Integrating Materials and Manufacturing Innovation (2023) 12:301–320
https://doi.org/10.1007/s40192-023-00311-9

THEMATIC SECTION: 7TH WORLD CONGRESS ON INTEGRATED COMPUTATIONAL MATERIALS ENGINEERING

Tempered Hardness Optimization of Martensitic Alloy Steels

Heather A. Murdoch[1] · Daniel M. Field[1] · Benjamin A. Szajewski[1] · Levi D. McClenny[1,2] · Andrew Garza[3] · Berend C. Rinderspacher[1] · Mulugeta A. Haile[1] · Krista R. Limmer[1]

Received: 28 July 2023 / Accepted: 12 September 2023 / Published online: 26 October 2023

Abstract
A simple Gaussian process regressor (GPR) model is employed to predict steel hardness and toughness response for tempered martensitic steels. A dataset of over 2000 hardness values from over 250 distinct alloys was compiled, with the aim of incorporating a diverse set of quenched and tempered martensitic steels. The Izod impact toughness was included for over 450 of these alloy/temper conditions. The GPR exhibited an increase in accuracy for both the predicted hardness and Izod impact toughness over linear regression trained on the same dataset. Shapley additive explanations (SHAP) were used to assess the importance of the input features of tempering temperature, tempering time, and 15 elements. Tempering temperature and carbon content were the most important input features in all models. The relative importance of the other 14 alloying elements varied depending on the target property. The SHAP analysis highlighted the complex relationships between composition and mechanical properties that are able to be captured by machine learning approaches.

Keywords Tempered steel · Izod impact toughness · Shapley additive explanations · Martensitic steel

Introduction

Many empirical relationships have been developed over the years to describe processing conditions for steels. Most notably, equations for the martensite start (M_S) temperature [1–5], but also the austenitization temperature [1, 2, 6], tempering parameter [7], as-quenched hardness and hardenability [8], some of which are even included in ASTM standards. The appeal of traditional empirical models (e.g., additive models of single-factor terms such as linear regression) over physics-based thermodynamic calculations and data-driven complex machine learning (ML) approaches, is their interpretability, reusability, and applicability to small datasets. Furthermore, they can be easily extended by adding another term, e.g., for elements that were not considered previously or cross-correlative terms. More complex models like artificial neural networks (ANNs) or nonparametric models, such as Gaussian process regression, may require considerable amounts of data, fitting time, or both and remain largely black boxes [9]. On the other hand, very complex relationships can be modeled; handling the tradeoff between accuracy and complexity has a long history [10, 11]. A common problem of complex models is overfitting to the data, which can be mitigated by regularization and cross-validation methods [10].

A countervailing limitation of traditional empirical models is their failure to extrapolate correctly beyond the underlying composition ranges for which they are devised. Hence, a patchwork of partially overlapping composition ranges has been used to fit such models, leading to competing parameterizations in regions of overlap. One example is the more than twenty equations for the M_S temperature [3–5]. Nonetheless, confined to interpolation, these empirical equations often perform very well on the metric to which they were fit. In fact, a comparative study between ML models trained on a comprehensive compositional dataset and a simple linear regression model fit to the appropriate subsets of the data found that the more complex models did not significantly reduce the relative error over the simple linear regression,

✉ Heather A. Murdoch
heather.a.murdoch.civ@army.mil

[1] DEVCOM Army Research Laboratory, Aberdeen Proving Ground, MD 21005, USA

[2] Oak Ridge Associated Universities (ORAU), Aberdeen, MD 21005, USA

[3] University of California, Merced, USA

while simple random forest regression (RFR) and adaptive boosting showed improvement on both training and test sets [12]. The same study showed that traditional empirical models from the literature performed better with respect to mean absolute error on the subsets on which those models were trained, although it is not clear how much of the data should be considered outside of the training set [12]. In a study predicting hardness of low alloy steels, ANN and support vector regression (SVR) models performed significantly worse than a traditional empirical equation developed from the same dataset [13]. It is worth noting here that the SVR fit a statistically robust linear model, which is a considerably simpler approach than the traditional model. Furthermore, the ANN did not incorporate in its final activation that Vickers hardness may not be negative, which inhibits the ANNs ability to model the relationship well. On the other hand, RFR and k-nearest neighbor (kNN) models made slight improvements over the already very accurate traditional model's prediction [13]. In other cases, ANNs sometimes performed better and sometimes worse than linear regression in predicting tensile strength of low alloy steels [14]. In general, it may be concluded that complex ML models require judicious application to be of general use.

As intimated above, more complex models require larger datasets to model more complex relationships. These datasets must not just be larger, but also representative of the modeling domain and its complex relationships. For non-Bayesian models, the dataset also has to be large enough for a meaningful cross-validation and generalization error estimates [10]. A rule of thumb is an order of magnitude more samples than input features. Despite the recent proliferation of ML studies predicting properties of steel (and other metals), many of these models are trained on compositionally limited datasets, constraining their applicability in the same way as existing traditional empirical models. As an extreme example, some ML studies have even limited their dataset to *a single* alloy with multiple processing conditions, employing an ANN to predict tempered hardness of AISI 1045 with only 18 training points [15]. Another example is the application of an ANN with 30 parameters to predict hardness in 9 wt% Cr steel with only 36 training points [16]. Obviously, overfitting is a considerable risk with complex, expressive models and such limited data. The low diversity of the dataset also greatly restricts the domain in which these models can be expected to give accurate predictions.

On the other hand, large datasets of one particular family of steel can provide multiple, high-fidelity property predictions. An example dataset is that of ferritic creep resistant (9-12Cr) which has been used extensively: for robust comparisons of ML methods, studying the impact of various input features including calculated and measured microstructures, and prediction of multiple properties including creep life [17], yield [18], rupture [19], and hot strength [20]. However, since the training data is comprised of only one class of steel with a consistent microstructure that produces a singular strengthening mechanism, the model is expected to falter when extended to other classes of steels. Even in a subset of steels with the same strengthening mechanism, e.g., low C stainless steels with intermetallic (rather than carbide) strengthening phases, training on the subset of the data with one dominant phase (R-phase) and then making predictions on the subset of data with different dominant phases (Ni_3Ti and Cu clusters) resulted in a steep decline in predictive capability of the support vector regression model. This occurred even while including physical metallurgy input parameters such as the equilibrium volume fraction of the intermetallic phase in addition to the compositions in order to describe the mechanism more specifically [21]. Even when a dataset is comprehensive compositionally, it may have other specific features that limit extensibility. For example, in their predictions of tensile strength, Jiang et al. [22] included 23 processing parameters including the power rate of multiple fans on a production line. Xie et al. also used multiple plant and production line inputs in their ML prediction of several mechanical properties of hot rolled low alloy steel plate [23].

In this work, we employ the simplest possible inputs and perform a direct comparison between traditional empirical modeling (i.e., linear regression) and an ML model trained on the same dataset. The dataset is comprised of numerous classes of quenched and tempered martensitic steels, from simple Fe–C to high-C tool steels, including both standard AISI grades and experimental alloys, and low strength through ultra-high strength steels (UHSS). We also publish our dataset for reference in future studies involving more sophisticated ML models.

In quenched and tempered martensitic steels, the steel is first heated above the austenite transformation temperature where 100% austenite phase is achieved; second, it is quenched rapidly, forming martensite when it reaches the martensite start (M_S) temperature on cooling. The austenite to martensite transformation proceeds until reaching the martensite finish (M_f) temperature possibly resulting in incomplete martensite transformation, and retained austenite, if the M_f is below room temperature. Third, the tempering process is applied to generate the final microstructure of tempered martensite that is generally toughened by finely dispersed phases. For low alloy quenched and tempered steel systems, strength is understood to be primarily derived from supersaturated carbon in the as-quenched condition. Carbon super-saturation within the martensite lattice octahedral sites causes a significant hydrostatic stress and leads to tremendous increases in strength. The decrease in strength during tempering is associated with the formation of transition carbides that reduce the degree of C super-saturation and is also accompanied by an increase in toughness.

There are a few existing relationships to describe the hardness of low alloy steels in their as-quenched [8] and tempered [7, 24, 25] state as a function of composition. The as-quenched hardness prediction, for example, is based solely on the carbon content with an upper bound of 0.61 wt% C after which the hardness is assumed to saturate [8]:

$$H[\mathrm{HRC}] = 33.087 + 50.723X_C + 33.662X_C^2 - 2.7048X_C^3 - 107.02X_C^4 + 43.523X_C^5 \quad (1)$$

The limit placed at 0.61 wt% C is due to the severe depression of the M_s and M_f temperature from further additions of C leading to an incomplete transformation of austenite to martensite. The amount of retained austenite increases nearly linearly as the transformation temperatures are lowered and has been shown to be directly correlated to the carbon content [26].

Elements other than C, such as Mo and V, can provide solid solution strengthening in low temperature tempering regimes; however, at higher tempering temperatures they contribute to particular carbide formations that can either decrease or increase hardness depending on their size and location [27, 28]. Compositional variations can be utilized to form advantageous carbides, or slow the formation of deleterious carbides, as in the case of Si which inhibits the formation of cementite by changing the activity for C [27]. Alloying strategies also include mitigating detrimental effects of residual impurity elements. For example, Ti and Al are utilized to reduce the effects of O and N by forming fine particles that are then used as grain pinning particles. The effects of S are commonly addressed by the addition of Mn to form a ductile second phase, MnS, that can be relatively innocuous to the steel due to its ability to be deformed during hot working of the steel. Residual Cu can hinder a steels ability to be hot-rolled due to the phenomenon known as hot shortness [29]. On the other hand, Cu is purposefully added to form fine precipitates in some UHSS alloys where they provide the primary strengthening mechanism in contrast to the supersaturation of carbon compared to the low alloy steels. From this brief discussion, it can be seen that the role of individual elements can be multifaceted which provides a good example for comparing the application of traditional empirical (e.g., linear regression) to machine learning approaches.

Further, mechanisms that contribute positively to one mechanical property may negatively impact another. As an example, consider the tempered strength and toughness for AISI 4130 in Fig. 1 where the strength monotonically decreases with increasing tempering temperature, but impact energy has a complicated trend. As a consequence, empirical relationships for hardness and impact toughness do not have a simple inverse relationship due to their different physical mechanisms. While there are several existing empirical relationships for tempered hardness [7, 24, 25], the authors are unaware of any similar relationships for impact toughness due to its complexity. Several ML studies have explored toughness in a limited way, either with a small number of alloys [30, 31], or by only investigating a single class of steel [32, 33]. An ML study of Charpy impact toughness with a comprehensive compositional dataset resulted in good predictions only if additional mechanical test measurements were used as inputs in addition to the composition [34]. However, these additional inputs were the tensile strength and reduction of area, which leads to large material consumption to obtain representative test results. When using an existing dataset where tension tests were already performed for the material of interest this is a reasonable approach, but for a new alloy/temper not in the dataset, it would require less time and material to just perform the Charpy impact testing. In studies where multiple mechanical properties were predicted, impact toughness was by far the poorest prediction of the targets [23, 35]. A study with large amounts of industrial data was able to make fairly good predictions [36].

The existing empirical relationships for the hardness of the tempered state are discussed in more detail in a later section; however, they all have a similar functional form in that they predict hardness as a function of alloy content, time, and temperature. They all have generally good fits within the compositional ranges of the alloys used for the regression analysis but are not likely to be valid beyond this low alloy range. For example Kang [7] used steels with the elements C, Mn, Si, Ni, Cr, and Mo only, omitting many important alloying contributions for UHSS and other steels such as Co and V. Mukherjee [24] had the most comprehensive dataset from an elemental perspective, but most alloys had a maximum concentration of ~ 2 wt% for most elements; they also had some severe outliers with predictions using their low tempering temperature

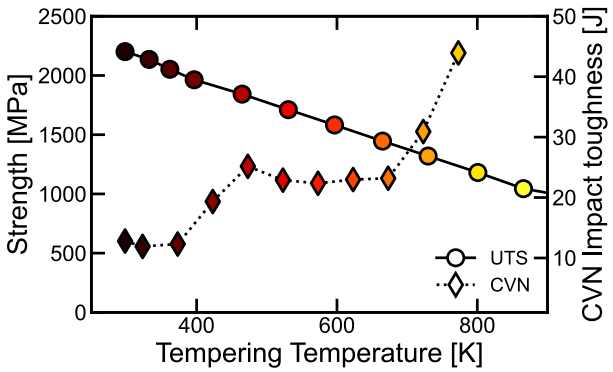

Fig. 1 Ultimate tensile strength (UTS) as a function of tempering temperature in AISI 4130 steel compared to Charpy (CVN) impact toughness. Points are colored by temper temperature. The UTS has a straightforward trend with respect to processing condition compared to the CVN. Data from [28]

(< 573 K) data. The model of Athavale [25] is somewhat limited in that it utilizes the incremental hardness increase of multiple alloying elements, but only uses data from one hour tempers (other temper times are predicted by integrating the Jaffe-Holloman tempering parameter).

The paper is organized as follows. First, we collect a comprehensive dataset of tempered martensitic steels of many classes that includes both low and high alloy steels. Second, we evaluate the effectiveness of a traditional empirical linear regression approach as a baseline. Third we apply a machine learning approach to the tempering dataset. Since the traditional empirical models exhibit clear contributions based on alloying content, i.e., in the form of the coefficients, we use an ML modeling approach where the impact of the input features can be quantified. The feature importance is discussed using Shapley values [37] which can be calculated for both linear regression and Gaussian process regressor (GPR). The bulk of the dataset and analysis is using the target property of hardness, but the approach to also extended to a smaller dataset of a more complex mechanical property, here Izod impact toughness.

Data Summary

The intent of this dataset is to incorporate the breadth of quenched and tempered martensitic steels, and as such, encompass a large composition space. Fifteen alloying elements are tracked to describe the composition. The target property of this dataset is the hardness. The secondary target property of Izod impact toughness was also collected from handbook data where it existed.

We attempt to incorporate the data used and cited in previous empirical [7, 24, 25] and other modeling (e.g., [21]) works to facilitate comparisons; however, most did not publish their datasets. In some cases, these previous modeling papers also included propriety industry data. Further, some of the literature data used in the previous datasets was discarded during our data assessment step described in a following section. A full description of the dataset is in the online supplementary material. The data in this dataset was collected from literature and handbook sources [25, 38–55] and is available on Materials Commons [56].

The handbook and literature data collected were assessed for compatibility with the desired model bounds. First, to limit this model to martensitic quench and tempered steels, mixed microstructure (e.g., microstructures of austenite + ferrite, martensite + austenite, bainite, etc.) effects were removed. Second, the remaining dataset was divided into two classes, low and high alloy steels, based on their alloying content.

Data Processing and Assessment

To rule out the effects of mixed microstructure on the mechanical properties, we assess the collected data (~ 3000 alloy, temper, time combinations) for adherence to martensitic structure. The original data source is checked for data on microstructural features. If the retained austenite fraction is reported in the original source, a threshold of ≥ 3% retained austenite is considered to be a mixed microstructure. This threshold is based on an observable fraction via X-ray diffraction (XRD). For example, it is noted in Grange et al. [53] that their high carbon (0.5, 0.72, 0.98 wt% C) steels had measurable amounts of retained austenite (3, 7, 13% RA, respectively); these datapoints are therefore dropped from the final dataset. Some of the data used in the Mukherjee empirical fit [24] was found to be mixed microstructure including reported mixtures of lath martensite and bainitic ferrite [38] or reported retained austenite fractions as high as 17% [57], so these were not included in our dataset despite being used in previous models.

Even if the microstructure is not directly reported, the likelihood of mixed microstructure being present can be estimated via the hardenability of the steel and the dimensions of the heat-treated part from which mechanical test coupons were extracted. For example, in the Modern Steels handbook [52] the mechanical property data is reported for various round diameters of steel bar, where the properties were reported as being measured from the center of the bar. Alloys with low hardenability and slow cooling rates (e.g., thick sections) will not form 100% martensite throughout the component, leading to a mixed microstructure in the sample incompatible with this model. To assess the probability of mixed microstructure in such parts, we consider the dimension of the component relative to the ideal diameter (DI) for the steel. The DI is calculated using the methodology of ASTM 255 [8]. If the dimension of the part is larger than the calculated DI, it is likely that the center of the part where samples are obtained is not 100% martensite and is discarded from the dataset. Data reported from Jominy end quench tests, e.g., [25], will also vary in microstructure along the length of the test coupon and the measured properties as a function of the cooling rate; therefore only the J1 position is utilized due to it being the highest cooling rate position.

The data were all converted to the same units: degrees Kelvin for the tempering temperature, seconds for the tempering time, Joules for the Izod impact toughness, and Vickers hardness (HV). The hardness values were all converted to Vickers using the methods from ASTM E140-12b [58]. In the cases that the data does not have an explicit temper time (e.g., some handbook data), we assume a time of 1 h, which is consistent with the median value of the data that has a reported time and is congruent with typical industrial

practice [59]. Data with assumed times are annotated in the dataset if other users would prefer to discard the time data.

Low Versus High Alloy

In processing the data, we partition the dataset into 'low alloy' and 'high alloy' steels, with the intent of thresholding between the steels that most empirical relationships have likely been fit to, and the more diverse range of steels which our dataset encompasses. The strengthening mechanisms for the low alloy steels will be similar to each other, in that they have a more limited number of possible carbide formers; in contrast, the high alloy steels could have any range of carbides, other non-carbide precipitates (e.g., Cu clusters etc.), etc. each of which influence the strength. As there is not a universally accepted definition for high alloy steels, we develop a quantitative threshold based on hardenability for whether a steel is considered low or high alloy to segment the dataset. There are two relevant empirical equations for hardenability in the ASTM A255 "Standard Test Methods for Determining Hardenability of Steel" [8] which we use to determine the low alloy threshold. The three criteria that must be met to be considered a "low alloy steel" in the dataset are:

1. The carbon content must be below 0.61 wt%; this criterion is based on the saturation of the hardness of as-quenched 100% martensite microstructure in ASTM A255 Table 7 [8]. Above 0.61 wt% C the as-quenched hardness was no longer reported to increase as a function of carbon content as noted previously.
2. The composition must be within the range defined in Table 1, based on DI calculation limits from ASTM A255 with other allowables for elements not originally included in the standard: W, Co, Ti, and Al. The hardenability calculator was designed for a subset of commercial steels that did not include these now common additions in low alloy steels. Ti and Al are added in small amounts (< 0.03 wt%), termed "micro-alloying additions" to remove N in solution by forming TiN and AlN to act as pinning particles for fine grain practice [29]. Co and W are found as residuals from the scrap melting process and are removed to < 0.1 wt% during melt refining; these low levels are innocuous to the steel's performance. An allowance for Sulphur was also made as it can be purposefully added to alloys, e.g., the AISI 1100 series steels, to improve machinability without significantly affecting hardness.
3. The calculated DI must be less than or equal to 7″, the maximum dimension for which dividing factors have been tabulated in ASTM A255.

Table 1 Composition limits for low alloy steels, modified from ASTM A255 multiplying factors

Element	Maximum (wt%)
Mn	1.95
Si	2.0
Ni	3.5
Cr	2.5
Mo	0.55
Cu	0.55
V	0.2
C	0.9
W*	0.1
Ti*	0.03
Co*	0.1
Al*	0.03
B*	0.003
N*	0.02
S*	0.6

*These elements are not in ASTM A255

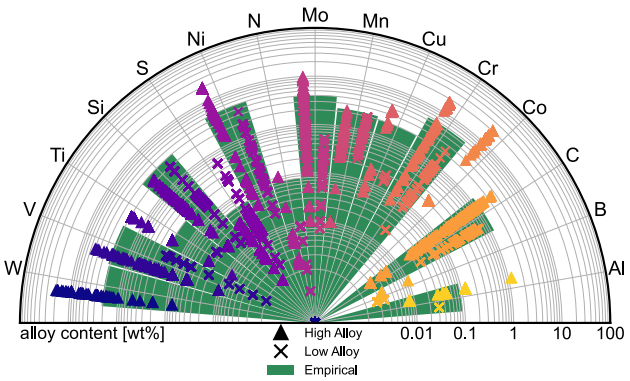

Fig. 2 Full dataset compositions ranges. The alloy content (in wt%) of each element contained in each sample is denoted with a ▵(high alloy) or an X (low alloy) with points colored by element for readability. The composition space examined within any of the previous traditional models [7, 24, 25] is highlighted in green

Final Dataset Description

After processing, the dataset is comprised of 2004 measured hardness values of quenched and tempered martensitic steel comprised of 61% low alloy steels and 39% high alloy steels. Within this dataset there are 269 distinct alloy compositions that have between 1 and 70 different tempered conditions associated with the alloy. The tempered processing conditions are described by temperature and time, and the composition of the alloy is tracked for 15 elements. A detailed summary of the dataset feature ranges is provided in the online supplementary material, Tables S1–S3. The compositional values are indicated in Fig. 2, with distinctions between high and low alloy subsets. The maximum ranges

for each element that have been included in one of the three existing empirical models (online supplementary Table S-4) are shown in Fig. 2 as shaded regions.

Empirical Models/Linear Regression

Existing Empirical Models

The foundational work of Grange et al. [53] established incremental changes to the hardness based on composition at several tempering temperatures. They proposed to predict the tempered hardness by a linear combination of these incremental alloy relationships. These relationships were only described for a tempering time of one hour and at discrete tempering temperatures. Athavale [25] incorporated the theories of Crafts and Lamont [60] and the Jaffe-Holloman tempering parameter [50] in order to extend the applicability of Grange et al.'s work. The tempering parameter, TP, is:

$$\mathrm{TP} = T\left(\log_{10}(t) + C\right) \tag{2}$$

where T is the temperature in Kelvin, t is the time in seconds, and C is a constant that is dependent on the composition. In Athavale's empirical model, C is described only as a function of the carbon content [25]. The model of Kang [7] also uses the tempering parameter, but fit the composition dependence of C to a linear relationship of the alloying elements within the overall expression for the tempered hardness, TH, in units of Vickers:

$$\mathrm{TH[HV]} = \left(1542.97 - \frac{25.31}{X_C}\right)\exp\left(-1.23\times10^{-4} * \mathrm{TP}\right) \tag{3}$$

Mukherjee [24] did not use the tempering parameter but did use the form of log(t) for their time inputs. They explored 16 different regression equation forms, with the selected equation being:

$$\ln(\mathrm{TH}) = A_0 + A_t \ln(t) + A_T T + \sum_{i}^{i=n} A_i X_i \tag{4}$$

where A_t and A_T are the coefficients for time and temperature, respectively, and A_i are the coefficients for each alloying contribution for element i. Their model had a high correlation coefficient ($R^2 = 0.924$) for the predicted and measured hardness for their training dataset but had notable extreme outliers at low tempering temperatures (< 573 K) and lower correlation for their test dataset ($R^2 = 0.796$).

The maximum ranges of the datasets used for these three empirical models are shown graphically in Fig. 2 as shaded regions with details provided in the online supplementary Table S-4. The notable differences between the coverage of the existing empirical equations and the dataset in this work are the complete absence of cobalt and the very low limits for W, V, Cr, Ti, Ni, and Mo. These elements are important components of martensitic stainless steels, UHSS with high fracture toughness K_{1C}, as well as high temperature tool steels needed for extreme environments, in which refractory alloy carbides are required to retain strength at elevated temperatures.

These three existing equations have been applied to our dataset, with the results shown in Fig. 3. First, we segment our dataset by the limits for each equation so as not to exceed the original bounds. This results in reasonable predictions, quantified in Table 2; Mukherjee's work had reported R^2 of 0.924 for their training set and 0.796 for their test set, while Kang reported a training set R^2 of 0.941 and did not use a test/validation set. Therefore, despite limiting our dataset to the original compositional limits of the empirical equations, there is a decrease in predictive capability compared to the original training set.

These empirical predictions break down, as expected, when used to predict the tempered hardness for the full dataset including high alloyed steels. For Athavale, the high alloys are generally underpredicted as the model does not have terms for several elements, resulting in zero contribution; however, the predictions are still within the realm of observable measurements. Conversely, for Mukherjee and Kang there are many unphysical predictions (resulting in extremely low R^2 below 0.03) such as hardness values in excess of 100,000 HV and negative hardness values. This is perhaps not unsurprising, due to the extreme extrapolation occurring for some of the elements in the empirical equations. We therefore fit a new empirical equation to our larger, more comprehensive dataset. This will also be used to compare directly to the GPR model trained on the same dataset.

New Linear Regression Models

The full dataset was split into training (80%) and test (20%) sets, with the empirical equation fit on the training set. A few different forms were considered for the empirical equation, in the manner of [24] with the best fit being of the following form with the coefficients shown in Table 3:

$$\mathrm{TH[HV]} = P_o + \sum_{i}^{i=n} X_i C_i + X_T T + X_t \log_{10} t \tag{5}$$

The predicted hardness from the traditional empirical Eq. (5) is shown versus the measured hardness in Fig. 4. Assessing the coefficients in Table 3, it is reassuring that the time and tempering temperature have a negative effect on strength that is obvious to the standard heat treatment of quenched steels. The strong positive effect of C on hardness

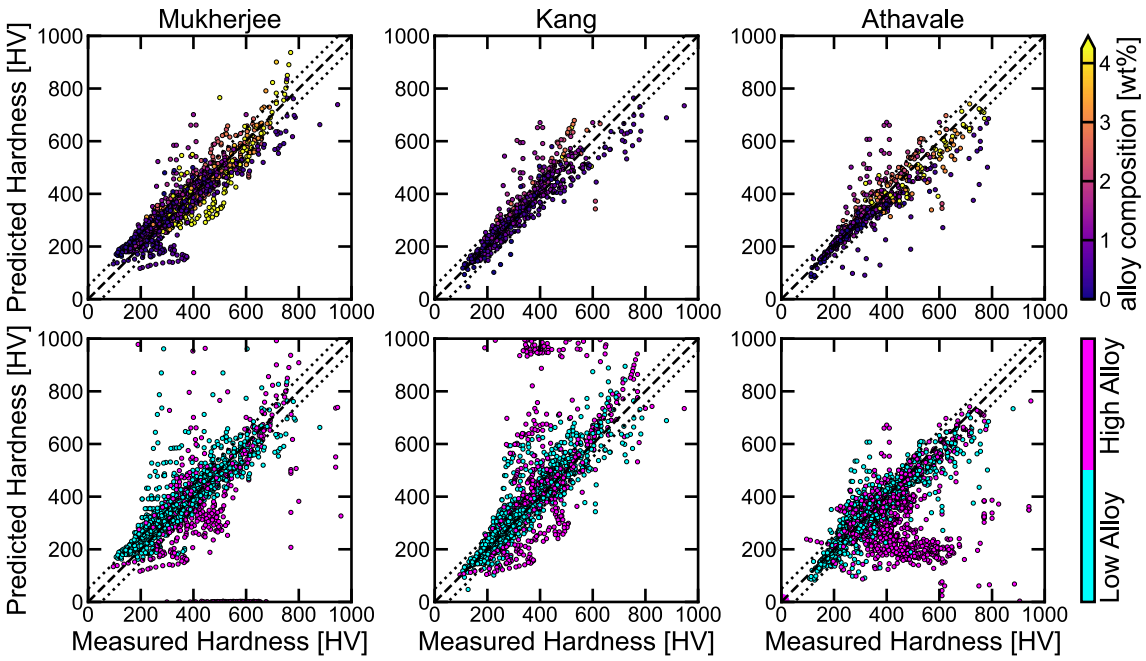

Fig. 3 Predicted vs measured hardness using existing empirical models for (top) the limits stipulated by the original model and (bottom) the full dataset colored by low-/high-alloy split. The metrics for the predictions are in Table 2

Table 2 Statistics for linear regression equations

Model	Original equation limit			Low alloys subset			Full dataset		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
Mukherjee	0.796	42.8	63.1	0.728	44.2	72.6	0.024	1278.0	11,562.7
Kang	0.826	37.1	59.5	0.814	39.7	62.4	0.001	1577.9	7279.8
Athavale	0.784	41.45	67.5	0.798	37.5	57.6	0.245	86.3	147.9
This work				0.835	40.2	53.9	0.796	49.5	67.7

is also an obvious result. We defer discussion of the impacts of different inputs from the linear regression model with the GPR model to the Feature Analysis section. In terms of statistics (Table 2), this equation is an obvious improvement over prior models, but it still has quite a large MAE and RMSE.

The linear regression for the impact data is performed for the following form, where the Izod value is scaled by the natural log to maintain a positive prediction:

$$\ln(\text{IZOD}[J]) = P_0 + \sum_{i}^{i=n} X_i C_i + X_T T \tag{6}$$

The coefficients for Eq. (6) are in Table 3; several are zero because the Izod dataset does not contain alloys with these elements, but they are left in for consistency with the full dataset and the other tables in the paper. It is immediately apparent that the relationships between many of the input features have opposite impacts for predicting hardness versus toughness. The coefficients for temper temperature and carbon have opposite signs from the coefficients for the hardness prediction, as expected, i.e., increased carbon content is known to increase strength but decrease toughness while increased temper temperature decreases strength while increasing toughness.

While the hardness prediction from Eq. (5) is obviously a vast improvement over existing empirical equations, particularly for high-alloy steels, it still has relatively high error. Further, the Izod prediction (Fig. 5), while also not an unreasonable R^2 value, has almost half of the predicted points outside of ± 10 J from the measured value (with a MAE of 12.3 and RMSE of 16.5). Therefore, we move on to performing Gaussian Process Regression on the dataset in order to improve the predictions of both hardness and Izod.

Gaussian Process Regression

The input features are comprised of tempering time, temperature, and alloy compositions in wt%. The target variables are hardness and Izod impact toughness. Since all of these

Table 3 Coefficients for empirical Eqs. (5) and (6)

Prediction	HARDNESS	IZOD
Coefficient	HV	ln (IZOD[J])
P_0	776.91	1.891
X_T [K]	−0.61	0.00444
X_t [s]	−18.41	0
X_i [wt%]		
C	193.29	−3.239
Mn	8.53	0.004
Si	63.73	−0.488
Cr	0.18	−0.019
V	46.84	−0.198
W	16.18	0
Mo	35.97	0.685
Ni	9.96	0.008
Co	1.31	0
Al	71.63	0
Cu	19.05	0
Ti	−18.11	0
N	−338.72	0
B	1991.27	0
S	−129.00	−0.943

variables describe a physical quantity with an associated length scale, we apply the following rescaling procedures to preprocess our data. First, since both target variables must be positive, we apply a logarithm scaling. Specifically, we rescale each target y as $x = \ln(y)$. In addition, we rescale time as $x = \ln_{10}(y)$ similar to convention in the literature [41] and consistent with Eq. (2). In [33], normalizing the time via logscale was necessary in order to enable their ANN to make predictions. Second, for all variables, for which we denote a sample by x, we apply a Z-score normalization such that the scaled distribution of a variable has a mean of zero and standard deviation of unity, i.e.,

$$Z(x) = \frac{x - \mu}{\sigma} \tag{7}$$

where μ and σ represent the mean and standard deviation of all samples for each variable, respectively. The scaled distributions for each of the input features and hardness are shown in Fig. 6; the distributions for the Izod impact dataset are shown in online supplementary Figure S-2.

With the rescaled data, we employ a Gaussian process regressor (GPR) to compute probabilistic predictions of targets given the input features. The GPR requires a predefined kernel function specifying the covariance matrix. We

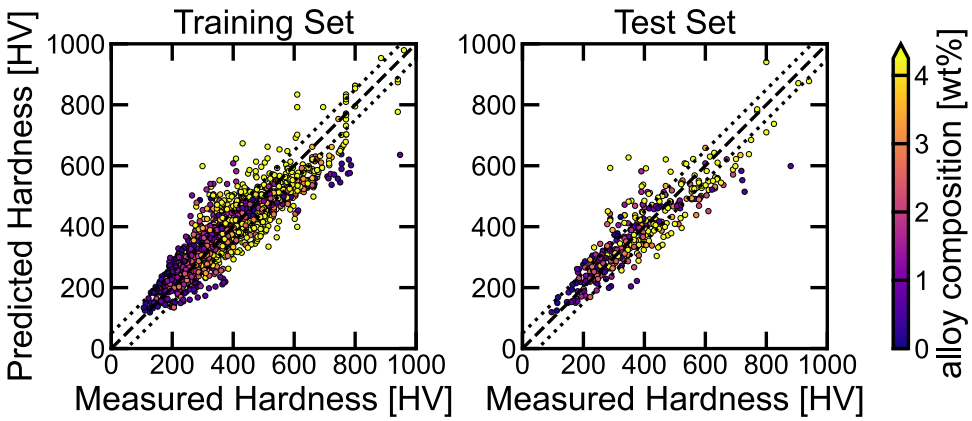

Fig. 4 Predicted vs measured hardness using Eq. (5) for the training ($R^2 = 0.795$) and test ($R^2 = 0.796$) set split of the full dataset

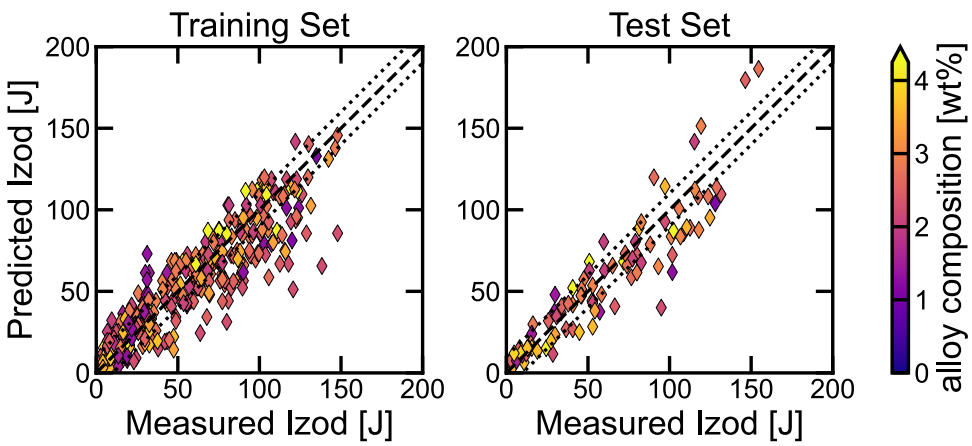

Fig. 5 Predicted vs measured Izod values from linear regression fit, Eq. (6). The R^2 value is 0.822 for the training set and 0.843 for the test set

employ a general kernel form comprised of three fundamental kernels and five hyperparameters:

$$\begin{aligned} K\left(x_i, x_j;\theta\right) =& \theta_1 \mathrm{RBF}\left(x_i, x_j; l_{\mathrm{RBF}}\right) \\ &+ \theta_2 \mathrm{Dot\ Product}\left(x_i, x_j; \sigma_0\right) \\ &+ \mathrm{White\ Noise}\left(x_i, x_j; l_{\mathrm{Noise}}\right), \end{aligned} \tag{8}$$

where the hyperparameters are denoted by $\theta \equiv \left[\theta_1, l_{\mathrm{RBF}}, \theta_2, \sigma_0, l_{\mathrm{Noise}}\right]$. Each of the three kernels may be expressed as:

$$\mathrm{RBF}\left(x_i, x_j; l_{\mathrm{RBF}}\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2l_{\mathrm{RBF}}^2}\right) \tag{9a}$$

$$\mathrm{Dot\ Product}\left(x_i, x_j; \sigma_0\right) = \sigma_0 + x_i \cdot x_j \tag{9b}$$

$$\mathrm{White\ Noise}\left(x_i, x_j; l_{\mathrm{Noise}}\right) = l_{\mathrm{Noise}}\delta\left(x_i, x_j\right) \tag{9c}$$

In Eqs. (9), ||·|| denotes the Euclidean distance, · denotes the dot product, and δ denotes the Kronecker delta which is unity when both arguments are equal, and zero otherwise. For training the GPR, fitting the covariance matrix to data via log maximum likelihood, and finally making predictions, we employ the python module Scikit learn [61].

Cross-Validation

The objective in optimizing Eq. (8) is twofold. First, the kernel hyperparameters, θ, must optimize the prediction quality of our model. Simultaneously, we must provide an accurate assessment of the performance of the optimized kernel against data which has been excluded from the optimization process. To accomplish these two tasks, we employ a nested cross-validation scheme comprised of an inner hyperparameter optimization step and a separate outer validation step with fixed hyperparameters.

The nested K-I cross-validation algorithm is comprised of three computational loops. The outer most loop cycles over all combinations of trial kernel functions, specifically initializing $\theta_1, \theta_2, l_{\mathrm{Noise}} \in \{0, 1\}$. For each superposed kernel function, the second internal loop begins by splitting the data into *K—folds* = 5, with test fraction: $1/K - folds = 20\%$ and training fraction $1 - 1/K - folds = 80\%$. The *K-fold* test fraction is set aside and the remaining *K-fold* training fraction is further partitioned in the third internal loop over *I-folds* ≡ 1 − *K-folds* = 4. The *K-folds* training data are partitioned into *1/I-folds* = 25% *I-fold* test data which comprise 20% of the dataset and 1 – 1/*I-folds* = 75% *I-fold* training data which comprise 60% of the dataset. The *I-fold* training data are employed to both optimize the hyperparameters, θ, via log maximum likelihood and populate the covariance matrix, simultaneously. The *I-fold* test data are subsequently employed to perform an initial model performance evaluation via mean squared error:

$$\mathrm{MSE}\left(\theta_i\right) = \frac{1}{n}\sum_{i=0}^{n}\left(y_i - \bar{y}\left(\theta_i\right)\right)^2 \tag{11}$$

Within this relation y_i and $\overline{y_i}$ denote observed and predicted target variables, respectively. This same loop over *I-folds* is repeated three additional times such that we obtain $\mathrm{MSE}(\theta_{\mathrm{i}})$ for $i = 1, \ldots, 4$ and the parameter set $\theta_{\mathrm{opt}} \equiv \mathrm{argmin}\ \mathrm{MSE}(\theta_i)$ which minimizes the mean squared error is stored for this trial kernel. The covariance matrix is subsequently populated with the *K-folds* training data, with θ_{opt} held fixed. An outer model performance test is performed employing Eq. (11), and calculating a similar MSE against the outer test data, i.e., *K-folds* test. Given the inner validation error and the outer test error, the total cross validation score can be computed as

$$\mathrm{CV[MSE]} = \frac{\mathrm{MSE}^{\mathrm{outer}}\left(\theta_{\mathrm{opt}}\right) + \mathrm{MSE}^{\mathrm{inner}}\left(\theta_{\mathrm{opt}}\right)}{2} \tag{12}$$

The inner validation and outer test dataset have the same number of samples, n. With *K-folds* = 5, five cross-validation scores are computed for each superposed kernel combination. The obtained cross-validation means and standard deviations are shown in Fig. 7 for each kernel tested and for the hardness, impact, and total datasets. In the interest of thoroughness, we have also included a constant kernel [Constant(), not shown in Eq. (8)] in this study.

A benefit of using GPR for predicting mechanical properties is the straightforward way in which it can account for noise/variance/spread in the experimental measurements. Each measurement is described by a Gaussian distribution with a given mean and variance, where the default assumption is essentially no variance. In Scikit-learn [61] the variance is passed as α with the default value being 1×10^{-10} for example. This parameter can be used to specify the variance for each individual measurement passed into the model; however, as only a few of the data sources in report the variance of their measurements we therefore make an estimate of the variability based on ASTM standards for measurement [62, 63]. For hardness, this variability is on the order of 1–5% of the measurement [62] and 3–6% for the impact toughness [63] depending on whether the measurement is on the low or high end of the scale. Assuming the observations are $y_i \pm \alpha_i$ with variance α_i (and then rescaled on a log-scale as defined above for the targets), the hardness predictions using an α based on the ASTM are compared with the default value for α (1×10^{-10}) [61] and shown in

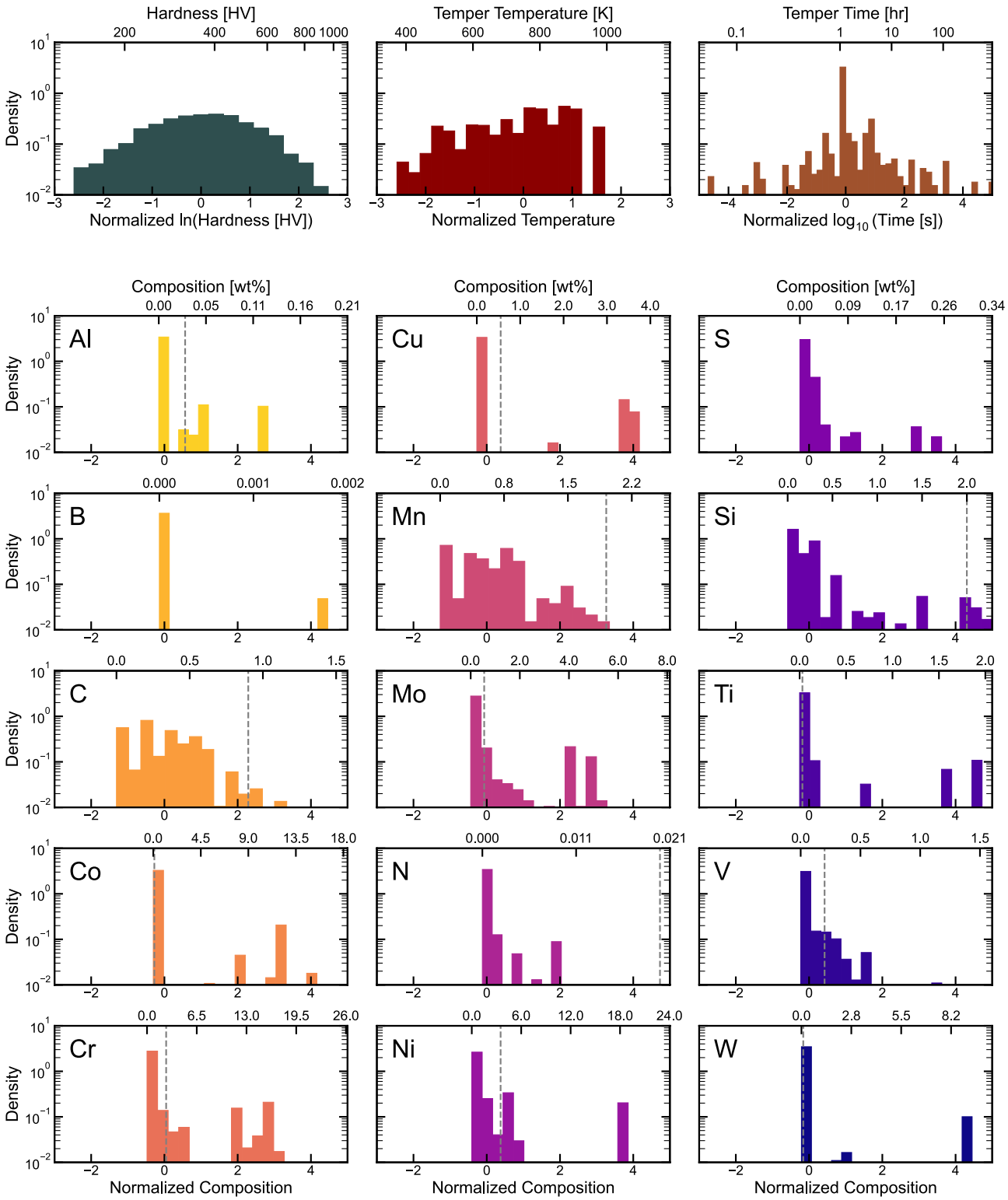

Fig. 6 Normalized distributions of input features and hardness. The elemental thresholds of Table 1 are indicated by grey dashed lines for reference. Normalized compositions greater than five standard deviations from the mean are excluded from the figure

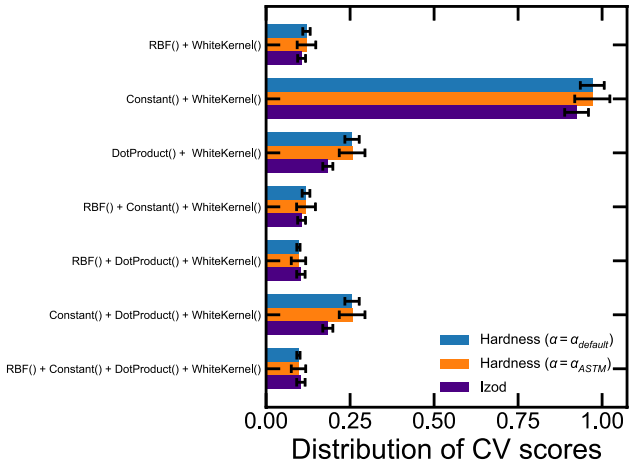

Fig. 7 Kernel comparison for predictions of hardness and Izod. Comparison of noise dependence for hardness prediction is included. A constant $\alpha = 1 \times 10^{-10}$ is compared to an observation dependent α

Table 4 Optimized kernels

Target value	θ_1	l_{RBF}	l_{Noise}
Hardness	1.88	1.76	0.05
Low alloy hardness	2.02	3.13	0.05
Izod	1.59	1.89	0.05

Fig. 7. In this case, the effect of α on the average CV for each kernel was found to be negligible, while the spread in the CV has slightly increased with increasing α.

Based on the results shown in Fig. 7, RBF() + WhiteKernel() is selected; this three parameter model provides the best compromise between minimal CV score and minimal model complexity. The optimized hyperparameters are shown in Table 4 for the hardness and Izod predictions. As shown in Table 4, the fundamental kernel prefactors (i.e., θ_1 and l_{Noise}) remain relatively constant over each of the four sets of target variables studied. Moreover, $\theta_1 >> l_{\mathrm{Noise}}$, suggesting that the WhiteKernel() does not contribute as significantly to the predictions.

Results of GPR

Using the optimized kernels of Table 4, the hardness and Izod impact toughness are predicted based on the composition and temper inputs. The predicted hardness and predicted Izod results for the test sets are shown in Fig. 8. The predicted hardness of the test sets has an R^2 value of 0.898 ± 0.023 and the predicted Izod test sets an R^2 of 0.919 ± 0.01. Both are significant improvements over the linear regressions trained on the same data and with the same input features (R^2 of 0.796 and 0.843, respectively). Further, the percent of predictions that are off by greater than ± 50 HV are reduced by the ML model to 10.9% down from 38.4% for the hardness prediction and those off by greater than ± 10 J are 31.2% down from 46.3% for the Izod prediction. In addition, the MAE is 25.7 HV down from 49.5 HV and 8.05 J down from 12.3 J for Izod. The RMSE for the Izod prediction is 11.1 J, which compares favorably to another ML model study (GA-NN and others) for low-alloy steels using Charpy impact data which had testing RMSE of between ~ 18 and 20 J [36]. We also train a GPR using only the low alloy subset of the hardness data; the R^2 for this prediction using just the low alloy subset is 0.939 ± 0.01, better than both the full dataset prediction and the empirical predictions using just low alloys. Further, only 8.1% of the predictions are greater than 50HV from the expected value and the MAE is 19.2HV.

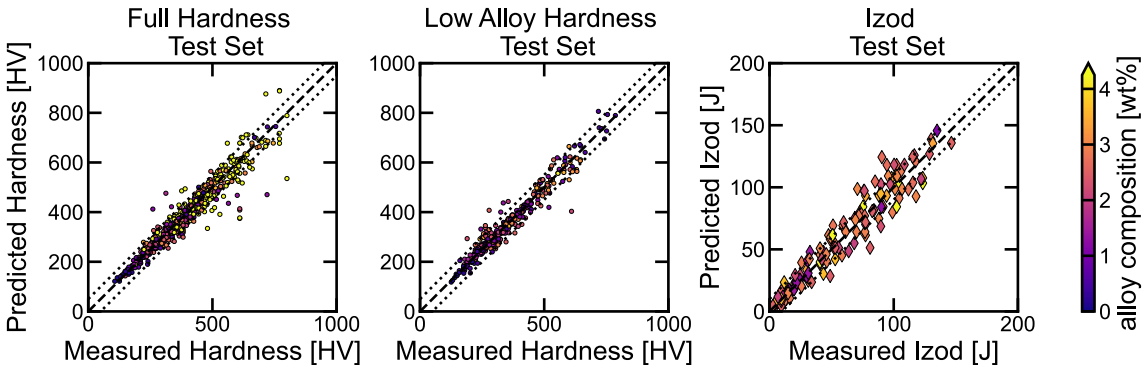

Fig. 8 Predicted versus measured hardness with full dataset hardness with low alloy subset Izod for an example test set using optimized kernels from Table 4

Feature Analysis

To describe some of the differences between the linear regression and GPR and interpret the trends in properties, we quantify the significance of the various input features. Feature importance is assessed using Shapley values calculated using the SHAP (SHapley Additive exPlanations) module in python [37]. Shapley values are representative of the significance of each feature on the predicted property and can be calculated for both the linear regression and GPR. The SHAP values plotted in Fig. 9 show the increase or decrease to a baseline property prediction value where each point represents the value of the feature for a given alloy and temper input. For the GPR kernel and test set presented in Fig. 9, the baseline hardness is 376 HV for example, and a point at a SHAP value of + 50 for Cr would indicate the input value of Cr for that alloy added 50HV to the baseline hardness. General trends can be observed, such as higher contents of most elements tending to increase the hardness; however, it is clear that these are not straightforward relationships between composition and hardness increase from the variation in colors representing the input feature value. Ni and V, for example, have some high compositions that decrease the hardness.

The input features are ordered by importance, which is calculated by mean(|SHAP|). Different kernels or test train partitions on the same data will end up with different rankings of feature importance. Here, since we have optimized the kernel, we quantify the feature importance variability by

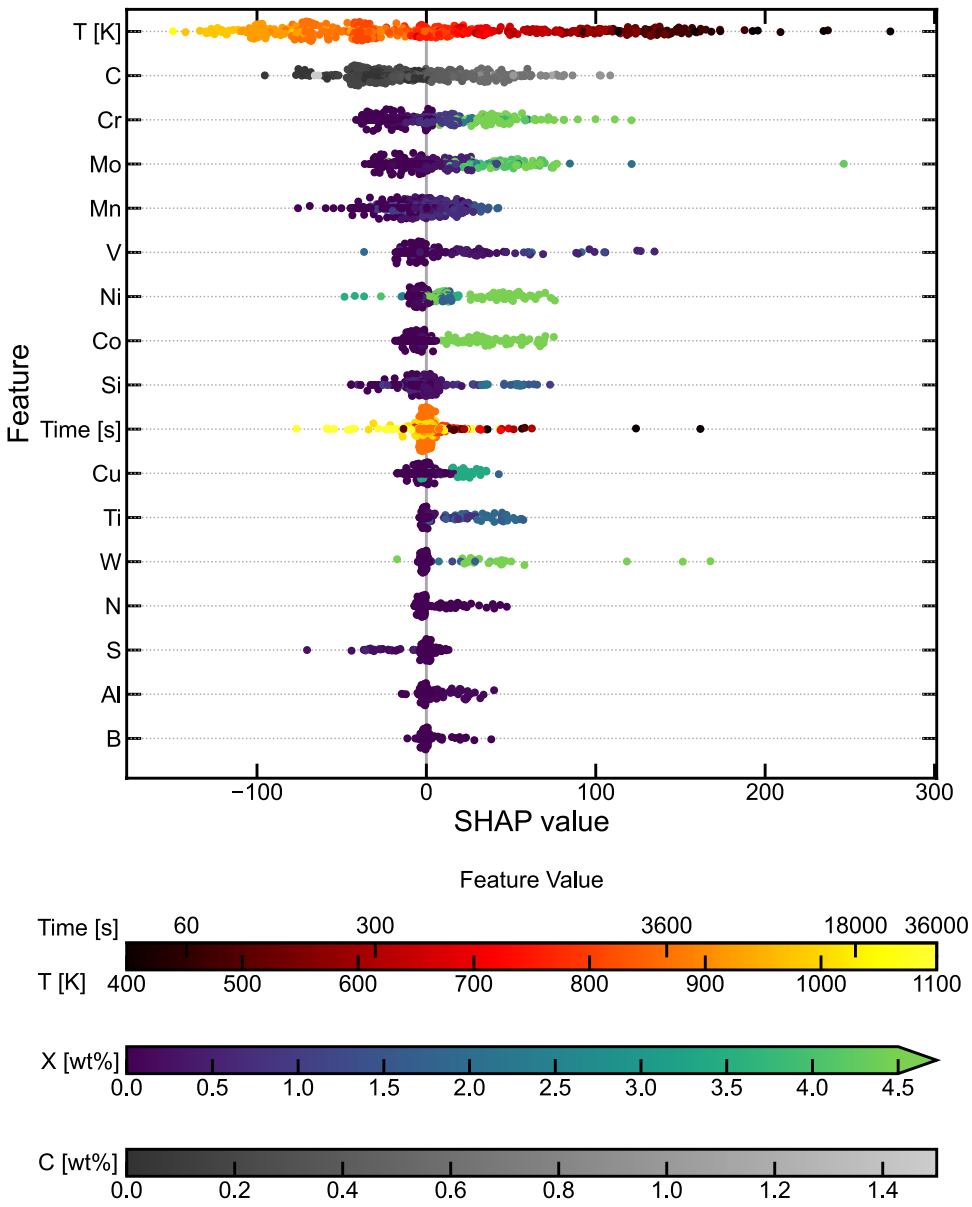

Fig. 9 Shapley plot for hardness prediction from GPR using the optimized kernel. Points are colored by the input feature value

varying the test and train split. Ten different random seeds for a 67%/33% test/train split are used with the optimized kernel, and the average Shapley values tabulated in Fig. 10. These results are compared to the mean(|SHAP|) values for the linear regression equations. While there is agreement for the relative importance of the first three input features for both predictions, the ranking of elements fluctuates thereafter. Several of the elements have essentially no importance according to the linear regression, but relatively high importance in the GPR, e.g., Cr. Details will be discussed in the following section.

Hardness

The temper temperature is by far the most important feature, regardless of test and train split. Carbon is unsurprisingly the most important of the alloying elements for predicting both hardness and toughness. For hardness, the next most important elements are Mo and Cr, with almost equal impact. Interestingly, the linear regression equation found Cr to have essentially no impact on the hardness. We illustrate the differences between the feature importance of the GPR and the empirical equation by directly comparing the SHAP values in Fig. 11. The SHAP values for the linear regression, Eq. (5), are indicated by the teal lines, and compared to the points representing the SHAP values for the GPR (points colored by temper temperature). The SHAP values for the GPR for Cr show approximately a parabolic trend, where the importance of Cr increases up to around 5 wt%, and then decreases at higher compositions. Further, at the leanest compositions, e.g., < 1 wt% Cr, Cr tends to decrease the hardness from baseline, whereas at all other compositions it provides some amount of increase. Another element where the GPR and linear regression particularly diverge is Co. Empirically, Co has not been found to have a large positive or negative effect on the predicted hardness of the steel, but it can be seen here to have a strong positive effect in certain high composition regions. There are many high Co steels that are typically tempered at or above 673 K as Co has been noted to reduce the activity of C [64] and promote formation of refractory carbides; this appears as an increased hardness at high temperatures.

For Co and Cr, it is clear that the linear regression cannot capture the complex relationship between the element and its effect on hardness. In many other cases, the overall trend between the two models is in general agreement, e.g., an increase in V content increases the hardness, but the details can differ significantly. However, in the case

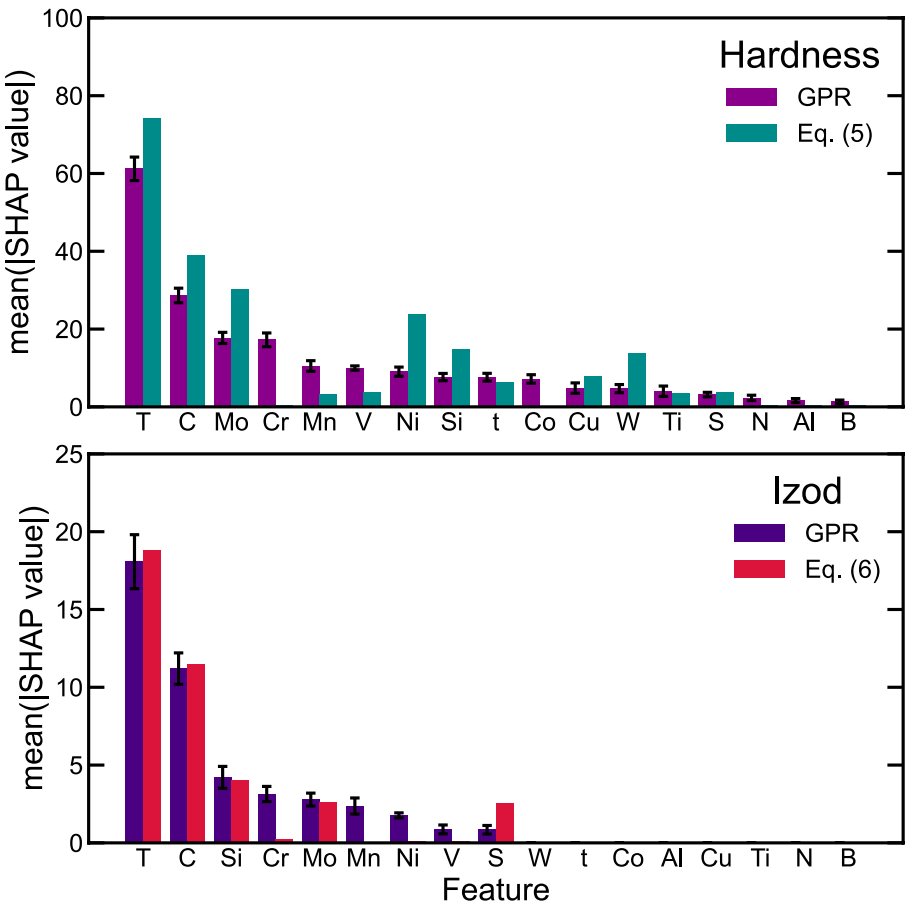

Fig. 10 Mean SHAP values for the hardness and Izod predictions

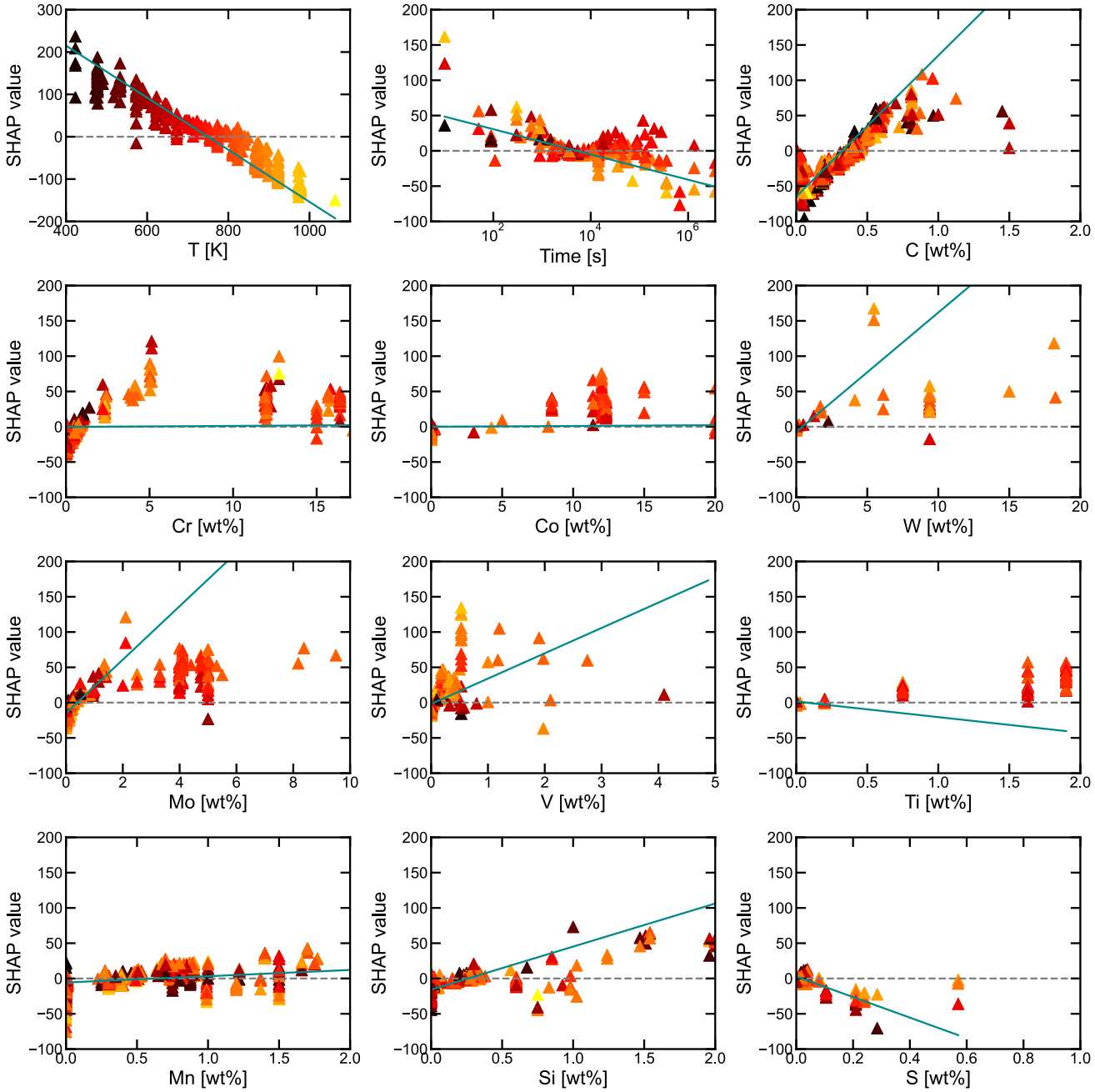

Fig. 11 Some SHAP value comparisons for the elements, colored by temper temperature

of Ti, the linear regression predicts that increasing Ti will decrease the hardness, whereas the majority of the GPR SHAP values show an increase. There are several elements for which the average SHAP value is low, but many individual SHAP values are quite high, in particular for W and Co. This is likely due to a large number of zero composition points. W has some very high importance values at high temperature tempers, which is likely correlated to its implementation in W-bearing tool steels for machining operations as well as for high temperature tool steels in which hardness at high operating temperatures are required.

Mo is considered a strong strengthener by both our linear regression and GPR models, with a positive impact on the hardness for a large range of composition. Mo and V are both known to provide solid solution strengthening during low temperature tempering ($\leq$473 K) and a secondary hardening affect from carbide formation (e.g., M_2C, MC) at higher temperature tempers ($>$773 K) [27, 59]. For low concentrations of Mo, i.e., where there is likely enough Mo

present to contribute to solid solution strengthening but not necessarily enough to form carbides, a temperature effect can be seen on the SHAP values, shown in Fig. 12 where the higher temperature points have a negative influence on the hardness and the lower temperature points are neutral to increasing hardness. At concentrations greater than ~ 0.25 wt%, Mo is unambiguously providing an increase to the hardness. Similarly, V generally provides an increase to the hardness other than at very low compositions and the highest compositions. It is interesting to note that the Mukherjee empirical model had negative coefficients for both these elements, in opposite relationship to both the linear regression and GPR models here. The Mukherjee model also performed poorly for low temperature predictions, which likely precludes describing the solid solution strengthening effect of these elements; further, they may have not had enough points at higher compositions of Mo and V (while their maximum values were 4 and 2.2 for Mo and V, respectively, their average values were only 0.15 and 0.04 wt%). Interestingly, for a GPR trained only on the low alloy dataset, the temperature trend does not appear (Fig. 12).

The influence of S is generally negative, which is not surprising due to its affinity to bond to effective strengthening agents at high temperatures, e.g., formation of MnS consuming Mn that could be used elsewhere. Mn itself has a variety of significant SHAP values making it one of the more important features but has no consistent positive or negative correlation with temperature or composition. A distinctive feature of the SHAP values for Mn is that there is a large spread at 0 wt%; other than at the lowest temperatures, the absence of Mn from the alloy is generally considered a negative impact on the hardness. Mn is present in most commercial steels to some degree, with many purposes, including: a desulfurizing agent, for its ability to promote cementite formation, and as a lower cost hardenability agent to replace Ni. Si is interesting, in that, like Mn there is a lot of individual variability, but no consistent positive or negative impact. Si is another intrinsic alloying element included in most steels and is used as a secondary deoxidizer. At high alloying concentrations Si has been shown to inhibit the formation of cementite at lath boundaries and reduce its embrittling effect.

Ni and Cu effects are not pronounced, which could be due to the lack of Ni or Cu containing carbides. Alloys containing high Cu are typically precipitation strengthened through the formation of Cu particles on the martensitic matrix. In maraging steels with high Ni concentrations, the formation of intermetallic Ni_3Ti particles is used as a strengthening phase rather than a carbide. These high Ni or Cu steels that are non-carbide precipitation strengthened also tend to have very low C content, as can be seen in Fig. 13, with the SHAP value data points colored by C content rather than temper temperature.

A full SHAP plot for the low alloy dataset can be found in online supplementary Figure S-4. In terms of feature ranking (online supplementary Fig. S-5 and Fig. 15), the temperature and carbon are still the most significant, but the importance of time has increased considerably relative to other input features; the importance of several elements has also decreased considerably (not including W and Co which are not in any alloys of the low alloy steel subset). Mo is far less important, for example, as the low alloy data subset does not include the composition range where it was a consistently positive contributor to the hardness. An ANN model with a dataset consisting of low alloy steels found Cr to be the next most important element after C in predicting hardness [13]. In fact, the importance of Cr increased in the low alloy subset over the full hardness dataset in our model as well, shown in Fig. 10. Similarly, they found Si, Mn, and Mo to have about the same impact, and Ni to have almost no impact; however, they also limited their alloys to only these six elements. It can be seen from Fig. 13 that most of the influence of Ni derives from the low-carbon high alloys. Overall, the SHAP values for the GPR trained on the low alloy dataset, shown in Fig. 14, exhibit simpler trends, which make them more likely to align with that of the linear regression. This is also reflected in the improvement in the statistics for the low alloy dataset prediction from the linear regression model in Table 2.

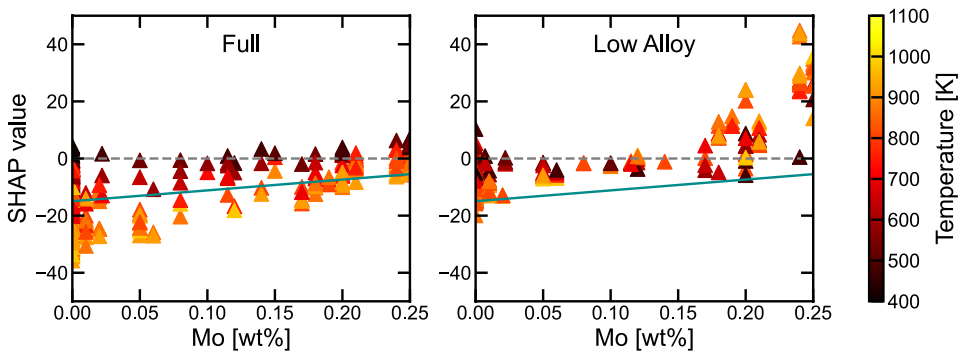

Fig. 12 Comparison of the SHAP values at low compositions of Mo for the GPR trained on the full dataset vs only the low alloy subset to predict hardness

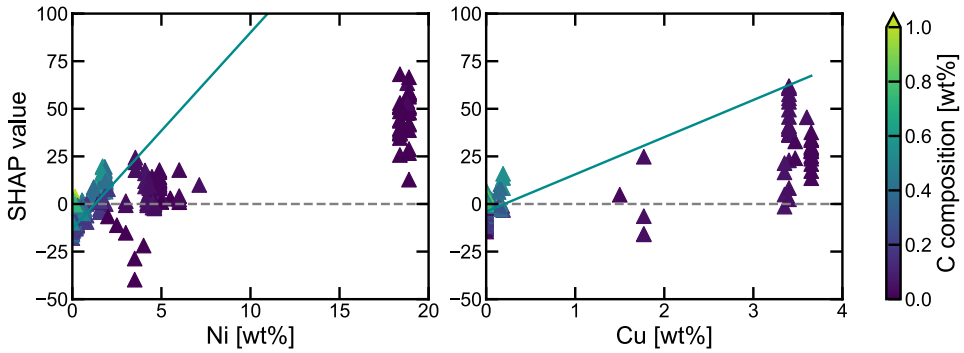

Fig. 13 SHAP values as a function of Ni and Cu content colored by carbon content for the GPR (points) and linear regression (line) trained on the full dataset to predict hardness

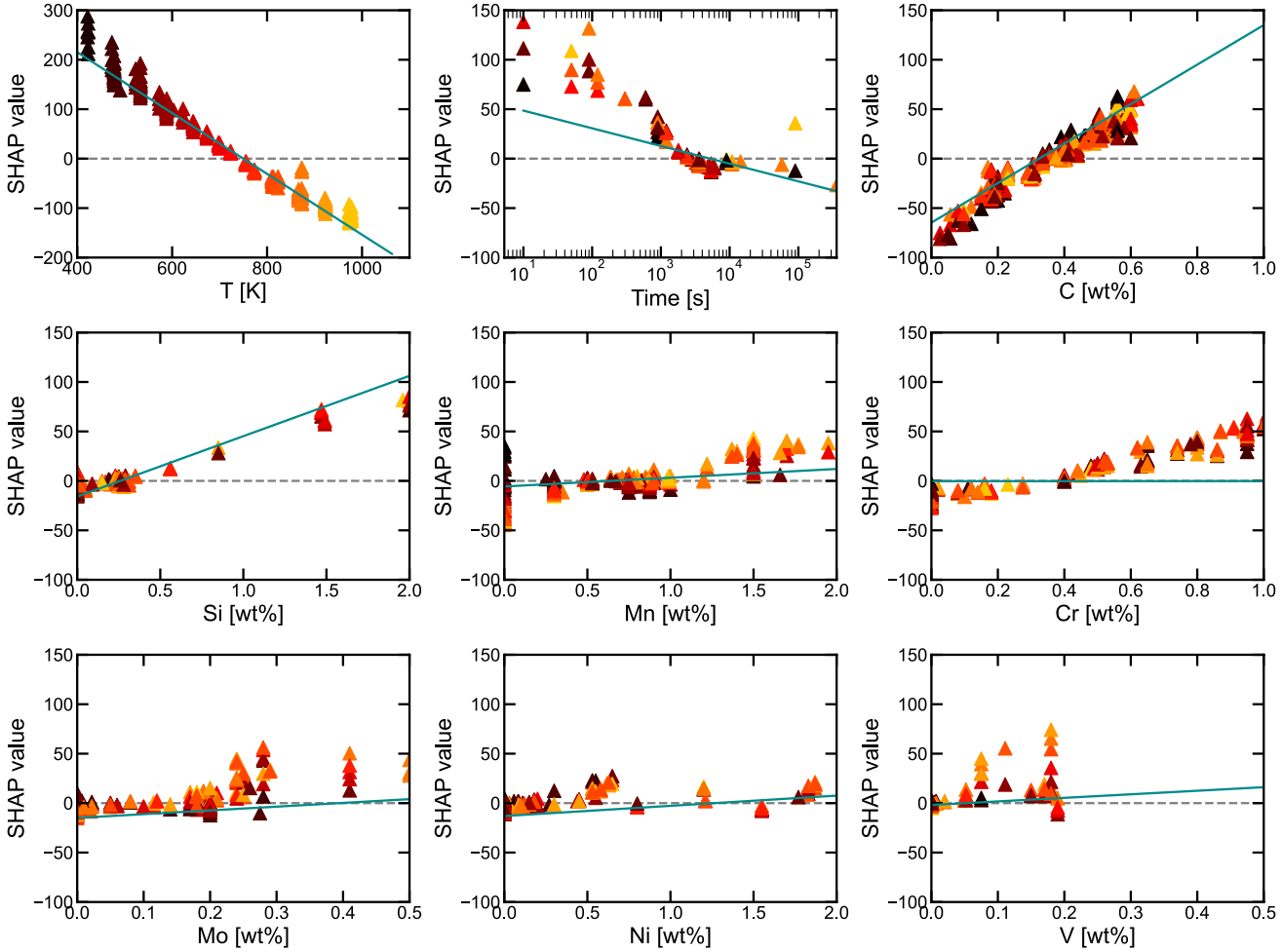

Fig. 14 Some SHAP values for the GPR trained on the low alloy dataset only. Points are colored by tempering temperature. Lines are the SHAP values for the linear regression

Izod Predictions

As expected, temperature and carbon are still the most important features for the prediction of Izod impact toughness (Fig. 15). They also have the anticipated inverse relationship, i.e., increasing carbon content increases hardness but decreases toughness, as can be seen in a comparison of the SHAP values. This general trend has been well established experimentally where prioritizing strength (by either low temperature tempering or increasing carbon) comes at the cost of toughness. It is also captured by the empirical equation, with fairly good alignment between the SHAP

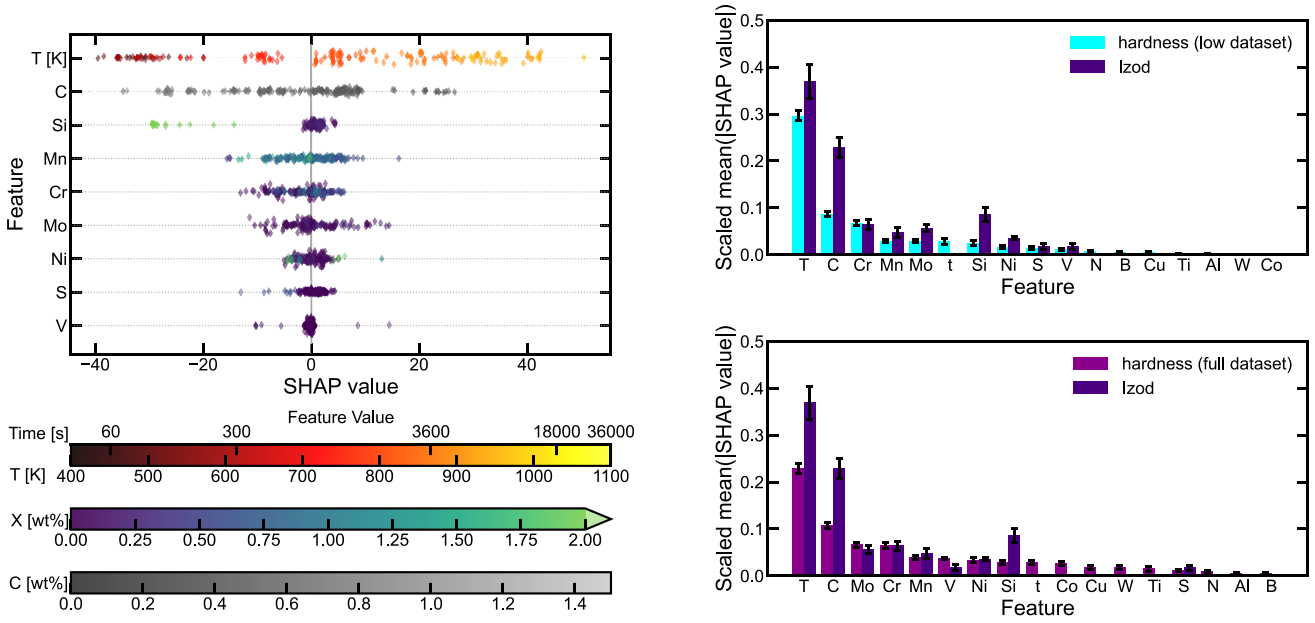

Fig. 15 SHAP values for Izod prediction, and feature importance compared to hardness predictions. Feature importance is scaled by max SHAP value for ease of comparison

values of the GPR and linear regression for the temperature and carbon content (Fig. 16). The Izod dataset is primarily comprised of low alloy steels (88% low alloy, compared to 61% for the hardness dataset), so the feature importance is compared to the low alloy subset as well as the full dataset (Fig. 15). The mean SHAP values are scaled by the highest SHAP value in the prediction (both hardness and Izod) to enable easier comparison between the hardness and Izod predictions.

Cementite formation and morphology are understood to be the leading cause of tempered martensite embrittlement (TME) [26], which is most commonly expressed as low impact toughness, e.g., low Izod values. Therefore, the SHAP values for the Izod predictions in Fig. 16 are colored by carbon content (a strong correlative factor to cementite formation) to look for cross-correlations. Aside from carbon, the empirical equation to predict Izod impact toughness, Eq. (6), had very small coefficients for most of the elements, as is reflected in the SHAP values for the empirical equation in Fig. 16 (red lines). Silicon is known to be effective at altering the activity for carbon and retarding cementite formation during tempering; however, other than a deleterious effect at high concentrations, there is actually very little influence on the Izod. The few high Si content points skew the average importance of this element, when comparing the average SHAP values.

Cr, Mn, and Mo are the next most important features, but have no evident trend as a function of composition. Manganese is typically noted as having a complicated effect on the toughness of steels. At low carbon levels ($\leq$0.3 wt% C) the addition of Mn can be beneficial to the toughness of the steel, whereas for higher carbon steels (>0.5 wt% C) it was found that increasing the Mn concentration produced a drop in toughness [65]. This dichotomy can be observed in the SHAP values for the large cluster of Mn concentrations less than 1 wt% in Fig. 16 where low values of C (purple to blue points) have a slight positive influence on the Izod whereas higher C values (blue to green points) have a slight negative influence, and the highest carbon content having the most negative influence. In a GA-NN model predicting Charpy impact toughness [36], there was a clear trend of decreasing impact energy with increasing carbon content, as also seen with our GPR model. They also observed an increase in impact toughness with an increase in Cr content with some confidence between the Cr compositions of ~ 1–2 wt%; however, below 1 wt%, the range present in our dataset, their confidence interval was too large to confirm any trend.

Trends in the influence of composition on Izod are far more difficult to discern that for hardness, despite the comparatively constrained compositional ranges. The elements all have varied positive and negative SHAP values and very little cross-correlation with carbon content, as might have been expected from known TME mechanisms. Historically hardness has been predicted with some degree of accuracy in low alloy steels; however, a unified model incorporating both low and high alloy steels have not been successful using standard empirical relationships. This is seen in the complicated effect of the input features via the SHAP values. Toughness has eluded even limited empirical relationships in literature because of the complicated nature of

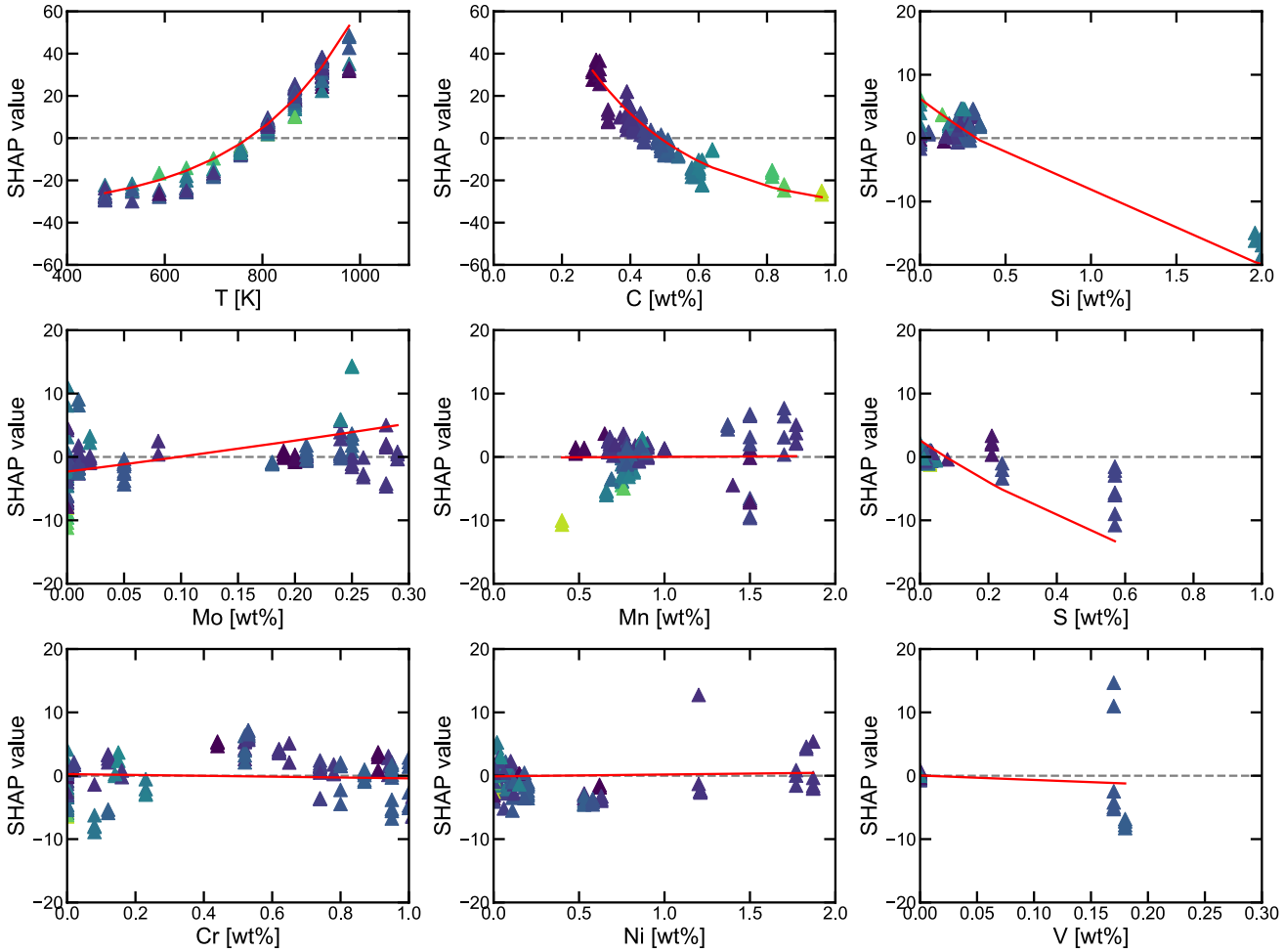

Fig. 16 SHAP values for Izod prediction from GPR (points) and linear regression (line). Points are colored by carbon content

steel fracture; while there exists qualitative assessments of individual elements on toughness an overarching model has not yet been identified. From these modeling frameworks we have shown that the effect of processing (tempering time and temperature) as well as composition have complicated nonlinear effects on the resultant toughness.

Conclusion

We have presented a framework for Bayesian Gaussian process regression modeling to predict the hardness and Izod impact toughness of low and high alloys quenched and tempered steels given alloy composition and processing history. This approach more accurately captures complex alloying behavior far beyond the range of previous regression models. Shapley feature analysis on the input parameters for both GPR and linear regression showed that there are complicated interactions between the elements and the target properties, particularly for the Izod impact toughness. The most important input features for both target properties are tempering temperature and carbon content, as expected. The dataset gathered here also provides a starting point for future, more complex, ML models and analysis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40192-023-00311-9.

Acknowledgements A. Garza was supported by the Department of Defense (DOD) Historically Black Colleges and Universities and Minority Serving Institutions (HBCU/MI) summer research program sponsored by the Office of the Under Secretary of Defense for Research and Engineering/Research, Technologies, and Laboratories (OUSD(R&E)/RT&L). L.D. McClenny was supported by funding through the Oak Ridge Associated Universities contract #W911NF-16-2-0008. H. Murdoch would like to thank Dr. D. Magagnosc for discussions on figure presentation.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Andrews K (1965) Empirical formulae for the calculation of some transformation temperatures. J Iron Steel Inst 721–727
2. Kim H, Inoue J, Okada M, Nagata K (2017) Prediction of Ac3 and martensite start temperatures by a data-driven model selection approach. ISIJ Int 57(12):2229–2236
3. Ingber J, Kunert M (2022) Prediction of the martensite start temperature in high-carbon steels. Steel Res Int 93(5):2100576
4. Gramlich A, van der Linde C, Ackermann M, Bleck W (2020) Effect of molybdenum, aluminium and boron on the phase transformation in 4wt% manganese steels. Results Mater 8:100147. https://doi.org/10.1016/j.rinma.2020.100147
5. Park J, Shim J-H, Lee S-J (2018) New equation for prediction of martensite start temperature in high carbon ferrous alloys. Metall and Mater Trans A 49(2):450–454. https://doi.org/10.1007/s11661-017-4436-8
6. Kasatkin O, Vinokur B, Pilyushenko V (1984) Calculation models for determining the critical points of steel. Met Sci Heat Treat 26(1):27–31
7. Kang S, Lee S-J (2014) Prediction of tempered martensite hardness incorporating the composition-dependent tempering parameter in low alloy steels. Mater Trans 55(7):1069–1072
8. International A (2003) A255: standard test methods for determining hardenability of steel. ASTM International, West Conshohocken
9. Lipton ZC (2018) The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57
10. Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin
11. Vapnik V (1999) The nature of statistical learning theory. Springer Science and Business Media, Berlin
12. Eres-Castellanos A, De-Castro D, Capdevila C, Garcia-Mateo C, Caballero FG (2021) Assessing the implementation of machine learning models for thermal treatments design. Mater Sci Technol 37(16):1302–1310. https://doi.org/10.1080/02670836.2021.2001731
13. Jeon J, Seo N, Son SB, Lee S-J, Jung M (2021) Application of machine learning algorithms and SHAP for prediction and feature analysis of tempered martensite hardness in low-alloy steels. Metals 11(8):1159
14. Tenner J (2000) Optimisation of the heat treatment of steel using neural networks. University of Sheffield
15. Taghizadeh S, Safarian A, Jalali S, Salimiasl A (2013) Developing a model for hardness prediction in water-quenched and tempered AISI 1045 steel through an artificial neural network. Mater Des 51:530–535. https://doi.org/10.1016/j.matdes.2013.04.038
16. Mukherjee T, DebRoy T, Lienert TJ, Maloy SA, Lear CR, Hosemann P (2022) Tempering kinetics during multilayer laser additive manufacturing of a ferritic steel. J Manuf Process 83:105–115. https://doi.org/10.1016/j.jmapro.2022.08.061
17. Verma AK, Hawk JA, Bruckman LS, French RH, Romanov V, Carter JLW (2019) Mapping multivariate influence of alloying elements on creep behavior for design of new martensitic steels. Metall Mater Trans A 50(7):3106–3120. https://doi.org/10.1007/s11661-019-05234-9
18. Peng J, Yamamoto Y, Hawk JA, Lara-Curzio E, Shin D (2020) Coupling physics in machine learning to predict properties of high-temperatures alloys. npj Comput Mater 6(1):141. https://doi.org/10.1038/s41524-020-00407-2
19. Mamun O, Wenzlick M, Hawk J, Devanathan R (2021) A machine learning aided interpretable model for rupture strength prediction in Fe-based martensitic and austenitic alloys. Sci Rep 11(1):5466. https://doi.org/10.1038/s41598-021-83694-z
20. Dimitriu RC, Bhadeshia HKDH, Fillon C, Poloni C (2008) Strength of ferritic steels: neural networks and genetic programming. Mater Manuf Process 24(1):10–15. https://doi.org/10.1080/10426910802539796
21. Shen C, Wang C, Wei X, Li Y, van der Zwaag S, Xu W (2019) Physical metallurgy-guided machine learning and artificial intelligent design of ultrahigh-strength stainless steel. Acta Mater 179:201–214. https://doi.org/10.1016/j.actamat.2019.08.033
22. Jiang X, Jia B, Zhang G, Zhang C, Wang X, Zhang R et al (2020) A strategy combining machine learning and multiscale calculation to predict tensile strength for pearlitic steel wires with industrial data. Scripta Mater 186:272–277. https://doi.org/10.1016/j.scriptamat.2020.03.064
23. Xie Q, Suvarna M, Li J, Zhu X, Cai J, Wang X (2021) Online prediction of mechanical properties of hot rolled steel plate using machine learning. Mater Des 197:109201. https://doi.org/10.1016/j.matdes.2020.109201
24. Mukherjee M, Dutta C, Haldar A (2012) Prediction of hardness of the tempered martensitic rim of TMT rebars. Mater Sci Eng A 543:35–43
25. Athavale VA (2019) Development of stage-I tempered high strength cast steel for ground engaging tools. Missouri University of Science and Technology
26. De Cooman B, Speer J (2011) Austenite decomposition in Fe-CX alloy systems. Fundam Steel Product Phys Metall 173
27. Leslie WC (1981) The physical metallurgy of steels. Hemisphere Publishing Corp, New York, p 396
28. Clarke AJ, Klemm-Toole J, Clarke KD, Coughlin DR, Pierce DT, Euser VK et al (2020) Perspectives on quenching and tempering 4340 steel. Metall Mater Trans A 51(10):4984–5005. https://doi.org/10.1007/s11661-020-05972-1
29. Lankford WT (1985) The making, shaping, and treating of steel. Association of Iron and Steel Engineers
30. Dunne D, Tsuei H, Sterjovski Z (2004) Artificial neural networks for modelling of the impact toughness of steel. ISIJ Int 44(9):1599–1607
31. Wang C, Shen C, Huo X, Zhang C, Xu W (2020) Design of comprehensive mechanical properties by machine learning and high-throughput optimization algorithm in RAFM steels. Nucl Eng Technol 52(5):1008–1012. https://doi.org/10.1016/j.net.2019.10.014
32. Chen M-Y, Da L (2006) Impact toughness prediction for TMCP steels using knowledge-based neural-fuzzy modelling. ISIJ Int 46(4):586–590
33. Guo Z, Sha W (2004) Modelling the correlation between processing parameters and properties of maraging steels using artificial neural network. Comput Mater Sci 29(1):12–28. https://doi.org/10.1016/S0927-0256(03)00092-2
34. Chen Y, Wang S, Xiong J, Wu G, Gao J, Wu Y et al (2023) Identifying facile material descriptors for Charpy impact toughness in low-alloy steel via machine learning. J Mater Sci Technol 132:213–222. https://doi.org/10.1016/j.jmst.2022.05.051
35. Diao Y, Yan L, Gao K (2022) A strategy assisted machine learning to process multi-objective optimization for improving mechanical properties of carbon steels. J Mater Sci Technol 109:86–93. https://doi.org/10.1016/j.jmst.2021.09.004
36. Yang YY, Mahfouf M, Panoutsos G (2011) Development of a parsimonious GA–NN ensemble model with a case study for Charpy impact energy prediction. Adv Eng Softw 42(7):435–443. https://doi.org/10.1016/j.advengsoft.2011.03.012
37. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30

38. Dhua S, Ray A, Sarma D (2001) Effect of tempering temperatures on the mechanical properties and microstructures of HSLA-100 type copper-bearing steels. Mater Sci Eng A 318(1–2):197–210
39. Chi Y-C, Lee S, Cho K, Duffy J (1989) The effects of tempering and test temperature on the dynamic fracture initiation behavior of an AISI 4340 VAR steel. Mater Sci Eng A 114:105–126
40. Zhang Z, Delagnes D, Bernhart G (2004) Microstructure evolution of hot-work tool steels during tempering and definition of a kinetic law based on hardness measurements. Mater Sci Eng, A 380(1):222–230. https://doi.org/10.1016/j.msea.2004.03.067
41. Speich G (1969) Tempering of low-carbon martensite. Trans Met Soc AIME 245(12):2553–2564
42. Kwon H, Lee K, Yang H, Lee J, Kim Y (1997) Secondary hardening and fracture behavior in alloy steels containing Mo, W, and Cr. Metall Mater Trans A 28:775–784
43. Gojic M, Kosec L, Matkovic P (1998) The effect of tempering temperature on mechanical properties and microstructure of low alloy Cr and CrMo steel. J Mater Sci 33:395–403
44. Mesquita RA, Kestenbach H-J (2012) Influence of silicon on secondary hardening of 5wt% Cr steels. Mater Sci Eng A 556:970–973. https://doi.org/10.1016/j.msea.2012.06.060
45. Nam WJ, Lee CS, Ban DY (2000) Effects of alloy additions and tempering temperature on the sag resistance of Si–Cr spring steels. Mater Sci Eng A 289(1):8–17. https://doi.org/10.1016/S0921-5093(00)00928-X
46. Euser VK, Williamson DL, Clarke AJ, Speer JG (2022) Cementite precipitation in conventionally and rapidly tempered 4340 steel. JOM 74(6):2386–2394. https://doi.org/10.1007/s11837-022-05285-1
47. Ohmura T, Hara T, Tsuzaki K (2003) Evaluation of temper softening behavior of Fe–C binary martensitic steels by nanoindentation. Scripta Mater 49(12):1157–1162. https://doi.org/10.1016/j.scriptamat.2003.08.025
48. Tanino M, Nishida T (1968) On the secondary hardening on tempering in vanadium steels. Trans Jpn Inst Metals 9(2):103–110
49. Nishimura T (1967) On the tempering behaviour of 3Cr-W, 3Cr-W-Co and 12Cr-W-Co type tool steels for hot work. Tetsu-to-Hagane 53(2):116–130
50. Jh H (1945) Time-temperature relations in tempering steel. Trans AIM 162:223–249
51. Goodall AL (2020) Effect of initial microstructural conditions and tempering parameters on the carbide characteristics and hardness of alloyed quenched and tempered steel. University of Birmingham
52. Bethlehem Steel (1952) Modern steels and their properties. Bethlehem Steel Company, Bethlehem, PA
53. Grange R, Hribal C, Porter L (1977) Hardness of tempered martensite in carbon and low-alloy steels. Metall Trans A 8:1775–1785
54. GRANTA Materials Universe (2020) In: ANSYS. GRANTA: ANSYS
55. ASM International (1990) Metals handbook, 10th edn
56. Murdoch H, Field D, Szajewski B, McClenny L, Garza A, Rinderspacher B et al (2023) Hardness temp martensitic steels. https://doi.org/10.13011/m3-0pcb-2x35
57. Wang M, Wang Y, Sun F (2006) Tempering behavior of a semi-high speed steel containing nitrogen. Mater Sci Eng, A 438–440:1139–1142. https://doi.org/10.1016/j.msea.2006.02.202
58. International A (2012) E140–12b: standard hardness conversion tables for metals relationship among brinell hardness, vickers hardness, rockwell hardness, superficial hardness, knoop hardness, scleroscope hardness, and leeb hardness. ASTM International, West Conshohocken
59. Krauss G (2015) Steels: processing, structure, and performance. ASM International, Materials Park
60. Crafts W, Lamont J (1947) Effect of alloys in steel on resistance to tempering. Trans AIME 172:222–243
61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
62. ASTM (2022) E18-22: Standard test methods for Rockwell hardness of metallic materials. ASTM International, West Conshohocken
63. ASTM (2016) E23–16b: Standard test methods for notched bar impact testing on metallic materials. ASTM International, West Conshohocken
64. Speich G, Dabkowski D, Porter L (1973) Strength and toughness of Fe-10Ni alloys containing C, Cr, Mo, and Co. Metall Trans 4:303–315
65. Davis JR, Mills K, Lampman S (1990) ASM handbook. In: Properties and selection: irons, steels, and high-performance alloys, Vol 1. ASM International, Metals Park