Title
A physics-informed feature engineering approach to use machine learning with limited amounts of data for alloy design: shape memory alloy demonstration

Authors
Sen Liu,1 Branden B. Kappes,1 Behnam Amin-ahmadi,1 Othmane Benafan,2 Aaron P. Stebner,1* Xiaoli Zhang1*

Affiliations
1Mechanical Engineering, Colorado School of Mines, Golden, CO 80401 USA
2Materials and Structures Division, NASA Glenn Research Center, Cleveland, OH 44135 USA
*Corresponding author: astebner@mines.edu, xlzhang@mines.edu

Abstract
Machine learning using limited data from physical experiments is shown to work to predict new shape memory alloys in a high dimensional alloy design space that considers chemistry and thermal post-processing. The key to enabling the machine learning algorithms to make predictions of new alloys and their post-processing is shown to be a physics-informed feature engineering approach. Specifically, elemental features previously engineered by the computational materials community to model composition effects in materials are combined with newly engineered heat treatment features. These new features result from pre-processing the heat treatment data using mathematical relationships known to describe the thermodynamics and kinetics of precipitation in alloys. The prior application of the nonlinear physical models to the data in effect “linearizes” some of the complex alloy development trends a priori using known physics, and results in greatly improved performance of the ML algorithms trained on relatively few data points.

1. Introduction
Predictive design of the performances of alloys based upon their processing, structure, and properties (the so-called “process-structure-property-performance” paradigm [1]) is a time-consuming process because of high-dimensional design space and relevant physics that span length scales of $10^{-10}$ m, the length scale of atomic bond, to $10^6$ m, the length scale of metallic components, and time scales of $10^{-14}$ s, the time scale of atomic vibrations, to $10^7$ s, the time scale of aging and corrosion. Decades of global research and development initiatives such as Integrated Computational Materials Engineering (ICME) [2][3] and the Materials Genome Initiative (MGI) [4] have demonstrated the ability for both physics-based and data-driven computations to accelerate the discovery and deployment of new alloys. It is established that machine learning (ML) can model process-structure-property relationships of alloys [5][6]. Of equal or greater impact, ML can greatly reduce the number of physics-based experiments and calculations needed to discover and design new materials with optimal properties [7][8][9]. However, the robust prediction of a new alloy and its processing designed to meet a desired, yet not previously achieved performance remains an open challenge; one that is met in this work.

In other sects of materials science and engineering where new materials have been successfully predicted, the formulation of effective data descriptors, or “feature engineering,” has emerged as a critical data pre-processing step to enable better performances from ML. Most such studies have focused on using high-throughput physics-based calculations together with chemical element descriptors to assist ML prediction [7][9]. For example, density functional theory (DFT)
calculations have been used to generate large amounts of data, which are then organized and indexed within materials databases such as Aflowlib [10], OQMD [11], and materialsproject.org [12], which are then mined to build ML models capable of predicting the properties of new compounds. Recently, the development of data descriptors has been the key to data-driven models to predict the glass-forming ability of metallic glasses [13], band gap energy of thermoelectric [14], formation enthalpy of semiconductors [15], properties of inorganic crystals [16], critical temperatures of superconductors [17][18], and the structures and band gaps of Heusler compounds [19][20] and perovskites [21]. While these methodologies have proven the promise of using data-driven ML for materials design, these models were constructed using descriptors of data from first-principle calculations or basic element descriptors such as bond energies or atomic radii of elements. Recently, the formulation and integration of new thermodynamic descriptors that consider both entropy and enthalpy, such as the “entropy density of states,” has led to breakthroughs in the discovery of ultra-high temperature ceramics [22]. Still, the thermodynamic data that these descriptors are based upon are generated in large amounts using CALculation of PHase Diagrams (CALPHAD) approaches in combination with DFT.

A vast frontier of discovery and development still remains in moving beyond DFT and CALPHAD generated databases for ML-informed materials discovery and development. Alloys are one material class where this is especially true; most engineering alloys are composed of three or more elements, with the most prolific engineering alloy class, steels, often having 8 or more critical alloying additions and impurities dictating their behaviors [23]. Today, DFT calculations of ternary alloys and compounds are at the extent of tractable calculations in terms of model sizes and computation times; a calculation of a steel considering all of its constituent elements is still a decade or more away from being routine. Furthermore, many alloys behave poorly without thermomechanical post-processing. The integration of thermomechanical constitutive models of post-processing, together with DFT and CALPHAD calculations of the properties of the base chemistries, to automatically search across the process-structure-property space of alloy design is still to be attained – in the best documented examples, the calculations of composition and thermomechanical post-processing effects of the processing are still hierarchical and bespoke [24]. Still, experimental data - based ML is equally challenging – largely because the number of data points in regard to any one composition is usually very limited – on the order of ones to tens. Hence, the number of data are insufficient to inform ML algorithms of complex relationships, especially those mathematically described by nonlinear relationships.

In this paper, we develop a feature engineering – based approach to enable machine learning to assist in the development of alloys and their processing using limited experimental data. Specifically, we will show that by using feature engineering practices established by the DFT community to model alloy chemistries, together with new, simple, established constitutive models of precipitation physics to pre-process thermomechanical post-processing data, ML can work better for alloy design than it could using alloy chemistries in isolation. We will also show that considering the initial processing technique in the ML model leads to further improvements. We will demonstrate the ML framework by verifying its ability to predict a new shape memory alloy composition – post-processing combination results in phase transformation temperatures outside of the range of the training data set in regard to the combination of thermal hysteresis and mean transformation temperature.

Shape memory alloys (SMAs) make an challenging and necessary test case for developing ML based upon physical experiments, largely because computational methods are still being established to computationally predict shape memory behaviors from alloy chemistries and processing [25][26]; hence, physics-based experiments still largely drive the development of SMAs [27][28][29]. Furthermore, many SMAs do not exhibit shape memory behaviors at all unless they are thermomechanically post-processed with very specific treatments; in fact, NiTi, the most prolific SMA to date, exhibits poor shape memory properties sans thermomechanical post-
In this work, we will focus specifically on thermoelastic shape memory alloys; alloys that recover their shape in response to thermal or mechanical load changes via a reversible martensitic (first-order, diffusionless) phase transformation between high temperature, high symmetry austenite phases and low temperature, low-symmetry martensite phases [30]. The chemistry approach to tune TTs of SMAs for high and low temperature applications is to alter stoichiometries or introduce new alloying elements; for example, within the range of 50 to 52 at.% Ni in NiTi, 0.1 at.% change in Ni changes transformation temperatures by 20 K [32], while Co, Cr, V, Fe, and Mn can be added to NiTi to lower the TTs [33], whereas Hf, Pd, Pt, Zr and Au increase temperature [34]. In addition to chemistry, post-processing such as mechanical work and heat treatments play important roles to modify TTs [30][31]. Hysteresis defines the differences between forward (austenite-to-martensite) and reverse (martensite-to-austenite) transformation temperatures. Hysteresis often defines the efficiency of the performance of a shape memory alloy; high hysteresis leads to more efficient dampers, while low hysteresis leads to more efficient actuators. Hysteresis can also be tuned with chemistry and thermomechanical processing. In the absence of defects or secondary phases, it is established that altering chemistry to tune the lattice parameters of the austenite and martensite phases such that they can share an undistorted phase boundary reduces the hysteresis [29][35][36]. However, alloys with low hysteresis and high fatigue lives have also been developed through the use of secondary phases and defects, demonstrating the limitations of our current understanding and models for hysteresis engineering for SMAs [28][37][38]. While gaining physical knowledge from ML is beyond the scope of this work, it is the lack of physical knowledge for hysteresis engineering with defects that further motivates the desire to use ML as a tool to create models where physical models do not currently exist.

Finally, NiTiHf alloys have emerged as one of the more promising class of ternary SMAs due to their ability to be strengthened sufficient to exhibit repeatable shape memory behaviors using only thermal post-processing treatments, sans mechanical work. For this reason, they also make a desirable system for the development of ML methods for alloy design based on physical experiments – mechanical work is not needed to attain good shape memory behaviors, hence reducing the dimension of the design space. However, while single-step heat treatments may be sufficient to evoke shape memory behaviors from alloys containing moderate (15 – 30 at.%) amounts of Hf [34][39], we have found that multi-step aging treatments are more effective for moderate (3 – 15 at.%) Hf compositions [40][41][42]. Therefore, the ability to ignore mechanical post-processing comes with a need to consider a high dimension thermal post-processing design space. Relatively few NiTiHf alloys have been made to date, especially in moderate Hf compositions, hence, this class also has a lot of currently unexplored design space. Specifically, NiTiHf alloys with transformation temperatures between 220 and 275 K with low hysteresis have not been previously made, yet they are desirable for biomedical and aerospace applications. This transformation temperature range is the desired range for many medical implants; the Hf addition increases radiopacity relative to binary NiTi, making medical implants more visible with X-rays. Then, for aeronautics, these transformation temperatures would create the ability for an aircraft structure to morph autonomously, without an electrical control system, in going from takeoff/landing scenarios near the ground, where temperatures are usually above 275 K, to cruising at altitude, where temperatures are usually below 220 K. Binary NiTi alloys can attain the right mean transformation temperature, but not the required hysteresis – mean TT combination. Hence, the desire to discover new low-hysteresis NiTiHf alloys with low transformation temperatures defines the design space and motivates the demonstration of the physics - informed feature engineering ML approach for alloy design in this work.
2. Results

2.1 Data and predictor

2.1.1 Data collection and processing

The success of any data-driven ML method ultimately depends on access to reliable and plentiful data. The experimental data of NiTi and NiTiHf alloys in this work is extracted from 558 datasets [43], originated from previous published literatures and our unpublished data. Our training database includes 132 datasets for binary NiTi, 358 for high Hf (>12 at.%), and 68 for lower Hf (≤12 at.%) NiTiHf alloys. 33 among of 68 are our newly experimental data. It can be seen there is insufficient data in low Hf region where we are of interest. The Ni composition ranges from 48.5 to 51.5 at.% and Hf ranges from 0 to 30 at.% as shown in Fig. 1(a-b). For the NiTiHf alloys, transformation temperatures $M_s$, $M_f$, $A_s$, and $A_f$ are of importance for determining the characters of alloys. We basically have two properties as demonstration for workflow, as shown in DSC (differential scanning calorimetry) curve in Fig. S1. One is for the average of $A_f$ and $M_f$, which implies the overall phase transformation temperature location. It targets to 230-260 K for low temperature applications. Another is the width of phase transformation ends $A_f$ and $M_f$, which indicates the range of thermal hysteresis, and lower hysteresis is more desirable for SMAs due to the higher fatigue resistance. Thereafter, TTs denotes the transformation temperatures $(A_f+M_f)/2$. Then relative hysteresis $\Delta RT$ was used to present the phase transformation width $(A_f - M_f)$.

Fig. 1(a) shows data distribution for Ni, Hf compositions against TTs, the brighter and larger size points, means higher transformation temperature. It shows regardless of what processing or other parameters involved, TTs increases with Hf content and generally decreases with Ni content. The TTs covers the range of 200-800 K. Generally, Hf-high (Hf > 10 at.%) alloys have TTs in the range of 400-800 K, while Hf-low (10 at.% > Hf > 0 at.%) alloys are in 200-400 K for low temperature applications such as aircraft actuator and biomedical implants. The binary NiTi alloys are with TTs for about 200-400 K, but with different underlying mechanisms to dominate TT as ternary NiTiHf alloys. Fig. 1(b) shows the distribution for Ni, Hf compositions against relative hysteresis $\Delta RT$. The range of $\Delta RT$ is more randomly scattered in 25-200 K. There is no trend exist purely rely on composition distributions because of additional processing features’ influence.

Fig. 1(c) shows the TTs variations against Hf content under different processing categories. It generally includes as-fabricated (Unproc), after solid solution (Sol), direct aging (Aging), solid solution plus aging (Sol + Aging) and solid solution plus pre-aging and final aging (Sol + PreAging + Aging). At specific Hf content, the property values randomly scatter along vertical axis due to the variations of Ni content and processing features. Even it obscures the tendency of property variations, it still can be seen that TTs seems keep as constant up to 10% Hf. It then increases with Hf content sharply beyond 10%. The overlap and cluster points at 20% Hf content shows a large variation of TTs, this is because 20% Hf alloys is well studied, and researchers generate a lot of data with various Ni contents and processing. Fig. 1(d) shows the dependence of $\Delta RT$ on Hf content under different heat-treatments. Purely examine Sol datasets, it seems that $\Delta RT$ increases with Hf content to a maximum around 8-10% Hf. Then it decreases to a minimum around 20% Hf, and sharply increase beyond 20% with Hf content. The Unproc datasets seem also present the similar tendency except the region beyond 20% Hf since large variations in Ni content and heat-treatments. All of these show the processing parameters have a significant impact on properties and there is not a general rule to describe the property change tendency. In addition, our experiments data in green diamond shows TTs can reach to lower values being about 250 K, and some of them exhibit relative lower $\Delta RT$ at around 8-10% Hf region compared with others. Besides the composition and processing related parameters, alloys synthesis ways and the applied stress are also important factors to dominate the alloys characters. The synthesis ways include different elements melting methods. Even though the most transformation temperatures though DSC measurements are under zero stress level. At stress state, the TTs ($M_s$, $M_f$, $A_s$, and $A_f$) are higher than that at zero stress due
to the increased energy of the lattice. Therefore, it also incorporates the features of synthesis ways and stress level (tension, compression, and stress value) for the data-driven models.

Fig. 1 The dependence of (a) TTs and (b) ΔRT on Ni and Hf content variations. The brighter and larger circle means higher property value and vice versa. The variation of (c) TTs and (d) ΔRT on Hf content under different processing conditions (Unproc: as fabricated and homogenized; Sol: homogenized and solid solution; Aging: homogenized and direct aging; Sol + Aging: homogenized plus solid solution and aging; Sol + PreAging + Aging: homogenized plus solid solution, pre-aging and final aging). (e) The feature importance ranking using mutual information (MI) score to select the high informative features in respect to output properties and (f) Pearson correlation matrix to select the composition features in high correlation clusters. Blue and red colors indicate positive and negative correlations, respectively.

2.1.2 Predictor

Constructing a machine learning model includes training data, a set of features to describe each dataset, and a machine learning algorithm to map features to properties. Generally, the features of data are represented as nonlinear functions with property. Considering the number of training data from experiments are limited and the complex relations among composition, processing and functional properties, gaussian process regression (GPR) model is one of the most powerful, versatile and widely used methods [44]. There are several advantages to make it suitable for this problem. First, Gaussian process is good at capturing high-dimensional feature space and non-linear relations for interpolation even for extrapolation. Second, Bayesian updates inhered in modeling provides a natural pathway for estimating predictive uncertainty in addition with prediction itself. This approach assumes that a Gaussian distribution of models fitting the available seen data. The mean and variance of these predictions—the natural outcomes of GPR—are the most likely predicted value and prediction uncertainty, respectively. They were used to help navigating and searching the designed property using fewer experimental trials. Other regression methods like support vector regression (SVR) and random forest (RF) methods, the scheme such as bootstrapping may also be utilized to estimate uncertainties, but at significant added cost [44]. The GPR overfitting problem with less experimental data was avoid by tuning hyper-parameters with conjugate gradients method.
The model predictive ability was evaluated using cross-validation. The details can be found in the Materials and Methods section.

2.2 Physics-informed feature engineering and down-selection

2.2.1 Physics-informed feature engineering

The high-throughput calculations in materials informatics, the features to describe the properties generally based on the material compositions such as elemental property attributes, electronic structure attributes, crystal structure representations, or the DFT calculated formation energy [14]. They either keep the synthesis path constant or consider it as neglectable, that is, rarely consider processing features in their model. But in practical alloys developments, the properties of alloys are not only dependent on chemical compositions, also on synthesis ways, processing parameters, and even measurements conditions. The descriptors only from traditional materials compositions are not enough to reflect the produced functionality of alloys.

As traditional ways of material informatics, we developed a set of physics-based composition features which based on the statistics of attributes of constituent elements. The composition attributes for each alloys, includes basic elemental attributes (e.g., period/group on periodic table, atomic number, atomic radius etc.), electronic structure attributes (e.g., total valence electrons, average fraction of electrons from s, p, d and f valence shells, and the unfilled electrons as well), reactivity attributes (valance, electronegativity Pauling, electron affinity, and ionization energy), and elemental thermal properties (e.g., melting point, heat of fusion, thermal conductivity etc.). Besides that, the synthetic ways, heat-treatments parameters and measurement conditions are contained in the model to capture as much formation mechanism of alloys. In addition, by adding physiochemical knowledge into the model, feature engineering approach processes domain knowledge as additional features which describes phase transformation in respect to heat-treatment parameters. We therefore covert the heat-treated temperature (T) and time (t) into more meaningful transformed features (T_T and T_t) by incorporating the knowledge on precipitate growth kinetics, which describes the underlying relations between processing and transformation temperatures. The main reason is that nonlinear relations between processing and properties were converted into the linear and simplified relations, which provides more physics meaningful features that could lead to robust predictions. The calculation of composition, processing-related features and physiochemical domain knowledge were detailly described in Materials and Methods Section.

2.2.2 Feature down-selection

We totally generated 48 features or descriptors for the model as shown in Table S1. Whereas some of the features are not directly associated with output properties. The feature down-selection process can reduce model complexity and make the model more interpretable. The results of modeling are then easier to be understand. Here, the feature importance ranking using mutual information (MI) score method [45] was shown in Fig. 1(e). It equals to zero if and only if the feature is independent with output property, and higher MI score mean higher dependency. It shows that composition features generally have higher score compared with other processing-related features. The electronic structure attributes Avg.s, Avg.Sp, Avg.p, Avg.Us, Avg.Up, Avg.Uf exhibit very low score and were deleted for ML modeling. Synthesis ways is not significant than composition features but still comparable with other processing features. The transformed processing features encoded with physiochemical knowledge are comparable or more significant than original ones. The pre-aging related features have low score because of the shortage of data with pre-aging conditions in our database. It will be of significant when more pre-aging related experiments are available. Even process features importance is less than composition features, these process features were still kept, because no significant low features exist except pre-aging.

After the feature importance ranking, there remain many composition features. Eliminating the composition feature pairs that are strong correlated will reduce the total number of features for
better understanding of model. Pearson correlation method [46] is employed to define feature pairs with correlation coefficient larger than 0.95 are highly correlated. The heatmap matrix for this purpose and is presented in the right top of Fig. 1(f). It was calculated using the training data with blue and red colors indicating positive and negative correlations, respectively. The lighter the tone, the less corresponding correlations. The final feature selection procedure goes through the high MI score feature in Fig. 1(e) and then check the heatmap in Fig. 1(f) to delete the redundant high correlated and low MI score features. It iteratively goes through several times and finally retain a subset of 11 composition features in the left bottom of Fig. 1(f).

2.3 Model with physics-informed features

2.3.1 Model with chemically composition and processing features

![Fig. 2](image)

Fig. 2 Model prediction for (a-c) TTs and (d-f) ΔRT. (a, d) trained only with chemically composition features, (b, e) trained with only processing features, and (c, f) full features model trained with pre-selected composition and processing features. The inset histogram shows trained models’ relative predicted error and predicted uncertainty. The insert text box shows training matrix with R², MAE and RMSE. The legend indicates the training data comes from published data (green square) and lab unpublished data (blue circle). Perfect model along with diagonal line, and dark and light-yellow shadows present the 1 and 2 times of mean predicted uncertainty or deviation. The mean uncertainties for (a-c) TTs are 44.14, 77.17, 19.97 K, (d-f) ΔRT are 24.14, 22.20, 12.04 K, respectively. To avoid the model overfitting, the bias-variance trade-off techniques were used to search optimized hyper-parameters as shown in Fig. S2.

1) Trained only on chemically composition features

As traditional materials informatic ways, the model performance fitted on chemically composition features are evaluated in the left column of Fig. 2. An ideal model would place all predicted values over the black diagonal line, we can see all points from our model following the line closely. For property TTs in Fig. 2(a), it shows the model performance for TTs is not bad and
with relative higher $R^2$ 0.85. The inset in top left shows model predicted 90% of data with relative error lower than ±20%. The mean predicted standard deviation of training data is 44 K in dark-yellow diagonal strap. The difference between the experimental observed TTs and predicted value lied along with diagonal line within the light-yellow strap. The lab experiments data (blue circles) position closely with diagonal even the large difference in processing parameters. Whereas the predicted performance on $\Delta RT$ is bad when it was only trained on composition features as in Fig. 2(d). The $R^2$ only reach about 0.5. The width of uncertainty strap in yellow and predictive error exhibit large compared with observed data. Not only the data from literature experiments but lab experiments align with horizontally and didn’t show predictability.

2) Trained only on processing features

Processing conditions are not negligible for alloys design. It trained with heat-treatments parameters, synthesis ways, physiochemical theories features. It shows when trained only on process features in the middle column of Fig. 2, the predicted performance of TTs is much worse than those trained on composition features. The trained $R^2$ is only reach up to about 0.6. The predictive error bar and uncertainty yellow strap are much larger than the model training on composition features. This may be consistent with domain knowledge intuition that alloys composition is more significant to determine alloys property than processing features. It seems that the trained model on process features shows no predictive capability in high TTs region. Surprisingly, the $\Delta RT$ training on processing features shows the opposite. It shows the training with process features ($R^2=0.60$) is better, at least equivalent with the training on composition features ($R^2=0.49$). The predictive error and other metrics MAE and RMSE exhibit lower than previous composition features model. This may contradict with expert intuition that alloys composition largely dominates alloys properties than processing features. There are some functional properties like $\Delta RT$, their value and change tendency along with composition preparation and processing parameters are more complicated. The processing parameters may exhibit a dominant position and are non-negligible for materials modeling.

3) Trained full features model

The right column of Fig. 2 shows the full features model trained both on chemically composition and processing features. It shows both properties exhibit much better than previous subgroup features model. The $R^2$ value of TTs can increase up to 0.98, the MAE and RMSE decrease to about 9 and 14 K, respectively. The inset of top left shows about 95% of data with relative predictive error lower than ±10%. The insets in right bottom shows predicted uncertainty is on the range of 18-25 K. The average uncertainty 19.97 K as yellow diagonal strap is much narrow than before. Full features model on $\Delta RT$ exhibit similar improvement. The model predictive $R^2$ can reach up to 0.93 from 0.5 on compositions model and 0.6 on processing features model. The inset shows about 90% of data predicted with relative error ±20%. The predictive uncertainty can also decrease to reasonable value as shown in yellow strap. In addition, it generally shows the $\Delta RT$ model is a little worse than TTs model, in which more predicted points are far away from the diagonal.

4) Verification of Physics-informed model

Table. 1 Performance of six approaches of including different features for TTs and $\Delta RT$ model on a cross-validation test. The evaluation metrics are obtained with the average of 10 times cross-validation on testing. Mean std indicates the average of predictive uncertainties. Full features: pre-selected chemically composition features + process features. W/O Physiochemical theories: full features but exclude physiochemical encoded features. W/O Synthesis ways: full features but exclude synthesis ways feature. W/ Raw composition features: only raw element composition fraction features (baseline model). W/ Chemically composition features: only pre-selected chemically composition features. W/ Process features: includes heat-treatments, synthesis ways, measurement conditions and physiochemical theories encoded features.
To elucidate the contribution of different features on modeling, we explored different modeling approaches for accounting for alloys compositions, processing parameters and synthesis ways in our machine learning model as shown in Table 1. Specifically, it performed a cross-validation on the training data with different features models. In the test, it withheld 10% of the data as test and trained on the remaining 90%. It evaluated the predictive accuracy of testing to see the generalizability of model on the data outside the training. The testing results can then be averaged to obtain the evaluation metrics as listed in Table 1. The lower MAE and RMSE and higher R² values are more desirable for prediction model. The full features approach means the training on composition and process features as shown in Fig. 2, which specifically includes the pre-selected composition features, raw processing features, physiochemical theories features, and synthesis ways feature. The synthesis ways feature is the scheme where we added a new attribute to the representation that feature value equal to 1 if performed with VIM and 2 if performed with VAM and 3 for others. They are simply added as attributes into the model.

The baseline model of raw composition features achieves a cross validated accuracy of R² about 0.8 and 0.3 for TTs and ΔRT. Whereas the chemically composition features model can increase the R² to 0.82 and 0.34. The predictive uncertainty (Mean std) are also slightly depressed. The full features model we adopted for final alloy design performs best for both properties TTs and ΔRT being about 0.92 and 0.68. The MAE and RMSE are lowest among six sub-features model. The mean of predictive uncertainty (std) are also lower than other incomplete features model. When no physiochemical theories feature or synthesis ways features included into the model, the model become worse with RMSE increase and R² decrease. Even the full features model achieves higher predictive accuracy than w/o physiochemical theories model, their predictive uncertainty both reach the lower value 28.5 K and 16.5 K for TTs and ΔRT. When only process or chemically compositions included, the cross-validated test error becomes much worse as training error demonstration in Fig. 2. This quantification presents the inclusion of synthesis method, physiochemical theories and chemically composition features can improve the predictive accuracy and depress predictive uncertainty as well. The full features model was used for the calculation of subsequent ternary property color maps thereafter.

| CV evaluation metrics | Model for TTs / ΔRT |
|-----------------------|---------------------|
|                       | Full features       | w/o Physiochemical theories | w/o Raw composition features | w/ Chemically composition features | w/ Process features |
| MAE. [K]              | 19.40 / 10.78       | 23.12 / 11.85               | 20.88 / 11.39               | 35.41 / 19.48                   | 34.54 / 19.10       | 48.45 / 14.91     |
| RMSE. [K]             | 31.07 / 16.68       | 38.70 / 19.64               | 37.62 / 18.53               | 46.17 / 26.16                   | 45.89 / 25.42       | 75.75 / 21.94     |
| R²                    | 0.9175 / 0.6794     | 0.8653 / 0.5484             | 0.8756 / 0.6345             | 0.8048 / 0.2924                 | 0.8185 / 0.3444     | 0.5235 / 0.4499  |
| Mean std. [K]         | 29.57 / 16.68       | 36.24 / 18.58               | 28.53 / 16.25               | 46.12 / 25.94                   | 44.86 / 24.91       | 80.64 / 22.91    |
2.3.2 Synthesis/manufacturing ways dependence

Fig. 3 The TTs prediction profile for VAM (a, c, e) and VIM (b, d, f) methods under heat-treatment conditions Sol (1050 °C/0.5 h, WQ), Sol (1050 °C/0.5 h, WQ)+Aging (550 °C/3.5 h, AQ), and Sol (1050 °C/0.5 h, WQ)+PreAging (300 °C/12 h, AQ)+Aging (550 °C/3.5 h, AQ). The prediction region are 0≤Hf≤30 at.% and 49≤Ni≤52 at.%. The predictive uncertainties are provided in Fig. S3.

Materials compositions and processing parameters play a critical role in determining characteristics of SMAs, while different synthesis methods may also have un-negligible influence. We therefore want to know whether the properties of SMAs were sensitive to different synthesis methods (particularly if VIM was different from VAM). Out of various alloys melting methods, vacuum induction melting (VIM) and vacuum arc melting (VAM) are widely used for production of SMAs. The cost of production by either method is similar and they both provide suitable material for current application requirements [47]. Double melting process using VIM primary melting followed by vacuum arc re-melting (VAR) is often used to get further fining with greater degree of chemical homogeneity. It was widely commercially exploited for binary NiTi and the ingots ranging from several grams to several thousand kilograms can be produced. In training data, the alloys either manufactured with VIM and VAM, and commercial binary Ni-Ti mostly manufactured through VIM/VAR double melting. Other synthesized methods such as selective laser melting (SLM), electron beam melting (EBM), plasma melting, and chemical film deposition are not common and are excluded.

For detail examination of how different synthesis methods affect TTs of SMAs, VAM and VIM prediction profiles were plotted in Fig. 3 under three different heat-treatments. The prediction region is 0≤Hf≤30 at.% and 49≤Ni≤52 at.%. It shows for three heat-treatments, comparison of TTs profiles indicate that manufacture ways have a strong influence on TTs. The VIM synthesized alloys generally lower than VAM alloys by about 50-100 °C. In addition, TTs of high Hf content alloys are significantly greater than those of low Hf content. The TTs generally decrease with Ni content at the same Hf content. In addition, the TTs of Sol produced samples generally large than Sol + Aging samples, then Sol + PreAging + Aging produced samples exhibit smallest value. The more about the dependence of property on compositions and processing were presented in Discussions 3.4.
2.3.3 Multiple data sources dependence

Fig. 4 (a) The histogram distribution of property TTs for three distinct data sources. The model trained on (b) binary NiTi, (c) Hf-high, and (d) binary NiTi + Hf-high datasets, and then tested on target Hf-low alloys datasets. The insert text boxes present the evaluation metrics of training. (e-g) The amount of variance explained by each of the selected number of loading vectors (or principal components (PCs)), the inset shows the 2D mapping of first two principal components, which explains majority content of feature space. The explained variance by PC1 and PC2 are 86%, 87% and 84% for full features, composition features and process features, respectively.

The training data comprises of binary NiTi, Hf-high and Hf-low NiTiHf alloys where different driving mechanisms dominating the characters of SMAs since the large variation of constituent compositions. The involved process features are also varied with each dataset. Although the dominant factors for determining TTs are not same, but they share some common and should be useful for target Hf-low alloys design. We therefore need to know whether the individual models in different data source are sensitive to other data predictions. Here, we use TTs model as demonstration. Fig. 4(a) shows the histogram distribution of property TTs covering the entire TTs range. The binary NiTi and Hf-low data have similar distributions. There is some overlapped region with high-Hf alloys which exhibits relative higher values. As was shown in Fig. 2(c), a mixed-family source model does reasonably well among the different data sources. It achieves $R^2=0.9835$ and lower predictive uncertainty. As a validation, two separate models are trained only on a specific family, namely the binary NiTi and Hf-high, then test on Hf-low data. Benchmarking on mixed-family model of Fig. 2(c), the model performed well on datasets belonging to their training set family while demonstrating a little poor predictive power on target low Hf datasets (Fig. 4(b, c)). The model trained on binary NiTi and Hf-high alloys generally underestimates the TTs of Hf-low alloys. Specifically, the model trained on binary NiTi family gives better prediction in low TTs region, whereas trained on high Hf family gives more predictive capability in high TTs region. While combing binary NiTi and Hf-high datasets as training (Fig. 4(d)), it can improve predictive capability in Hf-low alloys compared with merely make use of one of other data source. That means,
each group of datasets contribute to a specific TTs range or features, and as a result, the mixed-family model is better determined over large property range and feature space.

Fig. 4(e-g) show the 2D mapping using principal component analysis (PCA) to indicate the similarities of feature space for three data sources. Since the feature space is large, it is difficult to visualize and compare their difference. Principal components (PCs) visualization were employed in order to gauge the trends and the spatial variations in the data [48]. The original data is decomposed into a new set of eigenvectors based on criteria of maximum statistical variance between eigenvectors and orthogonality. The first eigenvector accounts for the largest amount of variance, the second represents the second largest variance in the data set, and so on for all subsequent eigenvectors. It allows examination of data structure without any underlying physical assumptions and allows identification of statistically relevant behavior. From the figures, it shows the first two PCs explains the majority content of feature space, about 86%, 87%, 84% for full features, composition features and process features. The 2D mapping of full feature space shows binary NiTi mainly locates at left quadrant while Hf-high at the right. The target Hf-low data in cyan circles are at the boundaries, and they overlap both with binary NiTi and Hf-high data. The similarities mainly come from the process features where there exists large overlapping region on the left quadrant of Fig. 4(g). The 2D mapping of composition features for three datasets are apart from each other in Fig. 4(f).

2.4 Alloys design through predictive profiles

For engineering materials SMAs, because the cost of manufacturing is high and the design of feature space is large, a down-selection of the materials systems need to be addressed in experiments. The compositions of alloys, phase constitution and microstructure need to be finely tuned to specific applications to reconcile many, and often contradicting properties. The already large unexplored composition space is thus further expanded by the processing parameter space. Alloys can be synthesized by different ways and many possible processing parameters that lead to different material microstructures and properties. The experiments of this study are primarily on the validation of developed model for acceleration alloys design rather than a specific application or solution. The developed physics-informed model can potential be used for different purposes and in many different fields. For example, in this paper, we target to design Hf-low alloys with low TTs. It can be used for aircraft actuator which requires phase transformation upon heating and cooling at different flying height to adapt external environment (A\text{f} \sim \text{ground temperature}). The biomedical implants SMAs require the A\text{f} is below body temperature for super-elasticity behavior upon loading and unloading (A\text{f} < 310 K). Here we target to design alloys with transformation peak TTs neither too high nor too low being about 230-260 K. In addition to the smaller thermal hysteresis is more desirable because the improvement of fatigue resistance.

It has been shown that NiTiHf alloys, its Ni content in between 50 and 51 at.% is potential for shape memory effect related applications, in 50 ~ 54 at.% Ni content is for superelasticity related applications, and in 54 ~ 57 at.% is for ultra-hard bearings [49]. As a demonstration of the potential of model for design low temperature SMAs, the models were applied to predict properties for searching candidates among generated composition space under specific processing parameters, a large fraction among them have not been tested. The constituent composition space with constraints of 49%\leq\text{Ni}\leq52% and 0\leq\text{Hf}\leq30% in atomic percent. The concentration of Ni and Hf variation step is 0.1% and 0.2%, respectively. For convenience of sample preparation, the final alloy composition only chooses Hf content with 1% step and Ni with 0.1% in atomic percent. The ternary property profiles of VAM synthesized alloys were predicted because it has the advantage, in contrast to VIM method, of providing more accurate designed alloy compositions without much TiC formation [50]. When searching optimal composition, the heat-treatment parameters were set as constant based on domain knowledge, they are Sol (1050°C/0.5h, WQ), Sol + Aging (1050°C/0.5h, WQ + 550°C/3.5h, AQ), and Sol + PreAging + Aging (1050°C/0.5h, WQ + 300°C/12h, AQ + 550°C/3.5h, AQ).
Fig. 5 shows the $\Delta RT$ prediction profiles for VAM synthesis ways and under various process conditions. The VAM predictive profiles for TTs have been shown in Fig. 3(a-c). The corresponding predictive uncertainties are provided in Fig. S(3,4). Fig. 3 (a-c) shows TTs keeps almost constant at Ni< 50 at.%, and then generally decreases with the Ni content increasing. The TTs of high Hf content alloys are significantly greater than those of Hf-low content alloys. It then presents TTs prediction profiles of Sol + Preaging + Aging condition are generally lower than Sol + Aging, then significantly lower than Sol process condition. Fig. 5(a-c) shows that in Hf-low content and Ni-rich region, the $\Delta RT$ is in the range of 40-70 K for Sol + Preaging + Aging condition. The Sol + Aging condition has lower value in 30-55 K, and are significantly smaller than Sol condition where exhibits being about 90-120 K. It also shows the tendency that Ni-rich alloys exhibits higher $\Delta RT$ than those of Ni-lean alloys, except for Hf-low region of Sol + Aging conditions which seems keep lower constant value being about 30 K. In addition, compared with Sol and Sol + Aging processing conditions in Fig. S(4), it shows in the Hf-high region of Sol + Preaging + Aging produced samples, the predictive uncertainties are very large either for $\Delta RT$ or TTs property since there are few data with Pre-aging processing in this Hf-high region. Correspondingly, the uncertainties of Sol and Sol + Aging conditions are largely depressed. After examining the training data, the Preaging produced samples are solely our lab experimental data wherein Hf content is about 8 at.%. Furthermore, Sol + Aging and Sol + Preaging + Aging process conditions in Fig. S4, the predictive uncertainties of $\Delta RT$ in low Hf region are about same value 60 K, whereas the Sol condition present lower about 25 K.

Based on the above prediction profiles, it finally chooses the alloys compositions with Ni50.7Hf3, which have not been synthesized, and near the top of priorities for target low-temperature related applications. The DSC measurements of Ni50.7Hf3 alloy were performed on different heat-treatments in Fig. 5(e). It shows that TTs of alloys is generally hit the designed property being about 250 K under different process conditions, at the same time, $\Delta RT$ reach to lower being about 80 K. The true hysteresis ($A_p-M_p$) is only about 35 K. We also synthesized other compositions (Ni50Hf3, Ni50.4Hf3, Ni50.4Hf12, Ni50.5Hf11) even they are not hit the lower TTs as shown in Fig. S8, but it can fill the gap of blank region in Hf-low, then make a comparison with experimental observation for model validation covering the entire Hf content. The Fig. 5 (f, g) shows the alloys properties of newly fabricated and connection with published data [51][52][53][54] with Ni content 50.3 at.% and Hf content from 0, 15, 20, 25 to 30 at.% and under almost same processing parameters Sol (1050 °C/0.5 h, WQ) + Aging (550 °C/3.5 h, AQ). It presents the properties variation against Hf content. The experiments generally agree well with the prediction curves considering the predictive uncertainty. The little difference in prediction and experiments may come from a little difference of Ni content and processing parameters. The TTs seems keep plateau or a little decrease with the Hf increasing up to 10 at. %, then increase significantly at 30% Hf content. The $\Delta RT$ seems messy when Hf content is below 15 at. %, then it increases drastically with Hf content. More predictive curves extracted from ternary heatmaps were presented at Fig. S5 and discussed in Discussions 3.4.
Fig. 5 (a-c) The $\Delta RT$ prediction profile for VAM synthesis ways under heat-treatment conditions Sol (1050 °C/0.5 h, WQ), Sol (1050 °C/0.5 h, WQ) + Aging (550 °C/3.5 h, AQ), and Sol (1050 °C/0.5 h, WQ) + PreAging (300 °C/12 h, AQ) + Aging (550 °C/3.5 h, AQ). The prediction region are $0 \leq \text{Hf} \leq 30$ at.% and $49 \leq \text{Ni} \leq 52$ at.%. The predictive uncertainties are provided in Fig. S6. (d) An Ashby plot shows transformation temperatures TTs against relative hysteresis $\Delta RT$ and designed properties of 50.7Ni46.3Ti3Hf (e) The DSC measurements of 50.7Ni46.3Ti3Hf under different heat-treatments. (f-g) The predicted property curves of 50.3Ni alloys under different Hf content and Sol (1050 °C/0.5 h, WQ) + Aging (550 °C/3.5 h, AQ), and some experiments to show the property variation against Hf content by connecting the previous publications in binary NiTi and Hf-high region [51-54].

3. Discussions

3.1 Model performs better considering both composition and processing features

The advancement of machine learning has extended its application throughout many problems in materials, high-throughput DFT calculations, chemistry and medicine discovery to guide experimental design. However, compared to other fields, the experimental data of alloys design are
typically much smaller and more diverse in composition and processing, which undoubtable degenerates the construction of data-driven models. We proposed the physics-informed feature engineering approach which incorporates with chemically composition and processing related features. In addition, the modeling through GPR with bias-variance tradeoff techniques for hyperparameters optimization, can significantly improve the model predictability with the limited experiments data.

For alloys design, both constituent compositions and process features are important to determine output properties. The final properties are driven by these two competitive and complex interaction mechanisms. In previous material informatics, people either neglecting the influence of processing when modeling or synthesize the samples under same conditions. Until now, there is no clear demonstration and quantification on the effects of composition and process on functional properties of alloys. By exploring a vast number of different feature combinations, it can better reflect the reality and procedure of alloys synthesis, to model the experiments data with higher accuracy, that complement the traditional chemically composition descriptors in the past.

In SMAs, TTs and △RT are very sensitive with alloy compositions; Heat treatments are also the most effective method to control transformation temperatures and increase the alloys strength by formation of precipitates. The precipitation characteristics and corresponding properties highly depend on heat-treatments temperature, aging time, cooling rate. The possibility to use SMAs for related applications is accomplished by a proper selection of compositions, synthesis ways and successive heat-treatment parameters. The combinatorial models for TTs and △RT in Fig. 2(c, f) built with chemically composition and processing related features both reach high prediction accuracy. The cross-validation metrics on test data in Table 1 is a little bad than the training, but it is still high enough and acceptable. It suggested the ML model with combinatorial features are powerful enough to fuse and learn the knowledge from experiments data and create a set of universal features for alloys design. Furthermore, model predictability trained on process features for property △RT seems better than that trained on composition features. This indicates the dominant position of process features on property in relative to composition features, which contradict with some domain expert intuition. In addition, the modeling on △RT with R^2 about 0.93 is lower than TTs modeling with R^2 being about 0.98. It indicates the modeling on △RT is more difficult than on TTs, because the physics mechanism to dominate hysteresis is more complicated and some are still in unresolved. Even that, we already have a model in our hands that will convert the raw experiments data into the formalized composition and process features for mapping alloy properties with error reasonable.

3.2 Physics-Informed Feature Engineering Improves Model Predictability

3.2.1 Physics-based composition features

The converted composition features are more physical meaningful compared with raw composition fraction features. The developed chemically composition descriptors includes subgroups such as elemental property, reactivity, thermal property, and electronic structure attributes. These chemically composition features are more representative and physical meaningful than raw composition features. They are suitable for ML modeling to capture the underling mechanism to dominate the properties of interest. Table 1 shows for both properties TTs and △RT, the model trained based on chemically composition features is better than raw composition features in terms of RMSE and R^2, and their corresponding predictive uncertainties are also depressed.

The goal of design chemically composition features is to create a uniform descriptor representation of alloys that relates with essential physics and chemistry to influence functional properties. For example, [14] describes material based on the fraction of element and physically motivated heuristic descriptors, such as the maximum difference in electronegativity, to build models for the formation energy of ternary compounds. [55] uses the element property, such as boiling point and bulk modulus that strongly correlates with a more complex target quantity, to
understand the underlying physical mechanisms of dopant stability in zirconia. [56] explores a wide range of features, includes ionic radii, bond valence radii to assist the discovery of novel ABX3 compositions with perovskite crystal structure. [57] applies more comprehensive attributes, encompasses stoichiometric attributes, elemental property statistics, electronic structure and ionic compound attributes to predict diverse properties of crystalline and amorphous materials, such as band gap energy and glass-forming ability.

Our developed chemically composition features provide a general and relatively simple representation that reflect physical and chemical aspects of contributions for predicting alloys properties. In SMAs, [58] has shown that the $M_f$ are strongly depends on whether valence electron $e$ is greater or less than 7. The dependence of transition temperatures and hysteresis on electron number indicate the trends of dependence on incomplete $d$–$d$ orbital overlap at occupancy <7 and complete overlap at 7 [8]. The Pauling electronegativity $\chi$ and valence electron $e$ capture the alloys chemical bonding and the changes in valence electron concentration, respectively. It is known that the strong chemical bonding gives rise to large resistance to shape/volume change, and consequently results in high bulk and shear moduli. The elastic modulus of parent phase influences the transformation temperature [57]. Larger elastic modulus of the parent phase, cooling should continue before critical temperature point is reached; therefore, the TTs is depressed and vice versa. The introduction of alloying elements like Hf in Ni-Ti is accompanied by a change in interatomic metallic bonding, which is captured by features electronegativity $\chi$ and valence electrons $e$. The atomic radius $R_{rad}$ and covalent radius $R_{cov}$, has been shown to influence the thermal hysteresis. The thermal hysteresis increases with atomic size $R_{rad}$ at almost same electron valence number [8][58]. The alloying elements have different atomic sizes, the Laudau model represents the effects of induced strain by doped elements on transformation temperature [57]. The change of transformation temperature is influenced by the modulus and the local strains due to the effects of the atomic radii $R_{rad}$ of doped and host elements. Form these, it concludes that a general alloy composition descriptor set should contain the statistics of a wide variety of physics and chemical elemental properties to be adaptive. Fig. S6 shows the scatter plot of TTs variation against representative composition features $R_{cov}, \chi$ and $e$, which gives straightforward interpretation.

### 3.2.2 Physics-based process features

The transformation temperatures of SMAs strongly depend on heat-treatments parameters. The schedules of heat-treatments for achieving specific size and volume fraction of precipitates and their average interparticle distance, can significantly encourage or depress the phase transformation temperature of alloys. The features through physics-based precipitates nucleation/growth models are expected to be beneficial for ML modeling. Therefore, besides the developed chemically material descriptors, feature engineering for incorporating the phase transformation kinetics theory into the post heat-treatments schedules. All datasets assumed the alloy has been homogenized. These physics-processing features includes transformed solution temperature ($T_{_Tsol}$), transformed solution time ($T_{_Tsol}$), transformed pre-aging temperature ($T_{_Tpre}$), transformed pre-aging time ($T_{_Tpre}$), transformed final-aging temperature ($T_{_Tage}$), and transformed final-aging time ($T_{_Tage}$). The domain knowledge makes process features more physics meaningful that are more relevant with alloys properties. Although the empirical phase transformations may not sufficiently describe alloys functionality, it at least provides qualitatively knowledge about targeted property. From the first two column of Table 1, when there is no physiochemical theories related process features incorporated into the model, the average of CV test evaluation errors MAE and RMSE increased, and $R^2$ value decreased, but the predictive uncertainties does not show large improvement, even lower than full features model. That means the incorporation of physiochemical theories features into the model, it assists the model prediction but at the expense of model complexity in terms of feature space dimensionality.
3.2.3 Manufacture ways dependence

Quantitatively analysis of the difference of synthesis ways would be beneficial for model and practical usage. Comparison of the predictive maps of Fig. 3 indicates that the synthesis methods have a strong influence in alloys properties. The training database contains 240 for VIM method, 201 for VAM method, and 117 for VIM/VAR method for commercial binary NiTi. VAM production procedure does not need any graphite crucible. The constituent elements are mixed and irradiated by argon arc through an electrode made of a tungsten rod. The melting and solidification are carried out in water-cooled copper container and hence contamination from the crucible is avoided. Acceptable homogeneity samples usually produced through multiple melting cycles. Whereas VIM melts all raw materials simultaneously in a graphite crucible under vacuum or an inert gas atmosphere. It uses currents induced through electromagnetic induction to melt raw elements by changing the magnetic field. Electromagnetic induction induces eddy currents to create heating effects to melt alloys. Also, electrodynamic forces result in stirring and mixing the melt to encourage the chemical homogeneity. Graphite crucibles are generally used for this process due to easier handled and least expensive. However, the carbon contamination of the ingot from graphite crucible is a major drawback of VIM \[50\][59]. The TiC form during solidification will increase Ni concentration, which in turns depresses the transformation temperatures, thereby affect the characters of SMAs. The prediction profiles in Fig. 3 indicate VIM prediction profiles are lower than those of prediction of VAM under various heat-treatments. Therefore, model prediction does sensitive to different synthesis ways and capture the underlying phenomenon of alloys fabrication.

3.3 Multiple data sources dependence

In traditional alloys design, the experiments data for a specific alloy usually is insufficient for data-driven ML modeling. The alloys manufacturing and property characterization is expensive. For example, we only have 33 experiments datasets for Hf-low alloys, in addition to 35 heterogeneous data from different literatures, which is not enough for ML model. One strategy to overcome the data deficiency is through the feature engineering to design the comprehensive and powerful features to describe alloys. The other way is fully utilizing the data from other relevant but not common data source. Our training data contains a list of alloys, including many closely related alloys varying in stoichiometry and processing parameters as well. Each family alloys or processing provides one aspect of SMAs mechanisms. Then the mix-family model can learn the knowledge each other to enhance prediction performance overall.

For further examination the difference of three data sources and interpretability of physical mechanism to dominate the characteristics of SMAs, the feature contribution on each data source are shown Fig. S7. The difference of feature importance across three models reflect the fact that distinct mechanisms that are responsible for TTs and \(\Delta RT\) variations. The Fig. S1(a) shows most of composition features have significant impact on properties in binary NiTi alloys. The feature importance for both properties seems synchronously. For Hf-low alloys, the pre-aging relevant features \(T_{\text{pre}}\) and \(t_{\text{pre}}\) becomes more significant. The composition feature score for both properties becomes more inconsistent; some features show higher influence on TTs and some higher on \(\Delta RT\). Whereas Hf-high alloys, the processing importance score becomes lower compared to other alloys families. The MI score variations of each feature for both properties become more consistent. The group features of elemental properties, reactivity and thermal properties exhibit much more influence on properties, whereas electronic structure attributes present lower score on properties compared to Hf-low alloys family.

Typically, the traditional data-driven model treats different tasks or data sources separately, and each was trained to have individual models. Here, it has proven the feasibility of treating these different but related data source together to cover more wider feature and property range. In this paper, the three data sources have different compositions and property response range, the processing involved are also different. However, the knowledge learned from each source can be
shared and transferred through data groups to aid alloys experiment design. This scenario is common in practice, new material discovery or synthesis is slow and expensive, the number of datasets for initial training is insufficient and hard to obtain with experiments. When trying take advantage of the available related data source, it is possible to fuse and apply the knowledge learned to target task, then by adding some informative experiment data through active prediction to correct model in target region.

### 3.4 Alloys design with machine learning

In this paper, SMAs were used for modeling from physics-informed feature engineering approach, it can easily generalize to other kinds of alloys. The workflow quantitatively elucidates previous ambiguous opinion on the effects of compositions and processing on properties of alloys. The physics-informed features in terms of composition and processing can boost model predictive performance and interpretability. Therefore, through property prediction profiles regarding the composition and processing variations, it can be used for assisting alloys design with selection of candidates for desirable properties.

Prediction profiles from Fig. 3 and Fig. 5 under different process conditions, the alloys demonstrate a significant compositional dependence. For a more explicit expression for property and composition relationships, Fig. S5(a-c) shows tendency curves extracted from under Sol process condition. Fig. S5(a) presents the Hf dependence on TTs under different Ni contents. For 50.3 and 50.5% Ni contents alloys, TTs almost stay constant up to 10% Hf addition and increases drastically beyond 10% Hf. Previous reports present similar trends in Hf dependence of $M_s$ [30][32]. For Ni-rich content such as 50.9%, TTs tends to initially decrease with Hf, reaching a minimum about at 10% Hf, and then increase continuously. It also shows the Ni-rich alloys exhibits lower TTs than Ni-lean alloys. Fig. S5(b) shows the effect of Ni content on TTs. For lower Hf content alloys (Hf=5, 10, 15%), TTs is roughly independent of Ni-content for Ni<50% and exhibits a plateau. However, when Ni-content above 50%, TTs exhibits a non-linear, inverse dependence on Ni content which becomes steeper with increasing of Ni. For 25% Hf, even considering the uncertainty band, the plateau is not obvious and TTs was observed firstly increase up to 50% and then decrease sharply. Fig. S5(c) shows the Hf dependence on relative hysteresis $\Delta$RT for different Ni contents, exhibiting a strongly non-monotonic behavior. With increasing of Hf content, hysteresis tends to increase and attain a maximum around 8% Hf, and then a minimum around 20% Hf. Beyond 20% Hf, hysteresis increases again. Even large predictive uncertainty, it generally shows Ni-rich alloys exhibit larger hysteresis than Ni-lean alloys.

In fact, some researchers have noticed trend curve both for transformation temperatures and hysteresis dependence on composition variations. Tejas et al. [54] experimentally observed the variations of property $M_s$, hysteresis ($A_f$ - $M_s$) versus Hf and Ni content at the same heat-treatment conditions. At low Hf=0-10% content, it seems hysteresis increase with Hf content, whereas the $M_s$ either decrease then increase again or keep constant values in lower Hf region. They reported that slight reduction of TTs with small Hf=0-10% additions in Ni-rich NiTi alloys which agrees with our prediction maps. These results indicate that, along with Ni anti-site defects in Ti sublattice, leading to a drop of TTs in binary TTs [32]. Hf atoms replacing Ti should contribute to the increase in local lattice distortion, and thus, the reduction in TTs of alloys. In addition, the increase in local lattice distortions is likely to stabilize the quasi-dynamic strain nano-domains in austenite, similar to those reported for Ni-rich NiTi and NiTiFe systems [60][61]. These mechanisms tend to hind the martensite transformation temperature ($M_s$), causing energy barriers to the nucleation and growth of martensite. As the consequence, amount of overcooling is required for $B2 \rightarrow B19'$, and thus decrease $M_s$. The $\Delta$RT tend to increase in low Hf=0-10% region, seems caused by the reason of $M_s$ decreasing and the stabilization of austenite. This TTs decreases can be proven with electronic structure calculation, to check the role of Hf on the Fermi surface nesting responsible for $M_s$.
However, such calculations would be computationally expensive due to the size of the supercells needed to mimic the composition variations.

Under processed by Sol + Aging heat-treatments, the Hf and Ni content dependence on TTs is shown in Fig. S5(d–e). The tendencies are generally similar with those of Sol condition but not exactly same. For example, Fig. S5(d) shows three Ni contents exhibiting similar plateau in low Hf region which does not show obvious concave shape around 8% Hf. Fig. S5(e) does not present obvious Ni dependence on TTs when Hf is 5 and 10%, it does not show large TTs change covering all of Ni contents compared to Fig. S5(b). Fig. S5(f) shows a similar trend curve with Sol conditions, even there are larger uncertainties bond. △RT seems shift to left and reach a maximum and minimum at 7 and 17% Hf. All of these prediction insights are explicit and expect to be helpful for alloy design for specific practical applications.

Materials and Methods

Experimental procedure

The base ingot of NiTiHf alloys were made by vacuum induction melting (VAM) or vacuum arc melting (VIM) with high-purity elemental constituents. The pre-trained lab owned unpublished data are manufactured by VIM method. The alloys investigated consisted of compositions with 50.3Ni-6Hf, 50.3Ni-8Hf, 50.3Ni-8.5Hf, 50.3-9Hf, 51Ni-8Hf, 51Ni-8.5Hf, 51Ni-9Hf, 51.5Ni-6Hf, 51.5Ni-6.5Hf and 51.5Ni-7Hf. All of these alloys and thereafter are in atom percent (at.%). For these VIM manufactured samples, the ingot was homogenized in a vacuum furnace at 1050 °C for 72 h with water quenching. Then the sectioned samples were initially solution-annealed at 1050 °C for 0.5 h in an evacuated quartz tube, water quenched, and then they were pre-aged at 300 °C for 12 h and air-cooled (AC). Finally, the samples were final aged at 550 °C for different times such as 0.5, 3.5, 7.5 and 13.5 h, AC. As comparison, other samples without pre-aging or purely solutioned samples were also characterized.

For the new alloys design experiments, the samples were manufactured by VAM method with several times flipping to make sure the chemical homogenization. The raw ingots were then homogenized in vacuum with 1050 °C and 24 h and water quenching. Then the specimens for DSC measurements were cut from ingot and then solution treated at 1050 °C and 0.5 h in an evacuated quartz tube, followed by water quenching (denotes Sol). Then samples were heat-treated with pre-aging with 300 °C/12 h, air cooling, and final aging under different temperatures and time (denotes Sol + Preaging + Aging). Some samples are not processed with pre-aging and directly to final aging (denotes Sol + Aging). During aging process, the specimens were wrapped with Tantalum foils to prevent the oxidation.

Differential scanning calorimetry (DSC) tests were performed using a TA Instruments Q100 V9.9 with heating and cooling rates of 10 °C/min and temperature range between -150 °C and 150 °C for three cycles. The third cycle cure were used to measure transformation temperatures. The temperatures were determined with tangent method as shown in Fig. S1.

Physics-informed feature engineering

The feature importance rank and selection were performed with mutual information (MI) score and Pearson correlation matrix. The visualization of composition and process features in 2-dimensional space was analyzed with principal component analysis (PCA). The Scikit-learn python implementation of these algorithms were used [63]. In collected datasets, each dataset is characterized via a representation of basic chemically composition and processing features and its property response. The composition features derived from element composition and its fundamental chemical properties. They represent aspects of structure and chemical bonding information of alloy compounds. The processing parameters include synthesis ways, solution, pre-aging, final aging temperature and time, and applied stress when take the measurements. Through these descriptors,
we hope to capture as much formation mechanism of alloys. Our machine learning method then learns a map or model connecting features $X$ to a specific material property $Y$, here is $\Delta RT$ and TTs.

Each composition feature is uniquely described as weighted fraction of each constituent element’ chemical character as calculated in Eq. (1). These chemical character values should be available for all elements, and are expected to be physically meaningful and should have a linkage with many alloys properties, such as phase transformation temperature $T$, shear $G$ and Young’s modulus $E$ etc.

$$A_i = A_{Ni}f_{Ni} + A_{Ti}f_{Ti} + A_{Hf}f_{Hf} \quad (1)$$

where $A_i$ is the $i$th basic composition feature of alloys; $A_{Ni}, A_{Ti}$ and $A_{Hf}$ represent the chemical characters of constituent elements, respectively; while $f_{Ni}, f_{Ti}$ and $f_{Hf}$ are their composition fraction in alloys.

The phase transformation temperature related with precipitate growth, which has a sigmoid function relation with process temperature as Eq. (2). It implies that with lower insufficient temperature, the precipitate growth is inhibited with approximate to zero probability contribution. With higher enough temperature, precipitate growth is encouraged with approximate to 1 probability contribution to phase transformation. The $\theta$ of transformation temperature for solution, pre-aging and final aging are defined as 850, 300 and 500 °C, respectively.

$$\sigma(T) = \frac{1}{1+e^{-(T-\theta)}} \quad (2)$$

For growth kinetics, JMAK (Johnson-Mehl-Avrami-Kologoromov) model describe the kinetics of phase transitions which proceed through nucleation and growth. It gives the relation between the fraction of phase transformed $Y$, relative to the time, $t$.

$$Y(t;T) = 1 - e^{Kt^n} \quad (3)$$

where $K$ is a temperature-dependent growth constant and $n$ describes the orders of the growth. The above function can be transformed as $\ln\ln(1-Y) = \ln K + n\ln t$.

Table S1 gives a summary all the developed composition and processes features.

**Machine learning and overfitting problems**

The trained Gaussian process regressor (GPR) on training samples is used to map features to property. The GPML Matlab code version 4.2 was used [64]. Here it briefly explains the GPR modeling and how the hyper-parameters are estimated to avoid the over-fitting or under-fitting problems. The basic goal is to learn a function $y(x)$ to describe the relationships between the output property $y$ and the pre-selected features $x$, where $x = \{x_1, \ldots, x_p\}^T$ is a vector of input variables (e.g., alloys composition and process features). The response property $y(x)$ is experimental observed at $n$ distinct locations $x$, that is $Y = y(X) = [y(x_1), \ldots, y(x_n)]^T$. Then the Gaussian process is modelled as:

$$y(x) = f(x) + \varepsilon(x) \quad (4)$$

Assuming additive independent Gaussian noise $\varepsilon$ with mean 0 and variance $\sigma_n^2$, then co-variance function with Gaussian noisy term becomes $\text{cov}(Y) = K(X,X) + \sigma_n^2 I$. $k(x_i, x_j)$ is co-variance function, which capturing the dependence between different locations $x_i$ and $x_j$ within the feature space. In this study, the isotropic squared exponential covariance function (covSEiso) was used:

$$k(x_i, x_j) = \sigma_f^2 \exp \left(-\frac{1}{2l^2}(x_i - x_j)^2\right) \quad (5)$$

It next follows a Bayesian framework to estimate the hyper-parameters of model. Let $\theta = \{\sigma_n^2, \sigma_f^2, l\}$ denotes the GPR model’s hyper-parameters needed to be calculated based on the observations datasets $\{X, Y\}$. Model parameters are treated as random variables that follow a joint prior distribution $p(\theta)$, thus, the posterior distribution of the parameters given observed data, $p(\theta | X, Y)$, is computed using Bayes’ rule:

$$p(\theta | X, Y) \propto p(Y | X, \theta) \times p(\theta) \quad (6)$$
where the \( p(Y|X, \theta) \) is the likelihood function, which represents the conditional distribution of property response \( Y \) given input features \( X \) and prior distribution \( p(\theta) \). Upon computing posterior distribution \( p(\theta|X,Y) \), it can be used to make inference about the parameters after feeding observed data. In function-space view, after introducing the noise term and co-variance function, we can write the joint distribution of the observed target values and the function values \( Y \) at new test locations \( X_\ast \) under the prior as:

\[
[\gamma_\ast] \sim \mathcal{N} \left[ K(X, X) + \sigma_n^2 I \quad K(X, X_\ast) \right] \left[ \begin{array}{c} K(X_\ast, X) \quad K(X_\ast, X_\ast) \end{array} \right] \]

(7)

Knowing the initial experiments design points \( X \), observations \( Y \) and optimal \( \theta \), the prediction \( Y_\ast \) given a specific unknown \( X_\ast \) is given by:

\[
Y_\ast|X, Y, X_\ast \sim \mathcal{N}(\hat{Y}_\ast, \sigma^2_\gamma(X_\ast))
\]

where the mean of prediction \( \hat{Y}_\ast = K(X_\ast, X)[K(X, X) + \sigma_n^2 I]^{-1}Y \), and the prediction variance/uncertainty \( \sigma^2_\gamma(X_\ast) = K(X_\ast, X_\ast) - K(X_\ast, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_\ast) \). Thus, prediction is given as a normalized distribution with mean \( \hat{Y} \) and variance \( \sigma^2_\gamma \), that is, the prediction \( \hat{Y}_\ast \) at the experiment candidate \( X_\ast \) is associated with its uncertainties \( \sigma^2_\gamma(X_\ast) \).

Bias-variance trade-off techniques was used on GPR model to avoid the over-fitting or under-fitting to search optimal hyper-parameters \( \theta \). It minimizes the negative log likelihood \( p(Y|X, \theta) \) using conjugate gradients [65]. The high bias region means the under-fitting of model and high variance region indicates over-fitting. The boundary of under-fitting and over-fitting in dash line presents the optimal hyper-parameters. Fig. S2 shows bias-variance trade-off plot explicitly represent the optimal hyper-parameters \( \theta = \{\sigma_n^2, \sigma^2_\gamma, l\} \).

**Evaluation metrics and cross-evaluation**

The evaluation of performance of the models can be achieved by different metrics. First, the coefficient of R-squared (R\(^2\)) is a statistical measure of how well observed outcomes are predicted by the model. R\(^2\) values range from 0 to 1, where 1 is a perfect agreement between model prediction and experimental observation. \( \hat{y}_i \) is the predicted property value, \( y_i \) is observed value on the \( i \)th data, \( \bar{y} \) is the mean of observed property, and R\(^2\) is defined as,

\[
R^2(y, \bar{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - y_i)^2}
\]

(9)

Second, root-mean-square error (RMSE) is a frequently used measure of the difference predicted values and observed values. RMSE is the square root of the average of squared errors. A lower RMSE is better than a higher one, which is expressed as,

\[
RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

(10)

Similarly, mean absolute error (MAE) is

\[
MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

(11)

These metrics can generally evaluate the overall performance of training model. One explicit representation to examine whether overfitting problem exist in modeling is cross-validation (CV). The model is not fit to the entire dataset but rather the data is first split into training and testing sets, while as the model is fitted to the training data, then predictions from the trained model are compared to the test data to approximate model error. In a \( k \)-fold CV, the original dataset is randomly partitioned into \( k \) subsets of roughly equal size, of one subset is retained as the validation data for testing the model, and of remaining \( k-1 \) subsets are used as training data. Each of the \( k \) subsets was used exactly once as the validation data. The \( k \) results can then be averaged to obtain the more accurate estimate of model prediction performance. The evaluation metrics on the testing data \( R^2_{test}, RMSE_{test}, \) and \( MAE_{test} \) can be derived in the same way as Eq. (9-11).
Data availability
The data are available on Citrination publication named “NiTiHf Shape Memory Alloys,” Citrination, 2018. Available at https://citrination.com/teams/45/resources.

Supplementary Materials
Fig. S1. Schematic of DSC measurements and properties characterization for TTs and ΔRT.
Fig. S2. The bias-variance trade-off techniques to present the searching GPR optimal hyper-
parameters for TTs modeling.
Fig. S3. The TTs predictive uncertainty profiles for VAM and VIM synthesis methods under different heat-treatment conditions.
Fig. S4. The ΔRT predictive uncertainty profiles for VAM synthesis ways under different heat-
treatment conditions.
Fig. S5. Predictive tendency curves extracted from predictive ternary profiles under Sol and Sol + Aging process conditions.
Fig. S6. The representative features’ distribution against the property variation.
Fig. S7. Feature importance score using mutual information (MI) method for training on binary NiTi, Hf-low, and Hf-high alloys.
Fig. S8. The DSC measurements summary for different compositions and processing paramet-
ers.
Table S1. The descriptors or features generated through physics-informed feature engineering
approaches.

References and Notes
[1] G. B. Olson, “Designing a new material world,” Science (80- ), vol. 288, no. 5468, pp. 993–998, 2000.
[2] J. Allison, D. Backman, and L. Christodoulou, “Integrated computational materials
engineering: a new paradigm for the global materials profession,” Jom, vol. 58, no. 11, pp. 25–27, 2006.
[3] J. H. Panchal, S. R. Kalidindi, and D. L. McDowell, “Key computational modeling issues
in integrated computational materials engineering,” Comput. Des., vol. 45, no. 1, pp. 4–25, 2013.
[4] N. S. and T. C. (US), Materials genome initiative for global competitiveness. Executive
Office of the President, National Science and Technology Council, 2011.
[5] T. M. Pollock, “Alloy design for aircraft engines,” Nat. Mater., vol. 15, no. 8, p. 809, 2016.
[6] A. D. Spear, S. R. Kalidindi, B. Meredig, A. Kontsos, and J.-B. Le Graverend, “Data-
driven materials investigations: the next frontier in understanding and predicting fatigue
behavior,” JOM, vol. 70, no. 7, pp. 1143–1146, 2018.
[7] J. Ling, M. Hutchison, E. Antonio, S. Paradiso, and B. Meredig, “High-dimensional
materials and process optimization using data-driven experimental design with well-
calibrated uncertainty estimates,” Integr. Mater. Manuf. Innov., vol. 6, no. 3, pp. 207–217, 2017.
[8] D. Xue, P. V Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, “Accelerated
search for materials with targeted properties by adaptive design,” Nat. Commun., vol. 7, no.
1, pp. 1–9, 2016.
[9] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, “The high-
throughput highway to computational materials design,” Nat. Mater., vol. 12, no. 3, pp. 191–201, 2013.
[10] S. Curtarolo et al., “AFLOW: an automatic framework for high-throughput materials
discovery,” Comput. Mater. Sci., vol. 58, pp. 218–226, 2012.
[11] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, “Materials design and
discovery with high-throughput density functional theory: the open quantum materials
database (OQMD),” Jom, vol. 65, no. 11, pp. 1501–1509, 2013.

[12] A. Jain et al., “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation,” Apl Mater., vol. 1, no. 1, p. 11002, 2013.

[13] F. Ren et al., “Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments,” Sci. Adv., vol. 4, no. 4, p. eaq1566, 2018.

[14] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, “A general-purpose machine learning framework for predicting properties of inorganic materials,” npj Comput. Mater., vol. 2, p. 16028, 2016.

[15] L. M. Ghiringhelli, J. Vybiral, S. V Levchenko, C. Draxl, and M. Scheffler, “Big data of materials science: critical role of the descriptor,” Phys. Rev. Lett., vol. 114, no. 10, p. 105503, 2015.

[16] O. Isayev, C. Ose, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, “Universal fragment descriptors for predicting properties of inorganic crystals,” Nat. Commun., vol. 8, no. 1, pp. 1–12, 2017.

[17] V. Stanev et al., “Machine learning modeling of superconducting critical temperature,” npj Comput. Mater., vol. 4, no. 1, pp. 1–14, 2018.

[18] B. Meredig et al., “Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery,” Mol. Syst. Des. Eng., vol. 3, no. 5, pp. 819–825, 2018.

[19] A. O. Oliynyk et al., “High-throughput machine-learning-driven synthesis of full-Heusler compounds,” Chem. Mater., vol. 28, no. 20, pp. 7324–7331, 2016.

[20] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, “Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling,” Phys. Rev. X, vol. 4, no. 1, p. 11019, 2014.

[21] P. V Balachandran, A. A. Emery, J. E. Gubernatis, T. Lookman, C. Wolverton, and A. Zunger, “Predictions of new AB O 3 perovskite compounds by combining machine learning and density functional theory,” Phys. Rev. Mater., vol. 2, no. 4, p. 43802, 2018.

[22] C. Oses, C. Toher, and S. Curtarolo, “High-entropy ceramics,” Nat. Rev. Mater., pp. 1–15, 2020.

[23] G. Krauss, Steels: processing, structure, and performance. Asm International, 2015.

[24] G. B. Olson and C. J. Kuehmann, “Materials genomics: from CALPHAD to flight,” Scr. Mater., vol. 70, pp. 25–30, 2014.

[25] J. B. Haskins, A. E. Thompson, and J. W. Lawson, “Ab initio simulations of phase stability and martensitic transitions in NiTi,” Phys. Rev. B, vol. 94, no. 21, p. 214110, 2016.

[26] L. Sandoval, J. B. Haskins, and J. W. Lawson, “Stability, structure, and suppression of the martensitic transition temperature by B19’ compound twins in NiTi: ab initio and classical simulations,” Acta Mater., vol. 154, pp. 182–189, 2018.

[27] R. Zarnetta et al., “Identification of quaternary shape memory alloys with near-zero thermal hysteresis and unprecedented functional stability,” Adv. Funct. Mater., vol. 20, no. 12, pp. 1917–1923, 2010.

[28] A. N. Bucsek, G. A. Hudish, G. S. Bigelow, R. D. Noebe, and A. P. Stebner, “Composition, compatibility, and the functional performances of ternary NiTiX high-temperature shape memory alloys,” Shape Mem. Superelasticity, vol. 2, no. 1, pp. 62–79, 2016.

[29] J. Cui et al., “Combinatorial search of thermoelastic shape-memory alloys with extremely small hysteresis width,” Nat. Mater., vol. 5, no. 4, pp. 286–290, 2006.

[30] K. Otsuka and X. Ren, “Physical metallurgy of Ti–Ni-based shape memory alloys,” Prog. Mater. Sci., vol. 50, no. 5, pp. 511–678, 2005.

[31] M. Asai, “Graduation Thesis,” University of Tsukuba, 1982.

[32] J. Frenzel, A. Wieczorek, I. Opahle, B. Maaß, R. Drautz, and G. Eggeler, “On the effect of
alloy composition on martensite start temperatures and latent heats in Ni–Ti-based shape memory alloys,” *Acta Mater.*, vol. 90, pp. 213–231, 2015.

[33] Y. Zhou et al., “Strain glass in doped Ti50 (Ni50– xDx)(D= Co, Cr, Mn) alloys: Implication for the generality of strain glass in defect-containing ferroelastic systems,” *Acta Mater.*, vol. 58, no. 16, pp. 5433–5442, 2010.

[34] J. Ma, I. Karaman, and R. D. Noebe, “High temperature shape memory alloys,” *Int. Mater. Rev.*, vol. 55, no. 5, pp. 257–315, 2010.

[35] R. D. James and Z. Zhang, “A way to search for multiferroic materials with ‘unlikely’ combinations of physical properties,” in *Magnetism and structure in functional materials*, Springer, 2005, pp. 159–175.

[36] Y. Song, X. Chen, V. Dabade, T. W. Shield, and R. D. James, “Enhanced reversibility and unusual microstructure of a phase-transforming material,” *Nature*, vol. 502, no. 7469, pp. 85–88, 2013.

[37] H. Hou et al., “Fatigue-resistant high-performance elastocaloric materials made by additive manufacturing,” *Science (80-. ).*, vol. 366, no. 6469, pp. 1116–1121, 2019.

[38] L. Casalena et al., “Structure-Property Relationships of a High Strength Superelastic NiTi–1Hf Alloy,” *Adv. Eng. Mater.*, vol. 20, no. 9, p. 1800046, 2018.

[39] H. E. Karaca et al., “Effects of nanoprecipitation on the shape memory and material properties of an Ni-rich NiTiHf high temperature shape memory alloy,” *Acta Mater.*, vol. 61, no. 19, pp. 7422–7431, 2013.

[40] B. Amin-Ahmadi, T. Gallmeyer, J. G. Pauza, T. W. Duerig, R. D. Noebe, and A. P. Stebner, “Effect of a pre-aging treatment on the mechanical behaviors of Ni50.3Ti49.7–xHfx (x≤ 9 at.%): Shape memory alloys,” *Scr. Mater.*, vol. 147, pp. 11–15, 2018.

[41] B. Amin-Ahmadi, J. G. Pauza, A. Shamimi, T. W. Duerig, R. D. Noebe, and A. P. Stebner, “Coherency strains of H-phase precipitates and their influence on functional properties of nickel-titanium-hafnium shape memory alloys,” *Scr. Mater.*, vol. 147, pp. 83–87, 2018.

[42] S. Mills, “PhD Thesis,” Colorado School of Mines, 2019.

[43] “NiTiHf Shape Memory Alloys,” Citrination, 2018. [Online]. Available: https://citrination.com/teams/45/resources.

[44] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[45] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, “MIFS-ND: A mutual information-based feature selection method,” *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6371–6385, 2014.

[46] J. Benesty, J. Chen, Y. Huang, and I. Cohen, “Pearson correlation coefficient,” in *Noise reduction in speech processing*, Springer, 2009, pp. 1–4.

[47] ASTM, “Standard specification for wrought Nickel-Titanium shape memory alloys for medical devices and surgical implants,” 2012.

[48] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, 1987.

[49] J. M. Jani, M. Leary, A. Subic, and M. A. Gibson, “A review of shape memory alloy research, applications and opportunities,” *Mater. Des.*, vol. 56, pp. 1078–1113, 2014.

[50] J. Frenzel, Z. Zhang, K. Neuking, and G. Eggeler, “High quality vacuum induction melting of small quantities of NiTi shape memory alloys in graphite crucibles,” *J. Alloys Compd.*, vol. 385, no. 1–2, pp. 214–223, 2004.

[51] S. Gollerthan, M. L. Young, A. Baruj, J. Frenzel, W. W. Schmahl, and G. Eggeler, “Fracture mechanics and microstructure in NiTi shape memory alloys,” *Acta Mater.*, vol. 57, no. 4, pp. 1015–1025, 2009.

[52] M. Moshref-Javadi, S. H. Seyedein, M. T. Salehi, and M. R. Aboutalebi, “Age-induced multi-stage transformation in a Ni-rich NiTiHf alloy,” *Acta Mater.*, vol. 61, no. 7, pp. 2583–2594, 2013.

[53] O. Benafan, G. S. Bigelow, and D. A. Scheiman, “Transformation behavior in NiTi-20Hf
shape memory alloys—Transformation temperatures and hardness,” *Scr. Mater.*, vol. 146, pp. 251–254, 2018.

[54] T. Umale, D. Salas, B. Tomes, R. Arroyave, and I. Karaman, “The effects of wide range of compositional changes on the martensitic transformation characteristics of NiTiHf shape memory alloys,” *Scr. Mater.*, vol. 161, pp. 78–83, 2019.

[55] B. Meredig and C. Wolverton, “Dissolving the periodic table in cubic zirconia: Data mining to discover chemical trends,” *Chem. Mater.*, vol. 26, no. 6, pp. 1985–1991, 2014.

[56] G. Pilania, P. V Balachandran, C. Kim, and T. Lookman, “Finding new perovskite halides via machine learning,” *Front. Mater.*, vol. 3, p. 19, 2016.

[57] D. Xue et al., “An informatics approach to transformation temperatures of NiTi-based shape memory alloys,” *Acta Mater.*, vol. 125, pp. 532–541, 2017.

[58] M. Zarinejad and Y. Liu, “Dependence of Transformation Temperatures of NiTi-based Shape-Memory Alloys on the Number and Concentration of Valence Electrons,” *Adv. Funct. Mater.*, vol. 18, no. 18, pp. 2789–2794, 2008.

[59] N. Nayan, C. N. Saikrishna, K. V. Ramaiah, S. K. Bhaumik, K. S. Nair, and M. C. Mittal, “Vacuum induction melting of NiTi shape memory alloys in graphite crucible,” *Mater. Sci. Eng. A*, vol. 465, no. 1–2, pp. 44–48, 2007.

[60] S. Sarkar, X. Ren, and K. Otsuka, “Evidence for strain glass in the ferroelastic-martensitic system Ti 50− x Ni 50+ x,” *Phys. Rev. Lett.*, vol. 95, no. 20, p. 205702, 2005.

[61] D. Wang et al., “Strain glass in Fe-doped Ti–Ni,” *Acta Mater.*, vol. 58, no. 18, pp. 6206–6215, 2010.

[62] J. Lee, Y. Ikeda, and I. Tanaka, “First-principles screening of structural properties of intermetallic compounds on martensitic transformation,” *npj Comput. Mater.*, vol. 3, no. 1, pp. 1–14, 2017.

[63] “Scikit-learn.” [Online]. Available: https://scikit-learn.org/stable/.

[64] “Gaussian process model.” [Online]. Available: http://www.gaussianprocess.org/gpml/code/matlab/doc/.

[65] “Minimize negative log likelihood.” [Online]. Available: http://www.gaussianprocess.org/gpml/code/matlab/doc/.

**Acknowledgments**

We gratefully acknowledge support from the Department of Defense, Office of Economic Adjustment (grant no. ST1605-17-02) and the Colorado Office of Economic Development & International Trade (grant no. CTGG1 2016-2166).

**Supplementary Materials**

Fig. S1 Schematic of DSC (Differential scanning calorimetry) measurements and properties characterization for TTs and $\Delta$RT. The design targets TTs between 230–260 K for low temperature applications, at same time minimize $\Delta$RT.
Fig. S2 The bias-variance trade-off techniques to present the searching GPR optimal hyper-parameters $\theta = \{\sigma_n^2, \sigma_f^2, l\}$ for TTs modeling. The high bias region means under-fitting of the model and high variance region indicates over-fitting. The boundary of over-fitting and under-fitting in dash line presents the optimal hyper-parameters obtained from the conjugate gradients method.
Fig. S3 The TTs predictive uncertainty profiles for VAM (a, c, e) and VIM (b, d, f) methods under heat-treatment conditions Sol (1050 °C/0.5 h, WQ), Sol (1050 °C/0.5 h, WQ)+Aging (550 °C/3.5 h, AQ), and Sol (1050 °C/0.5 h, WQ)+PreAging (300 °C/12 h, AQ)+Aging (550 °C/3.5 h, AQ). The prediction region are 0≤Hf≤30 at.% and 49≤Ni≤52 at.%.

Fig. S4 (a-c) The ΔRT predictive uncertainty profiles for VAM synthesis ways under heat-treatment conditions Sol (1050 °C/0.5 h, WQ), Sol (1050 °C/0.5 h, WQ)+Aging (550 °C/3.5 h, AQ), and Sol (1050 °C/0.5 h, WQ)+PreAging (300 °C/12 h, AQ)+Aging (550 °C/3.5 h, AQ). The prediction region are 0≤Hf≤30 at.% and 49≤Ni≤52 at.%.
Fig. S5 Predictive tendency curves extracted from predictive ternary profiles of under (a-c) Sol and (d-f) Sol + Aging process conditions. (a, d) The variation of TTs with Hf content for various selected Ni contents; (b, e) variation of TTs with Ni content for different Hf contents; and (c, f) relative hysteresis $\Delta RT$ variations against Hf content change.

Fig. S6 Presentative chemically composition features’ distribution against the property variation. The lighter color indicates higher property value. (a, b) The low average atomic radius $R_{rad}$ is
necessary condition for achieving low TTs. The dependence of $R_{rad}$ on $\Delta RT$ is similar with the Hf variation tendency in main text Fig. 1. In addition, the graphs of average atom mass $M$, atomic number $Z$ and covalent radius $R_{cov}$ against property have similar tendency with $R_{rad}$, and they are not shown as here as redundancy. (c, d) Another typical feature Electronegativity $\chi$ dependence on properties shows the inverse relations with $R_{rad}$. The high $\chi$ value corresponds to low TTs. (e, f) The third typical of features, the total valence electrons $Te$ has no significant dependency on TTs. But $C_v$, which defined as linear relations $Te/Z$, shows greater dependence on TTs and exhibits similar tendency with $\chi$.

Fig. S7 Feature importance score using mutual information (MI) method for training on (a) binary NiTi, (b) Hf-low, and (c) Hf-high alloys.
Fig. S8 The DSC measurements for (a) 50Ni3Hf, (b) 50.4Ni3Hf, (c) 50.4Ni12Hf and (d) 50.5Ni11Hf under heat treatments Solution (1050 °C/0.5 h, WQ) + PreAging (300 °C/12 h, AQ) + Aging (550 °C/3.5 h, AQ).

Table S1. The descriptors or features generated through physics-informed feature engineering approaches, which include raw compositions, chemically compositions, synthesis ways, raw processes and physiochemical theories features.

| Feature categories | Feature symbol | Feature full name | Feature categories | Feature symbol | Feature full name |
|--------------------|----------------|-------------------|--------------------|----------------|-------------------|
| Elemental properties | Z | 1. Atomic number | Avg.Sf | 25. Average of shell number from f valence shells |
| | Gro. | 2. Periodic table column | Avg.f | 26. Average of electrons from f valence shells |
| | Per. | 3. Periodic table row | Avg.Us | 27. Average of s unfilled electrons |
| | M | 4. Relative atomic mass | Avg.Up | 28. Average of p unfilled electrons |
| | Mend | 5. Mendeleev number | Avg.Ud | 29. Average of d unfilled electrons |
| | Rcov | 6. Atomic radius | Avg.Uf | 30. Average of f unfilled electrons |
| | Rcore | 7. Covalent radius | Ni | 31. Nickel at% |
| Reactivity | e | 8. Valence | Ti | 32. Titanium at% |
| | Eea | 9. Electron affinity | Hf | 33. Hafnium at% |
| | IE | 10. Ionization energy | Syn. | 34. Synthesis ways (Syn.) |
| | χ | 11. Electronegativity Pauling | Tsol | 35. Solution temperature |
| | k | 12. Thermal conductivity | T-Tsol | 36. Transformed solution temperature |
| | ρ | 13. Electrical conductivity | tSol | 37. Solution time |
| | ΔHfus | 14. Heat of fusion | T-Tsol | 38. Transformed solution time |
| | ΔHvap | 15. Heat of vaporization | Tpre | 39. Pre-aging temperature |
| Thermal properties | Raw processes and physiochemical theories attributes |
|--------------------|----------------------------------------------------|
| $T_m$ 16. Melting point | $T-T_{pre}$ 40. Transformed pre-aging temperature |
| $T_b$ 17. Boiling point | $t_{pre}$ 41. Pre-aging time |
| $T_e$ 18. Total valence electrons | $T-t_{pre}$ 42. Transformed pre-aging time |
| Avg.Sx 19. Average of shell number from s valence shells | $T_{age}$ 43. Final-aging temperature |
| Avg.s 20. Average of electrons from s valence shells | $T-T_{age}$ 44. Transformed final-aging temperature |
| Avg.Sp 21. Average of shell number from p valence shells | $t_{age}$ 45. Final-aging time |
| Avg.p 22. Average of electrons from p valence shells | $T-t_{age}$ 46. Transformed final-aging time |
| Avg.Sd 23. Average of shell number from d valence shells | $F-c$ 47. Stress conditions |
| Avg.d 24. Average of electrons from d valence shells | $F-v$ 48. Stress values |