signals in the model.

The molybdenum equivalent, $(Mo)_{eq}$, ranks among the stronger positive correlates $r \approx 0.24$. Although higher $(Mo)_{eq}$ values denote alloys chemically enriched in β -stabilizers, the relationship between β -stabilizer content and modulus is non-monotonic across the full compositional space [30]. Notably, the average atomic radius $\bar{r}$ is negatively correlated with Young's modulus $r \approx 0.22$, contrary to the geometric intuition that larger atoms stiffen the lattice. This inversion reflects the fact that the largest-radius alloying elements in this dataset, namely Nb, Zr, and Ta, are precisely the β -stabilizers responsible for driving modulus downward, so $\bar{r}$ functions here as an indirect compositional proxy rather than a pure structural stiffness indicator.

The SHAP analysis in Figure 6 extends these linear observations into the nonlinear regime, confirming and in some cases reordering the descriptor rankings from Pearson analysis. Fe and ΔVEC retain the top two positions in both analyses, indicating their influence is genuinely dominant rather than an artifact of linear correlation. Nb correspondingly appears as a high-impact feature in the SHAP beeswarm plot, where low Nb values shift predictions toward higher modulus and high Nb values consistently push predictions downward, in direct agreement with its well-documented role as a modulus-suppressing β -stabilizer [19]. Sn shows a tightly clustered SHAP distribution centered near zero, reflecting its compositionally narrow range across the dataset and its secondary role as a neutral or mildly stabilizing addition. Overall, the agreement between Pearson and SHAP rankings across the top descriptors supports the physical interpretability of the model rather than mere statistical fitting.

3.3. External Validation

To evaluate the capabilities of the model, external validation was performed on 24 titanium-based alloys taken entirely from the literature [17], [19], [20], [30], [31] and not included in the training set. The compositions were selected to cover a broad chemical space: classical Ti-Nb-Zr-Sn systems, Fe-stabilized alloys, refractory-inspired multicomponent systems, and a high-modulus Ti-Ta-Cu composition. This range provides a genuine test of transferability rather than interpolation within already sampled regions of compositional space.

Predicted and experimental Young's modulus values are given in Table 4. The parity plot and residual errors in Figure 7 show that across the full modulus range of 42–107 GPa, predictions sit close to the ideal line with no directional drift, and residuals fall on both sides of zero. Performance is strongest where training coverage is good. The Ti-Nb-Ta series (alloys 21–23) yields differences of 3.9, 6.8, and 8.8 GPa, and Alloy 24 ($\text{Ti}_{75}\text{Ta}_{20}\text{Cu}_5$, 107 GPa measured) is predicted at 97.3 GPa, a 9.7 GPa gap consistent with behavior elsewhere in the high-modulus range. In the Fe-stabilized series, Alloys 11 and 12 are within 7.6 and 3.2 GPa of their measured values, while Alloys 13–14 drift to around 11 GPa error, probably because incremental Sn additions in that range are thinly represented in the training data.

Larger deviations occur for the high-Zr refractory multicomponent alloys (alloys 16–19 in Table 4), where the model overestimates by 13–16 GPa. These systems have Zr contents of 40–45% alongside Hf, Mo, and Ta and sit at the compositional boundary of the training distribution. Incorporating additional descriptors that capture Zr-dominated lattice behavior or augmenting the dataset with more high-Zr entries, would likely reduce these deviations in future work.

Overall, the validation confirms that model predictions are reliable across chemically diverse Ti alloy families. Predictions for most alloys fall within 15 GPa of the experimental values,

and no systematic over or underestimation is observed across the full modulus range from 42 to 107 GPa.

4. Conclusion

We have presented an approach demonstrating that a machine learning framework trained on a large and diverse dataset with extended features can predict Young's modulus across a wide range of novel and complex Ti-based alloys, though with comparatively lower accuracy in the low-modulus regime that is most relevant to implant design. The model's robustness, validated externally, confirms that these inclusions are crucial for developing reliable predictive tools in materials science. Future studies should incorporate microstructural and processing descriptors to bring the framework closer to reliable targeted modulus prediction, the point at which a model becomes genuinely useful for guiding alloy design before any experiment is run.

Acknowledgements

The authors gratefully acknowledge the Department of Metallurgy and Materials Engineering at Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan, for providing research facilities.

Disclosure Statement

The authors declare no competing interests.

Funding

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

Data Availability Statement

The data and code supporting the findings of this study are openly available on Zenodo at <https://doi.org/10.5281/zenodo.20393039>.

References

- [1] M. Niinomi, “Recent metallic materials for biomedical applications,” *Metall. Mater. Trans. A*, vol. 33, no. 3, pp. 477–486, 2002.
- [2] J. Wolff, *The law of bone remodelling*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986.
- [3] R. Huiskes, H. Weinans, and B. Rietbergen, “The relationship between stress shielding and bone resorption around total hip stems and the effects of flexible materials,” *Clin. Orthop. Relat. Res.*, no. 274, pp. 124–134, 1992.
- [4] M. Niinomi and M. Nakai, “Titanium-based biomaterials for preventing stress shielding between implant devices and bone,” *Int. J. Biomater.*, vol. 2011, p. 836587, 2011.
- [5] C. Oldani and A. Dominguez, “Titanium as a biomaterial for implants,” in *Recent advances in arthroplasty*, S. Fokter, Ed., INTECH Open Access Publisher, 2012, pp. 149–162.
- [6] Y. L. Hao, S. J. Li, S. Y. Sun, and others, “Elastic deformation behaviour of Ti-24Nb-4Zr-7.9Sn for biomedical applications,” *Mater. Sci. Eng. A*, vol. 441, no. 1–2, pp. 112–118, 2006.
- [7] M. Geetha, A. K. Singh, R. Asokamani, and others, “Ti based biomaterials, the ultimate choice for orthopaedic implants, a review,” *Prog. Mater. Sci.*, vol. 54, no. 3, pp. 397–425, 2009.
- [8] B. Dipankar and J. C. Williams, “Perspectives on titanium science and technology,” *Acta Mater.*, vol. 61, no. 3, pp. 844–879, 2013.

- [9] M. J. Jackson, G. M. Robinson, M. A. Ali, and others, “Titanium and titanium alloy applications in medicine,” in *Surgical tools and medical devices*, M. J. Jackson and W. Ahmed, Eds., Cham (Switzerland): Springer, 2016, pp. 475–517.
- [10] Q. Wang, C. Dong, and P. K. Liaw, “Structural stabilities of β -Ti alloys studied using a new Mo equivalent derived from $[\beta/(\alpha+\beta)]$ phase-boundary slopes,” *Metall. Mater. Trans. A*, vol. 46, no. 8, pp. 3440–3447, 2015.
- [11] M. Niinomi, M. Nakai, and J. Hieda, “Development of new metallic alloys for biomedical applications,” *Acta Biomater.*, vol. 8, no. 11, pp. 3888–3903, 2012.
- [12] T. Ozaki, H. Matsumoto, S. Watanabe, and others, “Beta Ti alloys with low Young’s modulus,” *Mater. Trans.*, vol. 45, no. 8, pp. 2776–2779, 2004.
- [13] X. Li, X. Liu, S. Wu, and others, “Design of magnesium alloys with controllable degradation for biomedical implants: From bulk to surface,” *Acta Biomater.*, vol. 45, pp. 2–30, 2016.
- [14] Y. Xin, T. Hu, and P. K. Chu, “In vitro studies of biomedical magnesium alloys in a simulated physiological environment: A review,” *Acta Biomater.*, vol. 7, no. 4, pp. 1452–1459, 2011.
- [15] C. T. Wu, D. H. Yang, J. C. Lin, and others, “Machine learning recommends affordable new Ti alloy with bone-like modulus,” *Mater. Today*, vol. 34, pp. 41–50, 2020.
- [16] F. Yang, Z. Li, Q. Wang, and others, “Cluster-formula-embedded machine learning for design of multicomponent β -Ti alloys with low Young’s modulus,” *npj Comput. Mater.*, vol. 6, p. 101, 2020.
- [17] G. Marković, V. Manojlović, J. Ružić, and others, “Predicting low-modulus biocompatible titanium alloys using machine learning,” *Materials*, vol. 16, no. 19, p. 6355, 2023.

- [18] L. C. Zhang and L. Y. Chen, “A review on biomedical titanium alloys: recent progress and prospects,” *Adv. Eng. Mater.*, vol. 21, no. 4, p. 1801215, 2019.
- [19] C. A. F. Salvador, E. L. Maia, F. H. Costa, and others, “A compilation of experimental data on the mechanical properties and microstructural features of Ti-alloys,” *Sci. Data*, vol. 9, no. 1, p. 188, 2022.
- [20] X. Liu, Q. Peng, S. Pan, and others, “Machine learning assisted prediction of microstructures and Young’s modulus of biomedical multi-component Ti alloys,” *Metals (Basel)*., vol. 12, no. 5, p. 796, 2022.
- [21] C. Chai, Y. Liang, S. Liu, and others, “Machine learning-assisted design of low elastic modulus β -type medical titanium alloys and experimental validation,” *Comput. Mater. Sci.*, vol. 238, p. 112902, 2024.
- [22] A. R. Miedema, P. F. de Châtel, and F. R. de Boer, “Cohesion in alloys fundamentals of a semi-empirical model,” *Physica B+C*, vol. 100, no. 1, pp. 1–28, 1980.
- [23] P. J. Bania, “Beta titanium alloys and their role in the titanium industry,” *JOM*, vol. 46, no. 7, pp. 16–19, 1994.
- [24] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.
- [25] F. Pedregosa and others, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [26] J. Bergstra, R. Bardenet, Y. Bengio, and others, “Algorithms for hyper-parameter optimization,” in *Advances in Neural Information Processing Systems*, 2011, pp. 2546–2554.
- [27] D. Boldini, F. Grisoni, D. Kuhn, and others, “Practical guidelines for the use of gradient boosting for molecular property prediction,” *J. Cheminform.*, vol. 15, no. 1, p. 73, 2023.

- [28] S. Feng, H. Zhao, H. Liu, and others, "Prediction of the mechanical properties of high-entropy alloys based on ensemble machine learning," *J. Mater. Res. Technol.*, vol. 18, pp. 3189–3199, 2022.
- [29] Y. Wang, J. Xiong, Q. Wu, and others, "Mining the mechanical properties of lattice structures by a stacking ensemble learning framework," *Mater. Des.*, vol. 220, p. 110884, 2022.
- [30] C. T. Wu, C. Y. Chang, C. H. Kuo, and others, "Revisiting alloy design of low-modulus biomedical β -Ti alloys using an artificial neural network," *Materialia (Oxf.)*, vol. 21, p. 101313, 2022.
- [31] S. Kwak, J. Kim, H. Ding, and others, "Machine learning prediction of the mechanical properties of γ -TiAl alloys produced using random forest regression model," *Results Mater.*, vol. 13, p. 100315, 2022.
- [32] S. Fang, X. Xiao, L. Xia, and others, "Relationship between the widths of supercooled liquid regions and bond parameters of Mg-based bulk metallic glasses," *J. Non-Cryst. Solids*, vol. 321, no. 1–2, pp. 120–125, 2003.
- [33] Y. Zhang, Y. J. Zhou, J. P. Lin, and others, "Microstructures and properties of high-entropy alloys," *Prog. Mater. Sci.*, vol. 61, pp. 1–93, 2014.
- [34] O. N. Senkov, G. B. Wilks, D. B. Miracle, and others, "Refractory high-entropy alloys," *Intermetallics (Barking)*, vol. 18, no. 9, pp. 1758–1765, 2010.
- [35] L. Pauling, *The nature of the chemical bond*, 3rd ed. Ithaca (NY): Cornell University Press, 1960.
- [36] L. C. Allen, "Electronegativity is the average one-electron energy of the valence-shell electrons in ground-state free atoms," *J. Am. Chem. Soc.*, vol. 111, no. 25, pp. 9003–9014, 1989.

- [37] M. G. Poletti and L. Battezzati, “Electronic and thermodynamic criteria for the occurrence of high entropy alloys in metallic systems,” *Acta Mater.*, vol. 75, pp. 297–309, 2014.
- [38] S. Guo, C. Ng, J. Lu, and others, “Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys,” *J. Appl. Phys.*, vol. 109, no. 10, p. 103505, 2011.
- [39] T. B. Massalski and U. Mizutani, “Electronic structure of Hume-Rothery phases,” *Prog. Mater. Sci.*, vol. 22, pp. 151–262, 1978.
- [40] A. Takeuchi and A. Inoue, “Classification of bulk metallic glasses by atomic size difference, heat of mixing and period of constituent elements and its application to characterization of the main alloying element,” *Mater. Trans.*, vol. 46, no. 12, pp. 2817–2829, 2005.
- [41] X. Yang and Y. Zhang, “Prediction of high-entropy stabilized solid-solution in multi-component alloys,” *Mater. Chem. Phys.*, vol. 132, no. 2–3, pp. 233–238, 2012.
- [42] Z. Wang, Y. F. Gao, M. C. Gao, and others, “A new parameter to predict the phase formation of high-entropy alloys,” *Scr. Mater.*, vol. 94, pp. 28–31, 2015.
- [43] Y. F. Ye, Q. Wang, J. Lu, and others, “A unified parameter for predicting the formation of solid solution phases in high entropy alloys,” *Scr. Mater.*, vol. 104, pp. 53–55, 2015.

Figures & Tables

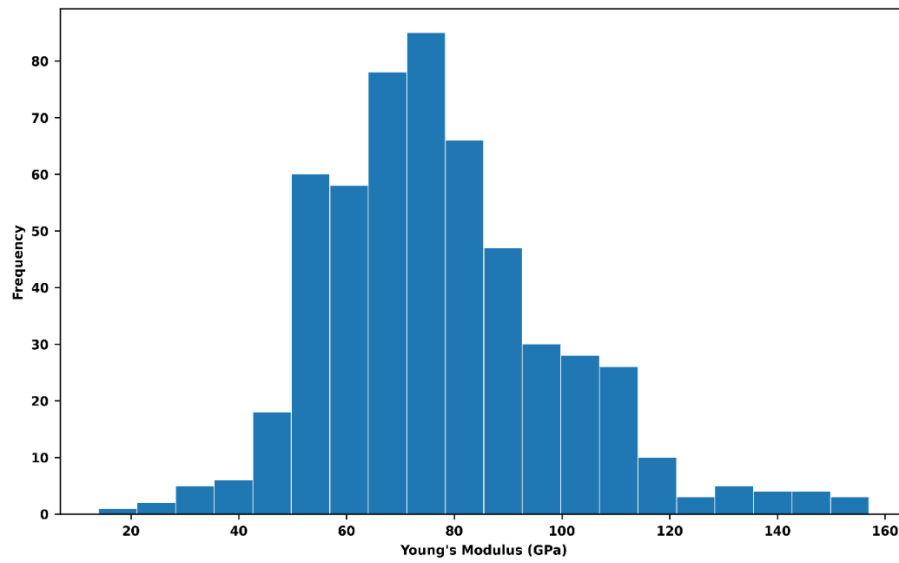


Figure 1. Frequency distributions of Young's Modulus in database.

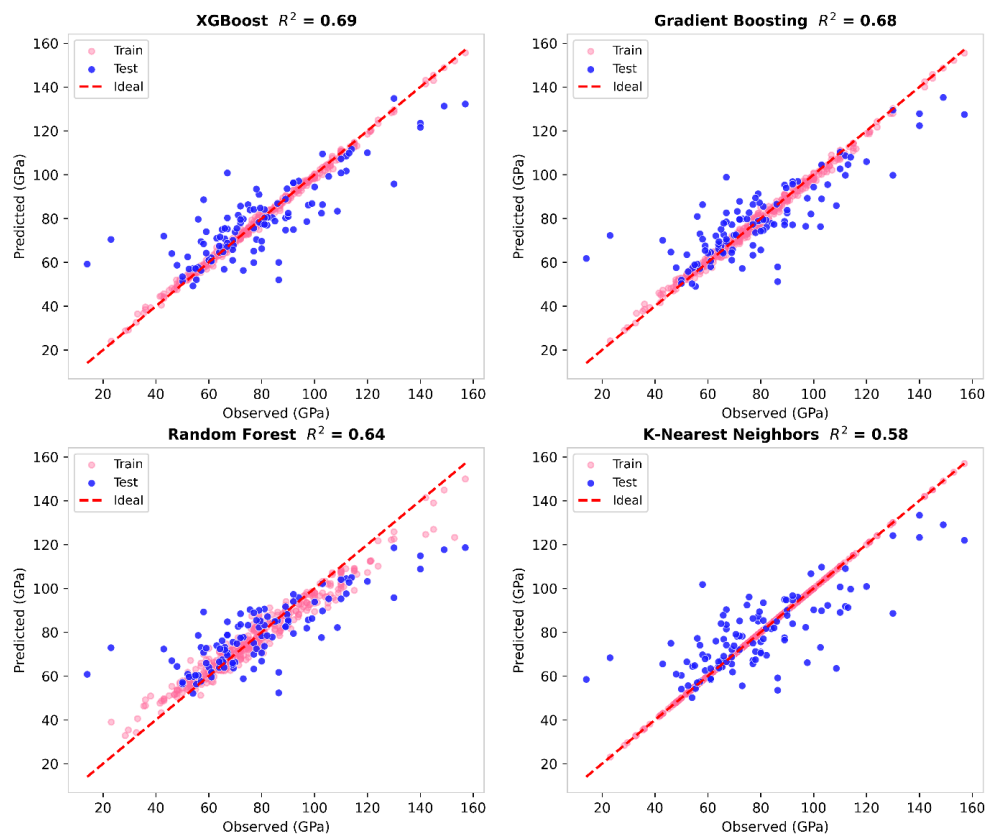


Figure 2. Top performing model's predicted vs actual values for Young's modulus.

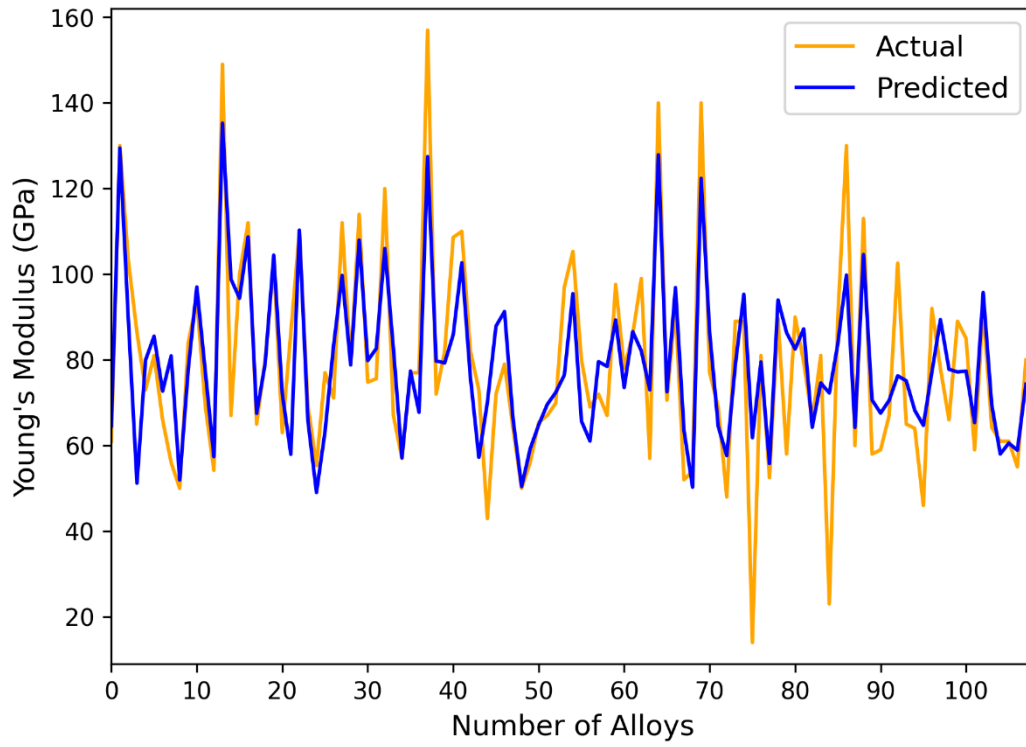


Figure 3. Model prediction accuracy and behavior across the training dataset.

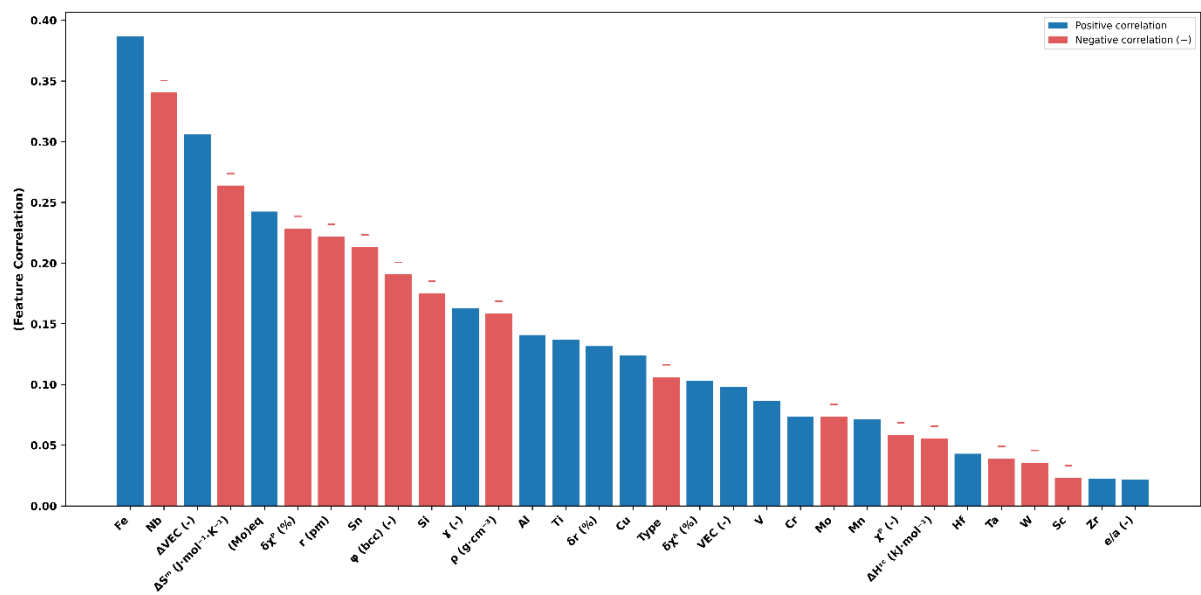


Figure 4. Pearson correlation coefficients of input descriptors with Young's modulus.

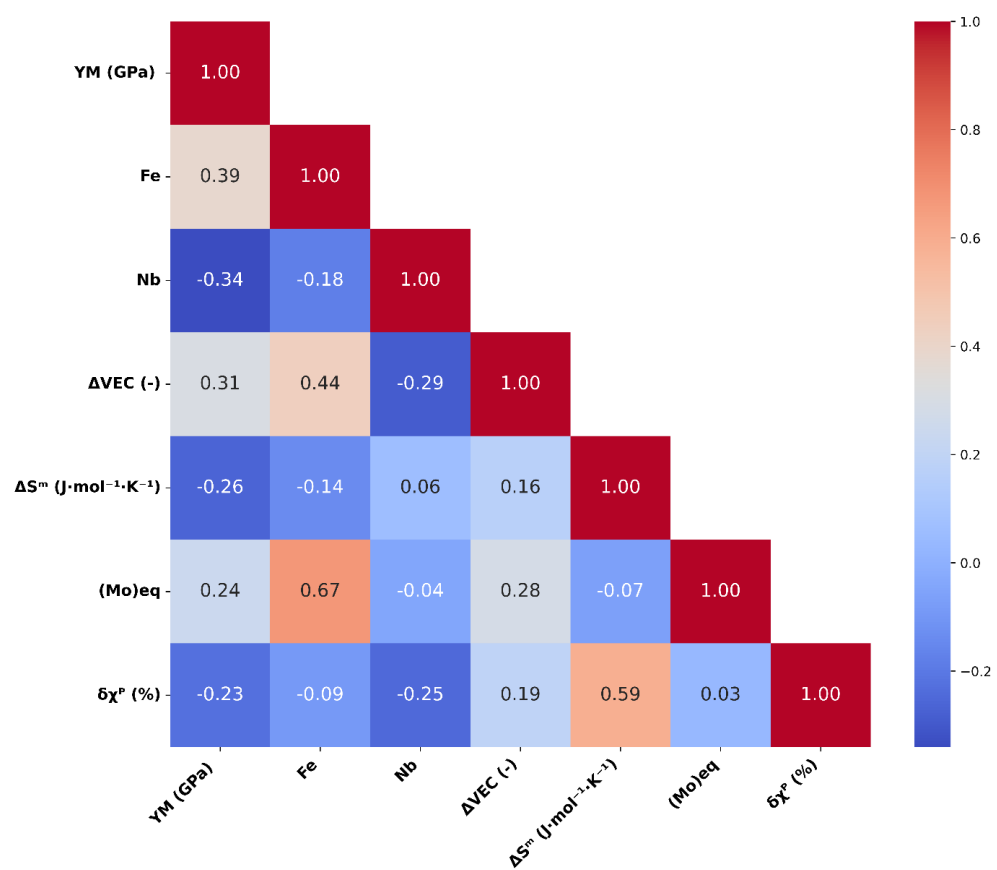


Figure 5. Correlation heatmap of top contributing descriptors with Young's modulus.

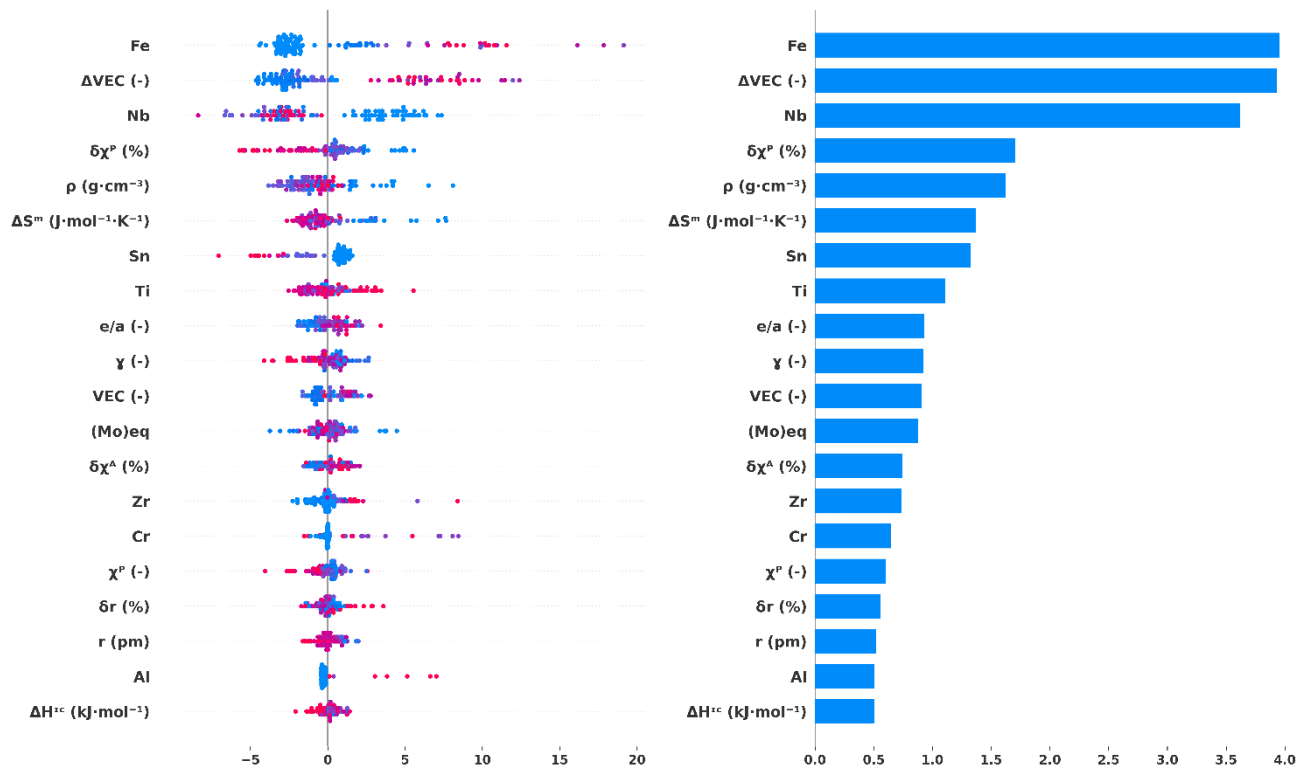


Figure 6. SHAP beeswarm summary plot and feature importance for Young's modulus prediction.

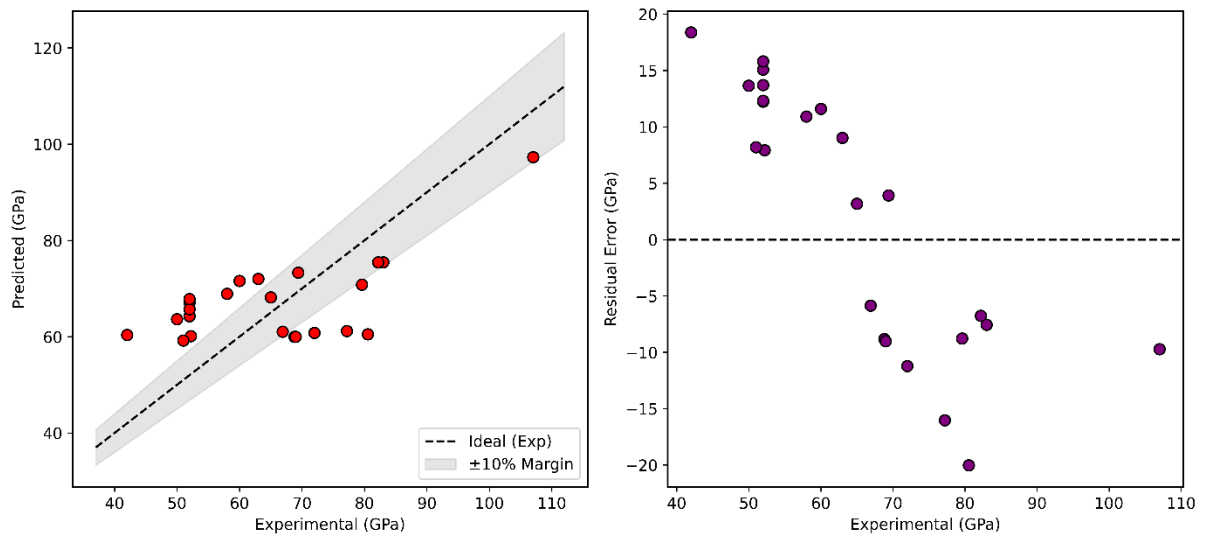


Figure 7. Parity plot and residual error plot for 24 validated alloys.

Table 1. Description of features used in the machine learning model.

Feature	Formula	Description	Ref
YM (GPa)	—	Young's modulus (elastic stiffness)	—
$\bar{\rho}$ (g/cm ³)	$= \sum x_i \rho_i$	Density by rule of mixtures	[32]
Thc (W/m·K)	$= \sum x_i (Thc)_i$	Thermal conductivity by rule of mixtures	[33]
$\bar{r}$ (Å)	$= \sum x_i r_i$	Average atomic radius	[34]
χ^P (—)	$= \sum x_i \chi_i^P$	Average Pauling electronegativity	[35]
χ^A (—)	$= \sum x_i \chi_i^A$	Average Allen electronegativity	[36]
$\Delta\chi^P$ (—)	$= \sqrt{\sum x_i (\chi_i^P - \bar{\chi}^P)^2}$	Std dev of Pauling electronegativity	[32]
$\delta\chi^A$ (—)	$= \sqrt{\sum x_i \left(1 - \frac{\chi_i^A}{\bar{\chi}^A}\right)^2}$	Electronegativity mismatch (Allen)	[37]
ΔVEC (—)	$= \sqrt{\sum x_i (VEC_i - \bar{VEC})^2}$	Spread in valence electron concentration	[38]
VEC (—)	$= \sum x_i VEC_i$	Average valence electron concentration	[38]
e/a (—)	$= \sum x_i (e/a)_i$	Average electrons per atom	[39]
ΔSm (J/mol·K)	$= -R \sum x_i \ln x_i$	Ideal configurational mixing entropy	[40]
ΔHm (kJ/mol)	$= \sum_{i \neq j} x_i x_j \cdot 4\Delta H_{ij}$	Enthalpy of mixing (Miedema-based)	[22]
Ω (—)	$= \frac{T_m \cdot \Delta Sm_{mix}}{ \Delta H_{mix} }$	Parameter for solid solution formation	[41]
γ (—)	$= \frac{1 - \sqrt{(\min(r_i) + \bar{r})^2 - \bar{r}^2}}{1 - \sqrt{(\max(r_i) + \bar{r})^2 - \bar{r}^2}}$	Lattice distortion parameter	[42]
$\varphi(\text{bcc})$ (—)	$= \frac{\Delta Sm_{mix}}{\delta r^2}$	Phase stability parameter (bcc)	[43]
$\varphi(\text{fcc})$ (—)	$= \frac{\Delta Sm_{mix}}{\delta r^2}$	Phase stability parameter (fcc)	[43]

Note: All summation terms use x_i to represent the mole fraction of element i .

Table 2. Optimized hyperparameters for the machine learning models.

Hyperparameter	Random Forest	Gradient Boosting	XGBoost
<i>n_estimators</i>	500	300	300
<i>learning_rate</i>	—	0.05	0.05
<i>max_depth</i>	20	5	6
<i>subsample</i>	—	0.8	0.8
<i>random_state</i>	42	42	42

Table 3. Performance metrics MAE and R² for Young's Modulus predictions.

Model	Train MAE	Test MAE	Train R ²	Test R ²
XGBoost	1.01	9.67	0.9963	0.69
Gradient Boosting	1.51	9.88	0.9928	0.68
Random Forest	3.90	10.57	0.9410	0.64
K-Nearest Neighbors	0.00	11.51	1.0000	0.58
Linear Regression	12.31	13.58	0.4404	0.38
ElasticNet Regression	12.54	13.82	0.4235	0.38
Lasso Regression	12.51	13.82	0.4270	0.37
Ridge Regression	12.39	13.70	0.4352	0.37

Table 4. External validation of predicted Young's modulus for Ti-based alloys.

No.	Composition (wt%)	Exp. (GPa)	Pred. (GPa)	Difference
1	Ti62-Nb16-Zr16-Sn6	72	60.784	11.216
2	Ti62-Nb10-Zr18-Sn10	77.2	61.175	16.025
3	Ti64-Nb6-Zr12-Sn12-Mo6	68.8	59.980	8.820
4	Ti63-Nb15-Zr6-Sn10-Ta6	52.2	60.121	-7.921
5	Ti66-Nb19-Ta5-Zr1-Sn9	80.54	60.524	20.016
6	Ti61-Nb13-Ta12-Zr10-Sn4	66.91	61.039	5.871
7	Ti64-Nb12-Zr12-Sn12	42	60.365	-18.365
8	Ti62-Nb14-Zr12-Sn12	51	59.188	-8.188
9	Ti67-Nb6-Mo3-Zr12-Sn12	52	64.232	-12.232
10	Ti64-Nb6-Mo6-Zr12-Sn12	69	59.980	9.020
11	Ti72-Nb26-Fe2	83	75.439	7.561
12	Ti70-Nb26-Fe2-Sn2	65	68.175	-3.175
13	Ti68-Nb26-Fe2-Sn4	58	68.912	-10.912
14	Ti66-Nb26-Fe2-Sn6	60	71.594	-11.594
15	Ti64-Nb26-Fe2-Sn8	63	72.018	-9.018
16	Ti45-Zr40-Nb4-Sn1-Mo5-Ta4-Hf1	52	67.072	-15.072
17	Ti50-Zr40-Nb4-Sn1-Mo2-Ta2-Hf1	52	64.313	-12.313
18	Ti50-Zr40-Nb6-Sn1-Mo1-Ta2	52	65.697	-13.697
19	Ti45-Zr45-Nb4-Sn1-Mo2-Ta2-Hf1	50	63.655	-13.655
20	Ti79.1-Mo15-Al3-Nb2.7-Si0.2	52	67.809	-15.809
21	Ti90-Nb5-Ta5	69.4	73.324	-3.924
22	Ti88-Nb7-Ta5	82.2	75.439	6.761
23	Ti82-Nb13-Ta5	79.6	70.816	8.784
24	Ti75-Ta20-Cu5	107	97.276	9.724

