Machine Learning-based estimation and explainable artificial intelligence-supported interpretation of the critical temperature from magnetic *ab initio* Heusler alloys data

Robin Hilgers [1,2,*] Daniel Wortmann [1] and Stefan Blügel [1,2]

[1]*Peter Grünberg Institute and Institute for Advanced Simulation, Forschungszentrum Jülich and JARA, 52425 Jülich, Germany*

[2]*Department of Physics, RWTH Aachen University, Aachen, Germany*

(Received 3 January 2025; revised 9 April 2025; accepted 14 April 2025; published 29 April 2025)

Machine learning (ML) has impacted numerous areas of materials science, most prominently improving molecular simulations, where force fields were trained on previously relaxed structures. One natural next step is to predict material properties beyond structure. In this work, we investigate the applicability and explainability of ML methods in the use case of estimating the critical temperature (T_c) for magnetic Heusler alloys calculated using *ab initio* methods determined materials-specific magnetic interactions and a subsequent Monte Carlo (MC) approach. We compare the performance of regression and classification models to predict the range of the T_c of given compounds without performing the MC calculations. Since the MC calculation requires computational resources in the same order of magnitude as the density functional theory (DFT) calculation, it would be advantageous to replace either step with a less computationally intensive method such as ML. We discuss the necessity to generate the magnetic *ab initio* results to make a quantitative prediction of the T_c. We used state-of-the-art explainable artificial intelligence (XAI) methods to extract physical relations and deepen our understanding of patterns learned by our models from the examined data.

DOI: 10.1103/PhysRevMaterials.9.044412

I. INTRODUCTION

Machine learning (ML) modeling has been shown to yield promising results in various scientific sectors and applications [1–3]. The ability of flexible learning algorithms to recognize patterns, adapt to data properties, and tackle challenges such as regression, classification, and clustering has established an additional scientific paradigm of data-driven science besides the traditional paradigms of experiments, theories, and simulations. Data-driven science essentially shifts scientific problem-solution strategies for predictions from problem-specific models to versatile data-based models [4–6]. This is also the case for a plurality of materials science applications including superconductivity [7], molecular dynamics [8], materials synthesis and design [9], knowledge discovery through data mining [10], entropy changes [11], and other topics for both properties and materials prediction [5,12,13]. For some of the mentioned applications, e.g., in some molecular dynamics simulation applications [8], lightweight and computationally inexpensive ML-based approaches were able to virtually replace established techniques, while in other applications ML-based approaches complement existing methodologies [5]. Data mining-related techniques have shown to be powerful tools in the hands of scientists to discover relations within data, even in the materials science community [10].

There are a multitude of magnetic properties to investigate, many of which are traditionally described by complex models based in part on the quantum mechanics of the many-electron problem. Within the set of magnetic properties, the critical temperature, also known as the Curie temperature in the context of ferromagnetic materials, represents a key characteristic in both fundamental physics and practical applications. It provides valuable insights into the transitions between different magnetic phases and guides the design and optimization of magnetic materials for technological use. For example, in the design of magnetic materials for the energy use sector of the economy [14], e.g., electric power generation, conditioning, conversion, transportation, or the information sector of the economy, e.g., spintronics [15] or magnetic storage devices (like hard drives), the critical temperature determines the maximum operating temperature where magnetic data storage remains stable. Typical application demands necessitate critical temperature values significantly exceeding room temperature [16]. Hence, in order to conduct application-oriented material screening studies at a high-throughput scale for materials discovery, a lightweight method is required to predict whether the critical temperature of a compound meets the requirements set by the applications. The ML-based prediction of magnetic materials and their corresponding properties is mostly covered by applications dedicated to certain material families, e.g., double perovskites [17], metallic glasses [18], multilayer film systems [19,20], and of course Heusler alloys [21]. Approaches in the field of high-throughput computational materials science currently aim to optimize computational starting parameters using computed data, essentially performing data-driven process optimization in materials research. [19] Existing works focusing on predicting magnetic properties of Heusler alloys, mostly try

*Contact author: robin.hilgers@rwth-aachen.de

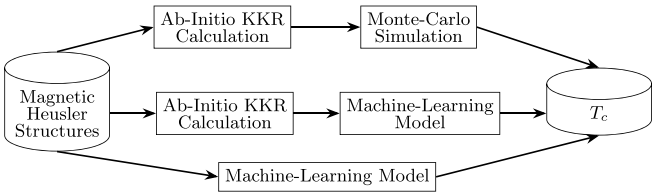

FIG. 1. Schematic depiction of the layered T_c determination with different ML integration levels.

to model the Curie temperature in ferromagnetic materials [22,23], while the more general concept describing a wide range of magnetic phases, including ferromagnetic, antiferromagnetic, ferrimagnetic, and spin spiral-type ordering is the critical temperature of the phase change transition of the ordered magnetic to a non-magnetic state represents the field of interest in this study. Other studies predicting magnetic properties in Heusler alloys predict, e.g., magnetic moments [21] and the magnetocaloric effect [24,25].

Within the phase space of magnetic materials, the Heusler (and Heusler-like alloys) alloys are known to represent candidate materials for various technical applications, as the material class of Heusler [26,27] alloys (e.g., the ordered $L2_1$ phase) and related disordered phases (such as A2 and B2 phases) are known to exhibit many interesting properties including superconductivity [28], piezoelectricity [29], rare-earth free permanent magnets [30], and half-metallicity [31]. The combination of multiple properties in a single compound such as both half-metallicity and magnetic stability allow for the occurrence of spin-polarized charge currents, which are a topic that is actively investigated by the scientific community for applications in spintronics [32,33]. By including not only the ordered but also disordered phases and quaternary Heusler alloys, the phase space of possible compounds increases drastically in comparison to existing works like [34], which restricts the phase space to pure transition-metal Heusler alloys. However, as a Heusler alloy's structure is defined by the individual compound's lattice site constituents, the lattice constant, and the symmetry group alone, the structural parameters that have to be considered by a model in order to describe such a system are very limited.

In this paper, we aim to demonstrate the advantages offered by ML, replacing traditional T_c determination using density functional theory (DFT) and Monte Carlo (MC) simulations. We focus on the prediction of the magnetic critical temperature for ordered (including the phases $L2_1$, $C1_b$, Y, and XA) as well as disordered (including the phases A2 and B2) magnetic Heusler alloys. The critical temperatures were determined in a two-step process of an *ab initio* KKR-GF [35] DFT simulation followed by an MC simulation of the T_c as depicted in the top path of Fig. 1. As both steps are comparable in computational cost, we apply our modeling for the whole process as well as only the MC step, taking advantage of magnetic results obtained in the *ab initio* step.

Beyond that, we discuss the impact of magnetic features for the prediction of high T_c materials and the usability in high-throughput materials screening applications, which do not include DFT-originated features in the first place. This discussion is heavily assisted by the use of explainable artificial intelligence (XAI) techniques, which we demonstrate to be able to explain model predictions based on materials science data and visualize relations in the training data captured by the ML model [36].

II. METHODS AND MATERIALS

A. Data processing & cleaning

The examined data was collected at our institute and published as the Jülich-Heusler-magnetic-database (JuHemd) [37]. It provides not only structural and stoichiometric information on the Heusler compounds but also magnetic data obtained by DFT and Monte Carlo simulations. The target quantity we want to predict in our modeling is the critical temperature T_c of the magnetic ordering. While the JuHemd contains experimental values as well as those based on DFT simulations using GGA [38] and LDA [39] exchange-correlation functionals, we restrict our analysis to the GGA-based values as these are provided for most compounds and provide the most homogeneous data quality.

As a first preparation step, we extract the T_c values together with a set of descriptors for each compound in the database. All information was encoded into a numerical representation and made available for the modeling process. Using the provided metadata to augment the information with additional atomic features, we finally obtain a set of 118 descriptors, as listed in Table I. [40] Before any ML modeling is performed, these descriptors $\{x_i\}$ are then transformed to a standardized form

$$\{z_i\} = \frac{\{x_i\} - \mu_i}{\sigma_i}, \tag{1}$$

using the mean μ_i and standard deviation σ_i of the i-th descriptor in the training set.

Only those compound entries have been included which contain all of the above-mentioned entry labels. Incomplete data points have not been used. Additionally, only magnetic alloys are selected. We chose the magnetic cutoff to be

$$\sum_i |m_i| > 0.1\mu_B, \tag{2}$$

where the m_i denotes the magnetic moment of the atom on site i in the compound's molecular formula. Similarly, we did not include compounds with a simulated $T_c = 0$ K. This leaves us with a final dataset size of 408 Heusler compounds.

Since, during the data processing, incomplete data points for Heusler compounds are removed, there are some elements from the periodic table that are contained in the original JuHemd but are not contained anymore in the processed data. The corresponding densities of these atomic numbers, which originate from these removed elements, represent descriptors with zero variance in every compound. Such descriptors are removed before further processing, as they are meaningless for the ML training and evaluation process. In this paper, of the 118 descriptors, there are 11 descriptors in the dataset with zero variance, which are hence removed. The whole data order has been randomized in order to avoid the clustering of similar data points due to the alphabetical order. This enforces the homogeneity of the dataset, which is necessary for the cross-validation (CV) [41] model evaluation to be meaningful.

TABLE I. List of all features which are contained in the processed data and their corresponding explanation. For all features that were directly derived from the JuHemd, the JuHemd label has been used. Also, JuHemd labels have been included which were used to construct processed quantities even though the original label is not included in the processed dataset due to the format, the quantity is given in the JuHemd.

Label	Description
Non-DFT originated features	
`lattice_constant`[a]	Lattice constant of the Heusler
`formula`[b]	Chemical formula of the compound
`Ferro Density`[b]	Fraction of ferromagnetic elements (Fe, Ni, Co) in the compound
`Rare earth Materials Density`[b]	Fraction of rare earth components in the compound
`Symmetry Code`[b]	An integer encoding the Heuslers symmetry group
`Stoichiometry`[b]	5-Digit integer encoding the stochiometry of the compound
`Density by Atomic Number`[b],[d]	Fractional density of each atomic number is encoded by an individual descriptor
`Atomic Number`[c]	Atomic number of the constituents Z_i
`Number of Neutrons`[c]	Number of neutrons of the constituents
`Nominal Mass`[c]	Nominal mass of the constituents atoms
`Number of Electrons`[c]	Number of electrons of the constituents
`Exact Mass`[c]	Exact mass of the constituents atoms
`Atomic Radius`[c]	Atomic radii of the constituents atoms
`Number of Valence Electrons`[c]	Number of valence electrons of the constituents atoms e^{val}
`Covalence Radius`[c]	Covalence radius of the constituents atoms
`Period`[c]	Period number in the PSE of the constituents atoms
`Electronegativity`[c]	Electronegativity of the constituents atoms χ_i
`Van der Waals Radius`[c]	Van der Waals radius of the constituents atoms r_i^{vdw}
`Electron Affinity`[c]	Electron affinity of the constituents atoms $E_{ea\ i}$
DFT originated features	
`Individual Magnetic Moments`[b]	Individual magnetic moments m_i of all constituent atoms
`Absolute Magnetic Moments`[b]	Individual absolute magnetic moments $\lvert m_i \rvert$ of all constituent atoms
`Total magnetic moment`[b]	$M = \sum_i m_i$
`Sum of absolute magnetic moments`[b]	$M_{\mathrm{Abs}} = \sum_i \lvert m_i \rvert$
`etotal (Ry)`[a]	Total energy of the compound E_{Tot}
`Magnetic State`[c]	Integer encoding the magnetic state (ferro, AFM, and spin spiral)

[a]Available directly from JuHemd.
[b]Constructed descriptors.
[c]Added atomic descriptors—most have four entries per compound.
[d]This feature has as many entries (columns) as the JuHemd contains a plurality of unique elements from the PSE.

The code of the data processing script, as well as the code used to generate the following results and figures, is available [42]. This allows us to reevaluate the models if more data is added to the JuHemd. Figure 2 shows the distribution of atomic numbers across different lattice sites in the Heusler compounds. One can see that manganese, chromium, and iron are contained in a large portion of compounds in the dataset.

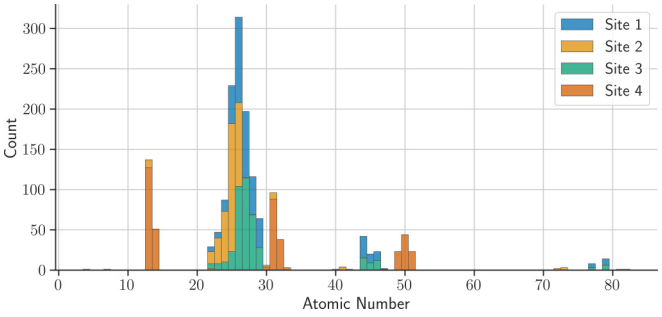

FIG. 2. Distribution of atomic numbers in the GGA dataset after processing and cleaning.

B. Model goals & evaluation

The prediction of T_c using the descriptors outlined in the previous section leads to a classical regression task. Such regression models aim at predicting T_c as accurately as possible. Different metrics are available to evaluate their performance. The evaluation method of choice is also determined by e.g., the error which is desired to be minimized and the importance and impact of outliers in the prediction. The metric used for regression models during this work is the coefficient of determination (Denoted as R^2) for test sets, as well as the CV scores. R^2 measures how well the describing features explain the change in the target variable. Hence, we can be sure to choose a model which properly links the descriptors to T_c.

Besides the regression, we also transformed our problem into a classification task. For the critical temperature, this can be done if one is interested in T_c to be in a certain

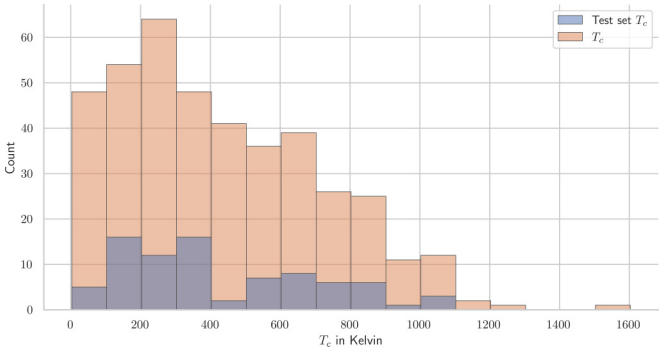

FIG. 3. Distribution of T_c values in the total dataset as well as in the test set only.

range. Industrial applications [43] as magnetic storage devices, for example, typically require magnetic materials to have a T_c above 60 °C in operating conditions. To maintain this comfortably and ensure long-time magnetic stability at those temperatures, we decided on a threshold of 140 K above 60 °C as T_c for a Heusler compound to be considered as "High" T_c [16]. This gives a classification threshold of 200 °C in total. For other thresholds (30 °C, 60 °C, and 90 °C) performance metrics have been computed and the results are referenced in the results section to get a broader view on different threshold choices. It is anticipated that the performance varies to some degree depending on the chosen binary classification threshold and the distribution of compounds within the classification ranges.

Classification typically represents an easier modeling task, as the predictive process is less demanding compared to a regression problem. Hence, if one is only interested in magnetic Heusler alloys, which are candidates for an industrial application, but the exact value of T_c is not of interest in the first place—as the exact value could still be determined in a later step using the established *ab initio* + MC method for the compounds classified as potentially relevant—one can stick to classifying model algorithms. This type of classification model can be used to filter a large number of potential compounds to determine which should be examined further, e.g., by a DFT calculation in a high-throughput materials screening context.

For the classification task, additional considerations on how to evaluate the model performance have to be made. The number of correctly predicted categories would be called the accuracy. However, the errors made in the classification do not have the same significance. If a compound is classified as a "low T_c" but actually has a "high T_c;" this means the model misses out on a material with a potential industrial application. The other error the model can make is classifying a "low T_c" compound as a "high T_c" compound [44], which in the worst case means a waste of computational resources in the example above. Therefore, the goal for a classification model in this application has to be to minimize data points falsely classified as "low T_c" while still keeping the number of falsely as "High T_c" classified compounds low in order not to waste too many computational resources on these false positives. Hence, we decided to continue with the balanced F1 score, which represents a trade-off between precision and recall.

The model performance is determined using 20 % of our data as a test dataset. This test set has been picked randomized out of the whole dataset and is used for calculating the test scores only. This gives an insight into how the model would perform on similar but unseen data. Fourfold CV scores were used in the course of this research in order to perform hyperparameter optimization using a grid search algorithm [45,46]. Hence, for this hyperparameter optimization, we again partition the training data into a 20% validation set for each individual CV fold and use only the remaining 60% for training. After the hyperparameter optimization, the validation set is included to train the model using the best-performing hyperparameters before proceeding with the testing.

The distribution of the T_c values in the test set is displayed in Fig. 3. The values above 1500 K can be considered outliers and are hence removed from the dataset before the data is used in an ML workflow.For all shown scores, the closer the score is to 1.0, the better the model's predictive performance is.

C. ML techniques

The zoo of ML models and techniques continues to grow year by year. It has already grown to such an extent that it is impossible to cover all possibilities and learning algorithms in a single paper. Hence, we limited our analysis to frequently used and established models. It is also worth mentioning that we excluded neural network models (NNMs) from our research on this dataset due to the tabular nature of the data [47,48].

Before training and evaluating models, it is usually not possible to anticipate which model will perform best on a given dataset. This is commonly referred to as the "no free lunch theorem" [49]. Hence, the regression models we evaluated are depicted in the following table:

Linear	Nonlinear	Ensemble
LASSO	K-Nearest Neighbors	Extra Trees
LASSOLars	Decision Tree	Random Forest
Linear Regression		

We have also examined classification models based on similar learning algorithms as some of the regression models depicted in the previous table, as well as a layered indirect classification based on the prediction of the regression models. The indirect classification has been performed to be able to compare the performance of the regression models to their classification counterparts. Since classification is an inherently less complex task than regression, the models would be hard to compare otherwise. The reason underlying this comparison is to determine the best-performing overall model to be used for the feature importance analysis.

Generally, we want our models to properly generalize to unseen data. As CV is used to choose a model's hyperparameters, one could argue that we essentially influence the CV scores by choosing a set of hyperparameters to maximize the score. Therefore, the most relevant metric to choose a model in this case is the model performances on unseen test

data, which is what we used to select models for the feature importance analysis. During this study, we depicted the coefficient of determination and F1 scores as relative indicators of model performance. However, in preliminary tests of this study, we observed that the mean absolute error as absolute measure in the regression case of the model performance on unseen data yields a similar model ranking. The automatic computation of the mean absolute error for regression models is also included in our code publication. [42] The data we produced by processing the original database is also publicly available. [50]

In a setting where these models would be used for materials screening on unseen and heterogeneous data, we would recommend other additional tests and evaluations beyond test set metrics. This could also include evaluation of the trained model's response to unreasonable inputs (such as edge cases) as well as inputs close to known data but with very minor deviations in key features.

D. Feature importance

ML algorithms can be used as black boxes, simply yielding a desired prediction. However, by not applying XAI techniques to understand the model's prediction, we could miss out on the opportunity to improve our understanding of the underlying physics and validate that the model, indeed, has learned physical key properties and relations. It is considered a best practice to perform feature importance analysis using the model which performs best. This is possible by using the SHAP package [51] including the inbuilt visualization options for the SHAP values. SHAP values represent an ML-specific case of the coalition game theory originated Shapley values [52]. SHAP values can be considered as the estimated average contribution of an individual feature—given a set of features—to the deviation of a predicted value from the mean prediction. Hence, Shapley values can be interpreted as a "driving force" of individual features away from the mean prediction. This allows us to explain the model's prediction locally for each individual prediction and globally for a set of predictions [51]. The SHAP package is, in principle, model agnostic but has routines optimized for certain model types such as the tree-based model [53].

III. RESULTS & DISCUSSION

In the following, we showcase the scores and results we achieved in training different ML models. In the spirit outlined in the introduction, we investigated the case in which we used descriptors, including results from the DFT simulations, to only learn the results of the Monte Carlo step first. In a second, independent analysis, we neglected all descriptors that are only available after the DFT simulation and tried to predict T_c values by using only the atomic data.

For the classification, we will discuss the best-performing model and the differences between direct and indirect classification for both the complete descriptor set and the reduced descriptor set.

TABLE II. Regression scores of trained models using the full dataset including *ab initio*-originated descriptors. The rows show the linear models, the next rows show the nonlinear models, and the final rows show the ensemble models.

	CV Score	Train R^2	Test R^2
LASSOLars	$<< 0$	0.77	0.65
LASSO	0.66	0.78	0.66
Linear Reg.	$<< 0$	0.77	$<< 0$
Decision Tree Regression	0.59	1.0	0.62
KNN	0.49	0.66	0.57
Extra Trees	0.77	1.0	0.85
Random Forest	0.74	0.97	0.82

A. Complete descriptor set

1. Regression

A first impression of the predictive performance of two different regression models can be obtained from Fig. 4. For a simple linear model (LASSO) as well as a more complex extremely randomized trees (extra trees) regression model, we report the predicted value of T_c in relation to the value obtained from the full simulation for our test set. While the Lasso results show a systematic error by underestimating the higher values of T_c while overestimating the critical temperature for the low T_c Heuslers, this deficiency is substantially reduced in the extra trees model. In addition, this model also reproduces the distribution of the values much more accurately and shows less spread around the ideal red line.

This can be confirmed by the plot (Fig. 5) of the difference between the simulated (true) value and the prediction, showing a low relative residual error for the whole temperature range except the very low T_c values. This error for low values arises from the scale of the very low-temperature values, which enlarges the relative error due to its fractional nature.

A more comprehensive overview of the results of various regression models can be obtained from Table II, in which we report some performance indicators for the linear, nonlinear, and ensemble models.

It is clearly visible that the ensemble models are performing best on the test set and in the CV scoring. However, these good predictions are accompanied by a high degree of overfitting to the training data, easily recognizable by the nearly perfect R^2 score on the training set, i.e., a very low or even close to vanishing bias [46]. The fact that only the ensemble models exhibit a reasonably low bias indicates that the complexity of other models does not meet the complexity of the quantity to predict and/or the data. In general, the model complexity has to be adjusted to the data complexity [54]. It is clear that a simple linear regression, as well as the K-nearest neighbor model, does not meet this requirement in our case. This finding reflects the complexity of the physical processes responsible for the emergence and stability of magnetic phenomena [55].

The typical approach to cope with overfitting is increasing the regularization [56]. However, even by applying regularization, we could not determine models with improved CV scores, which itself indicated that a lack of regularization is not the root cause for the overfitting. Moreover, when dealing

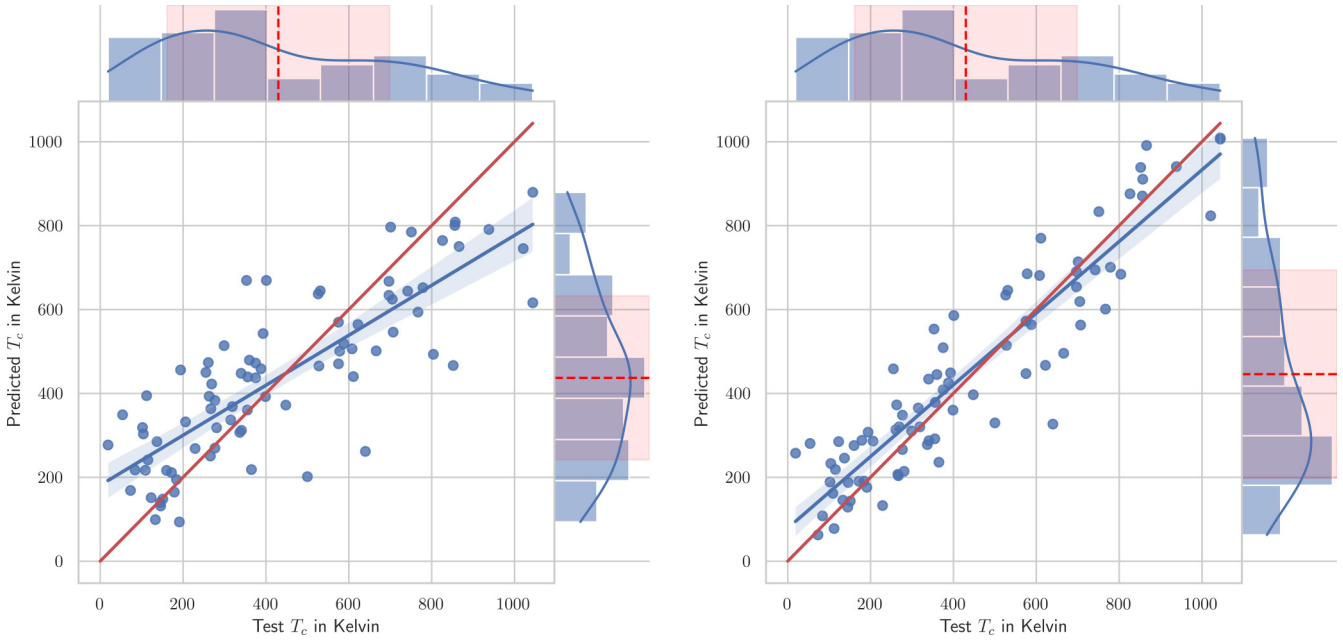

FIG. 4. Prediction result of two models for the test set. On the left, the data for the linear LASSO model is shown, while the right panel shows the data from the extremely randomized trees (extra trees) regression model [57]. The red line indicates a perfect match between the predicted and expected data and the blue line (with shade) is a linear regression through the predicted data points (with a 95% CI envelope for the regression, computed using a bootstrap [58] based method). On the side distribution plots of the test sets, true T_c values and the predictions are added.

with different iterations of the dataset over the course of this study we observed an improvement of the model performance, e.g., seen in a decreasing variance, with every increase of the total amount of included Heusler compounds. This indicated that a lack of training data causes a high variance for the more complex models. This also explains the higher test score compared to the CV score. The model in the test case had the full training data available, while for the calculations of the CV score each of the four CV scores—which are depicted here—had only 75% of the training set available for training.

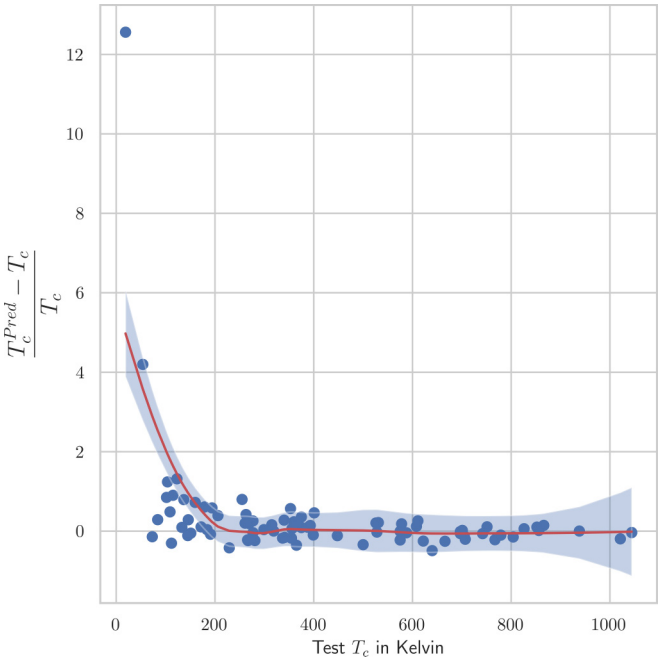

FIG. 5. Residual depiction of the extra trees regression shown on the right of Fig. 4 including a LOWESS smoothing applied to the data points with a pointwise 95% CI envelope.

2. *Classification*

Since classification is a significantly easier task than regression, we expect to see an improved model performance for each classification model compared to the regression case on this dataset. In Table III, the results of each linear, nonlinear, and ensemble classification model, as well as indirect classification models based on a linear model and an ensemble regression model from Sec. III A 1, are displayed.

As expected, the CV scores of the classification models are significantly higher than the scores of the regression models,

TABLE III. Direct and indirect classification scores of a model selection using the full dataset, including *ab initio*-originated descriptors. The rows show the linear models, the next rows the single tree-based model, and the final rows the ensemble-based model setups. Results from other threshold choices are available in Table VI.

	CV Score	Train F1	Test F1	Test Accuracy
Logistic Reg.	0.82	0.91	0.86	0.89
Indirect LASSO		0.86	0.81	0.85
Decision Tree	0.74	1.0	0.75	0.77
Extra Trees	0.82	1.0	0.91	0.93
Indirect Extra Trees model		1.0	0.89	0.92

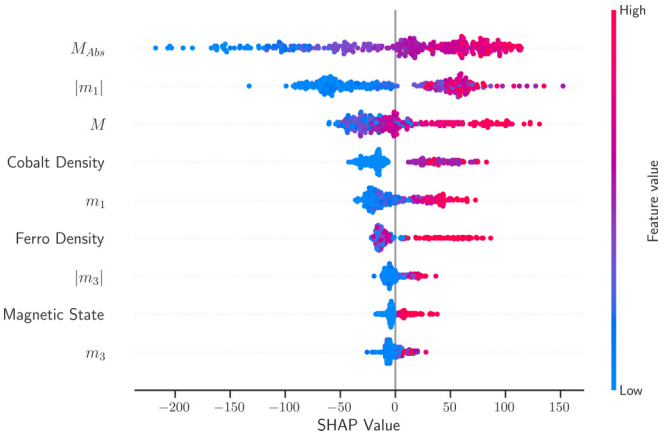

FIG. 6. SHAP beeswarm summary plot of the nine descriptors with the largest SHAP values [51,52].

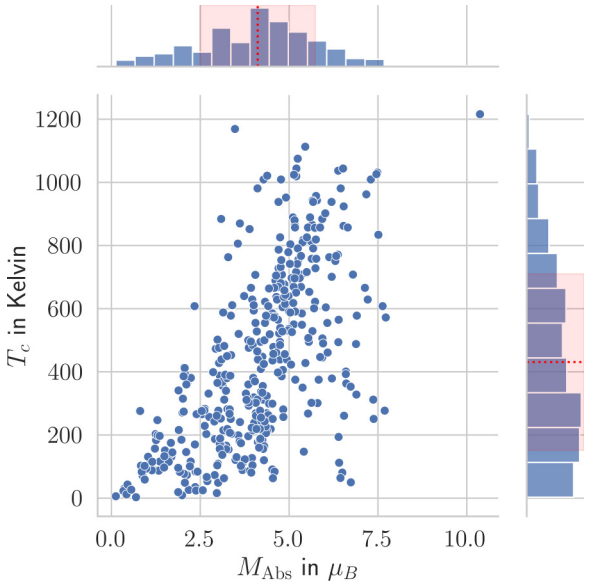

FIG. 7. Relation between absolute magnetic moment of the Heusler compounds to their T_c values for the whole dataset.

which corresponds to a lower bias. Similarly, the results for the test set are closer to the ideal prediction, as there is less variance occurring than for the regression models. This aligns with our interpretation of the overfitting in the regression case due to the fact that classification is an easier task, so there is less training data required to fit classification models to the data as the complexity of the quantity to predict is reduced from a continuous quantity to a binary value. This reduction is only possible as we know which minimum T_c values are required to be relevant to an industrial application.

3. *Feature importance*

After searching for a working set of hyperparameters in Sec. III A 1 using the training set, we used the determined hyperparameters and chose the extra trees regression [57] as our best-performing model to conduct a feature importance analysis using the SHAP package [51] and the corresponding SHAP values. The SHAP values have been calculated for the training dataset. These values for the most relevant features are shown in Fig. 6.

Besides the SHAP values, the color of the data points encodes the relative scale of the feature for each individual data point. This means that if there is a clear horizontal color fade visible, this implies a systematic impact of this feature for the predicted quantity.

From Fig. 6 one can see that for the extra trees model, the absolute magnetic moment of the compound has the largest impact on the T_c prediction. All nine most relevant quantities are either magnetic moments or indirectly related to magnetism (e.g., the Cobalt density of the compound), which confirms that the magnetic material-specific properties indeed have the largest impact on the value of the critical temperature. While all the nine quantities are positively correlated to T_c, i.e., have an increasing impact on the T_c when they increase too. For some of the quantities, this is of course an artifact of our descriptor construction. For example, we encode the magnetic state as an integer, with the smallest possible encoding 000 denoting that the material forms neither a ferromagnet, an antiferromagnet, nor a spin spiral. In contrast, the fact that the model assigns the most significance to the nine quantities listed here was obtained without providing any physical knowledge of the system, besides the fact that we included these descriptors in the first place. Thus, the modeling singled out that these parameters indicate the kind of "physical insight" that can be obtained from ML. For example, the high relevance of the absolute magnetic moment of the compound for the T_c is of course in line with the relation one would obtain from even very basic physics models of magnetism.Looking at Fig. 7, one can draw even more conclusions. It can be seen that T_c is not simply proportional to M_{Abs}. Instead, Heusler alloys with a higher M_{Abs} can show a higher T_c. However, while a high absolute magnetic moment does not guarantee the emergence of a high T_c, a low M_{Abs} prevents the occurrence of high T_c values. Therefore, it is safe to say M_{Abs} is acting as an upper boundary to T_c together with some constant factor C:

$$T_c \leqslant C M_{\mathrm{Abs}}, \tag{3}$$

Plausible values of C, replicating the observed upper boundary behavior, in this study range from $150\frac{\mathrm{K}}{\mu_{\mathrm{B}}}$ to $170\frac{\mathrm{K}}{\mu_{\mathrm{B}}}$. M_{Abs} is defined as the sum of the individual absolute constituents' magnetic moments $|m_i|$.

A preliminary investigation of the feature importance using only features remaining after a LASSO shrinkage analysis showed notable robustness of the used models toward the large number of features used within this study's application given this feature selection method. Even though this feature reduction slightly decreased the performance of the tree-based models the corresponding scores remained in a similar range. This preliminary investigation was deepened upon reviewers' request using the 10 most impactful features selected by the presented SHAP-based feature importance approach. Only using these 10 descriptors together with a retrained extra trees regression model, including the DFT-originated descriptors resulted, in a slight change in the coefficient of determination (0.84) and a slightly lowered CV score (0.73). The train scores in this retraining process remained unchanged.

TABLE IV. Regression scores of promising models using the reduced dataset, excluding *ab initio*-originated descriptors.

	CV Score	Train R^2	Test R^2
Extra Trees	0.52	1.0	0.76
LASSO	0.31	0.58	0.63

4. Computational efficiency

It is worth mentioning that the computational cost of performing the *ab initio* calculations followed by the MC in order to obtain the simulated critical temperature can easily reach an order of 10^3 core hours on a high-performance computing cluster per compound. Compared to this, the resource consumption of any ML model presented within this study for training, tuning, and evaluation is significantly smaller. The models can be trained and tuned, with regard to the models' hyperparameters, within an hour on a regular laptop. The evaluation of the model requires a set of features, to predict or classify the critical temperature of a compound. This evaluation step consumes comparably neglectable computational resources and takes only a few seconds.

B. Dataset without DFT-originated descriptors

1. Regression

Retraining the extra trees regression model as well as the LASSO model to the reduced descriptor set and again performing hyperparameter optimization using a grid search algorithm, we achieved the regression scores displayed in Table IV.

As expected, one can observe a clear decrease in performance compared to the case where DFT-originated descriptors have been used. In particular, the LASSO results now have huge deviations such that one could question its fitness for any practical application. Therefore, we can already conclude, that a prediction of the critical temperature without the use of the basic magnetic properties predicted by a DFT simulation is not really possible in our scenario. Therefore, we decided not to analyze this further, but to concentrate on the easier classification task.

2. Classification

The achieved classification model results using no DFT-originated descriptors at all are displayed in Table V. This table contains exactly the same models as seen before in Table III.

From the test set of 82 compounds, our constructed indirect extra trees classification model managed to correctly classify 47 "Low" T_c and 29 "High" T_c compounds. Of each class, three compounds have been wrongly predicted. We consider falsely classifying a "Low" T_c compound as a "High" T_c not so relevant for practical application. The worst outcome in a potential use case is that the model suggests a "High" T_c compound, and when computing it using a more sophisticated—and hence computationally more intensive—approach, one finds that the "High" T_c label has been falsely assigned. However, if a "High" T_c compound is classified as "Low" T_c in a high-throughput screening process, it will prob-

TABLE V. Direct and indirect classification scores of a model selection using the reduced dataset, excluding *ab initio*-originated descriptors. The rows show the linear models, the next rows are the single tree-based model, and the final rows are the ensemble-based model setups. Results from other threshold choices are available in Table VII.

	CV Score	Train F1	Test F1	Test Accuracy
Logistic Reg.	0.68	0.75	0.75	0.79
Indirect LASSO	n/a.	0.75	0.8	0.88
Decision Tree	0.66	1.0	0.8	0.84
Extra Trees	0.74	1.0	0.84	0.87
Indirect Extra Trees model	n/a.	1.0	0.91	0.93

TABLE VI. Scores resulting from evaluating a different threshold in the binary classification process using also DFT-originated feature input.

Threshold	30 °C	
Model	Test F1	Test Accuracy
Extra Trees	0.88	0.85
Decision Tree	0.88	0.85
Log. Reg.	0.83	0.79
KNN	0.79	0.77
Threshold	60 °C	
Extra Trees	0.85	0.83
Decision Tree	0.84	0.82
Log. Reg.	0.87	0.84
KNN	0.78	0.77
Threshold	90 °C	
Extra Trees	0.83	0.82
Decision Tree	0.77	0.76
Log. Reg.	0.83	0.82
KNN	0.78	0.78

TABLE VII. Scores resulting from evaluating a different threshold in the binary classification process not using DFT-originated feature input.

Threshold	30 °C	
Model	Test F1	Test Accuracy
Extra Trees	0.78	0.73
Decision Tree	0.74	0.68
Log. Reg.	0.77	0.7
KNN	0.71	0.7
Threshold	60 °C	
Extra Trees	0.75	0.72
Decision Tree	0.73	0.68
Log. Reg.	0.74	0.66
KNN	0.72	0.73
Threshold	90 °C	
Extra Trees	0.73	0.71
Decision Tree	0.7	0.68
Log. Reg.	0.73	0.68
KNN	0.72	0.74

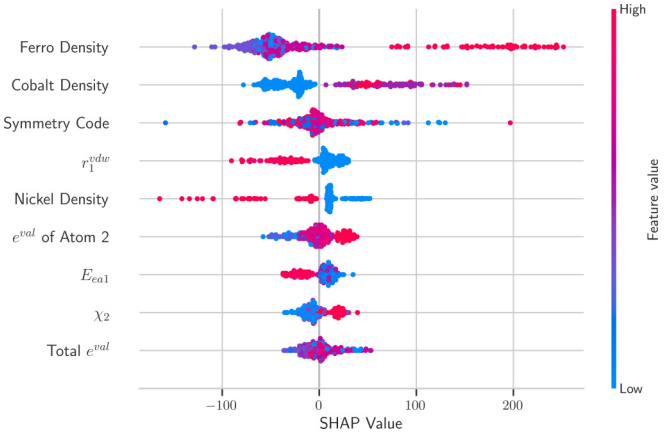

FIG. 8. SHAP beeswarm summary plot of the nine descriptors of the dataset without the DFT-originated data with the largest SHAP values.

ably never be computed with a more sophisticated approach, which causes this compound to be potentially "lost" for future research. In the case of this model, we saw that this crucial error for falsely classifying a "High" T_c Heusler as a "Low" T_c Heusler is below 5% and hence meets typical confidence criteria. This concludes that the indirect extra trees classification is capable of classifying the T_c in "High" and "Low" values even without the DFT-originated data. While "Low" means T_c is too low to be relevant for industry applications.

3. Feature importance

Computing the SHAP values for the reduced descriptor set and visualizing them as we did previously results in the beeswarm plot shown in Fig. 8. As we can see in Fig. 8, removing the DFT-originated descriptors and, therefore, the calculated magnetic moments caused other quantities to become more impactful. As one can expect, these are very closely related to the magnetic moments (e.g., the density of ferromagnetic materials in the compound as well as the cobalt and nickel densities). However, now we observe more complex relations than in the previous feature importance plot, demonstrating the lower significance of these quantities for the critical temperatures. We can see a negative correlation between the van der Waals radius of the atom on site one (r_1^{vdw}), the nickel density in the compound, and the electron affinity of the atom on site one (E_{ea1}) with a decreasing T_c as these quantities increase. For the fraction of ferromagnetic atoms, the effect is much less obvious. We can see that very high densities of ferromagnetic atoms in the compound contribute to a largely increased T_c prediction. However, on the other hand, a low density of ferromagnetic atoms does not lead to an equally decreased prediction of T_c. Interestingly, this reflects our previous result that a large absolute magnetic moment corresponds to an upper boundary for the T_c. Since a large amount of ferromagnetic compound constituents is highly correlated with a large magnetic moment. The required and obtained model complexity is also observed in the SHAP values of the symmetry code. Since this is an arbitrary-ordered label for the symmetry group of the compound, there is no clear order of the feature value that correlates with the T_c. However, the model seems to have learned that some feature values have a larger impact on T_c than others, which is indeed possible.

As the density of the ferromagnetic atoms, the cobalt and nickel atoms turn out to be relevant quantities in Fig. 8; we investigated their correlation with T_c in more detail as depicted in Fig. 9. The depicted fractional density histograms confirm the trends we were hinted at by the SHAP beeswarm plot. It is easily visible that a large density of ferromagnetic atoms in the compound is indeed contributing to a larger T_c value, with one exception: The antiferromagnetic case. We can see that there are a few materials that have no ferromagnetic atoms in the compound at all but still a very high T_c. These are strong antiferromagnets. This finding can be related to our previous result for the modeling, including the DFT-based magnetization values, in which we have seen that the SHAP values of M_{Abs} hint at a larger impact than those of M. As the antiferromagnetic compounds have a vanishing M but a large M_{Abs} while resulting in a stable antiferromagnetic state with a large T_c. The same relation is, in principle, true for the cobalt density. However, there are fewer compounds containing cobalt in the data than iron. The Nickel density has—for increasing densities—a negative impact on T_c according to Fig. 8, and as we can see in Fig. 9, this also emerges from the data. It seems that the presence of Cobalt is not as helpful in contributing to a stable magnetic state as e.g., iron.

Insights from feature importance analysis can be used to deepen our understanding of how the modeled target quantity (in this case the critical temperature) correlates to different features. There is no guarantee that this correlation has an underlying causal relationship. Nevertheless, the feature importance analysis can suggest further points of investigation into certain features over others which are not limited to, e.g., linear correlation.

Additionally, the insights from feature importance analysis can assist in selecting promising candidates prior to any computation or experimentation depending on which information is available beforehand. One example of such an insight is that an increasing nickel density seems not to favor a large critical temperature. The features found relevant should be considered as starting points for further investigation. Such observed trends and correlations could be used in a high-throughput simulation context to prioritize compounds prior to computation in order to compute correlative more promising compounds, based on the feature importance, first. While this can be applied for Heusler alloys and the critical temperature, this approach is not limited to a material class or this particular property to predict.

IV. SUMMARY AND OUTLOOK

This work can be seen as a small-scale sandbox-type case study in which lightweight ML algorithms can add value to existing *ab initio* data and eventually replace costly computational steps in layered calculation workflows in the future.

It was demonstrated that qualitative predictions for material-specific properties are achievable with very small errors, even for the limited dataset sizes common in materials science. Also, the expectation that the quantitative prediction is much more difficult and requires descriptors with much higher predictive power, has been confirmed. However,

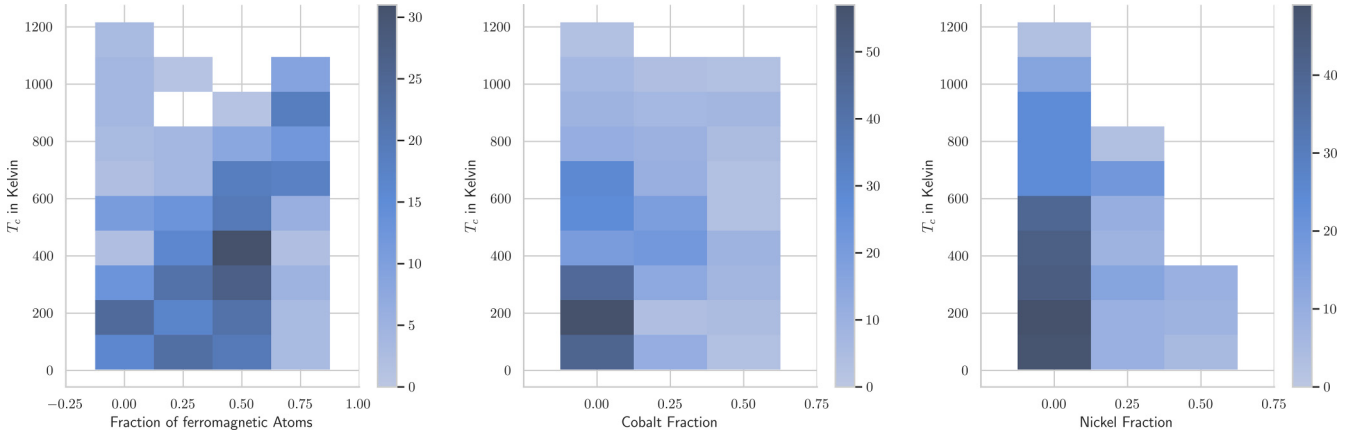

FIG. 9. Relation between T_c and the fraction of ferromagnetic elements (left), cobalt atoms (center), and nickel (right) shown as a heatmap-style histogram. The darker the color, the more compounds can be found in the colored region.

we could also demonstrate that even very simple and readily available descriptors not based on any actual calculation in combination with sufficiently complex models could be utilized in a classification task typically part of any high throughput screening. As demonstrated in this paper, there is a potential use for ML methods in materials science, even in quantitatively predicting properties as complex as the T_c. It is imaginable to perform similar studies on existing datasets of other material families beyond the Heusler alloys. However, one has to consider that the structural homogeneity of the material class we studied here simplified the complexity of the modeling task. This implies that if one would translate the methodical insights gained from this dataset of Heusler alloys to a different material class, there should either be a similar structural homogeneity, or if the structural complexity is increased one also has to choose descriptors with equally elevated descriptive value.

By performing feature importance analysis with XAI techniques—such as SHAP values—we gained physical insights about the relations of the target quantity to the included features, as well as the determining properties of the studied material class given in the examined dataset. Such analysis can provide a link between a complex ML process with a hard-to-expose underlying mechanism and true physical insight and the gain of knowledge of the system. In this study, we rediscovered dependencies expected from simple physical models without actually providing such knowledge to the process.

Finally, we would like to stress that the methodical approach described in this paper is not limited to predicting T_c or any other magnetic quantity, but that it can be transferred to any other material-specific property. We believe it is even possible to discover that known materials have currently unknown properties using predictive modeling.

ACKNOWLEDGMENTS

This work was performed as part of the Helmholtz School for Data Science in Life, Earth, and Energy (HDS-LEE) and received funding from the Helmholtz Association of German Research Centres. Since parts of the data processing have been performed and the displayed visualizations have been created using dedicated open-source packages, we acknowledge them here [59–67]. We acknowledge Stefano Sanvito for the inspiration to represent the fractional density of each atomic number in the dataset as a standalone descriptor, which we gathered from one of his talks. The authors thank Fabian Lux for initially hinting at the extra trees regression model. We acknowledge Dirk Witthaut for pointing out the advantages of SHAP values in XAI in comparison to model-bound feature importance methods. We want to acknowledge the productive work together with the reviewers to improve the clarity and content of the manuscript. The authors thank Roman Kováčik for discussing the structure and contents of the data published in Ref. [37]. Hence, we acknowledge the computing time granted by the RWTH Aachen University (Project jara0182) which was required in order to collect the data. As the original database as well as the ML-ready data was hosted by Materials Cloud, we acknowledge them [68].

DATA AVAILABILITY

The data that support the findings of this article are openly available [37,50].

APPENDIX

In order to provide a broader view on the presented approach we also computed binary classification scores of models using different thresholds. Previously, in the classification process the threshold has been chosen to be 200 °C. Table VI provides scores for models trained on thresholds 30 °C, 60 °C, 90 °C using the complete feature set, including DFT-originated features. Table VII shows the scores using the same thresholds for models trained on the reduced feature set, without the DFT-originated descriptors.

[1] H. Shi, S. Liu, J. Chen, X. Li, Q. Ma, and B. Yu, Predicting drug-target interactions using lasso with random forest based on evolutionary information and chemical structure, Genomics **111**, 1839 (2019).

[2] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, R. H. Mak, R. M. Tamimi, C. M. Tempany, C. Swanton, U. Hoffmann, L. H. Schwartz, R. J. Gillies, R. Y. Huang, and H. J. W. L. Aerts, Artificial intelligence in cancer imaging: Clinical challenges and applications, CA: Cancer J. Clin. **69**, 127 (2019).

[3] R. V. Shah, G. Grennan, M. Zafar-Khan, F. Alim, S. Dey, D. Ramanathan, and J. Mishra, Personalized machine learning of depressed mood using wearables, Transl. Psychiatry **11**, 338 (2021).

[4] T. Zhou, Z. Song, and K. Sundmacher, Big data creates new opportunities for materials research: A review on methods and applications of machine learning for materials design, Engineering **5**, 1017 (2019).

[5] H. J. Kulik, T. Hammerschmidt, J. Schmidt, S. Botti, M. A. L. Marques, M. Boley, M. Scheffler, M. Todorović, P. Rinke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A. P. Bartók, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit *et al.*, Roadmap on machine learning in electronic structure, Electron. Struct. **4**, 023004 (2022).

[6] Y. Igarashi, K. Nagata, T. Kuwatani, T. Omori, Y. Nakanishi-Ohno, and M. Okada, Three levels of data-driven science, J. Phys.: Conf. Ser. **699**, 012001 (2016).

[7] P. J. Garcia Nieto, E. García-Gonzalo, and J. Paredes-Sánchez, Prediction of the critical temperature of a superconductor by using the WOA/MARS, Ridge, Lasso and Elastic-net machine learning techniques, Neural Comput. Appl. **33**, 17131 (2021).

[8] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics, J. Phys. Chem. Lett. **9**, 2879 (2018).

[9] K. Takahashi and Y. Tanaka, Material synthesis and design from first principle calculations and machine learning, Comput. Mater. Sci. **112**, 364 (2016).

[10] A. Jain, G. Hautier, S. P. Ong, and K. Persson, New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships, J. Mater. Res. **31**, 977 (2016).

[11] H. Ucar, D. Paudyal, and K. Choudhary, Machine learning predicted magnetic entropy change using chemical descriptors across a large compositional landscape, Comput. Mater. Sci. **209**, 111414 (2022).

[12] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, Machine learning in materials informatics: Recent applications and prospects, npj Comput. Mater. **3**, 54 (2017).

[13] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater. **5**, 83 (2019).

[14] O. Gutfleisch, M. A. Willard, E. Brück, C. H. Chen, S. G. Sankar, and J. P. Liu, Magnetic materials and devices for the 21st century: Stronger, lighter, and more energy efficient, Adv. Mater. **23**, 821 (2011).

[15] V. Laliena, G. Albalate, and J. Campo, Stability of the skyrmion lattice near the critical temperature in cubic helimagnets, Phys. Rev. B **98**, 224407 (2018).

[16] T. Shang, E. Canévet, M. Morin, D. Sheptyakov, M. T. Fernández-Díaz, E. Pomjakushina, and M. Medarde, Design of magnetic spirals in layered perovskites: Extending the stability range far beyond room temperature, Sci. Adv. **4**, eaau6386 (2018).

[17] A. Halder, A. Ghosh, and T. S. Dasgupta, Machine-learning-assisted prediction of magnetic double perovskites, Phys. Rev. Mater. **3**, 084418 (2019).

[18] X. Li, G. Shan, and C. Shek, Machine learning prediction of magnetic properties of Fe-based metallic glasses considering glass forming ability, J. Mater. Sci. Technol. **103**, 113 (2022).

[19] R. A. Hilgers, Prediction of magnetic materials for energy and information combining data-analytics and first-principles theory, Dissertation, RWTH Aachen University, Aachen, 2024, doi:10.18154/RWTH-2024-09243.

[20] R. Hilgers, D. Wortmann, and S. Blügel, Application of batch learning for boosting high-throughput *ab initio* success rates and reducing computational effort required using data-driven processes, Electron. Struct. **7**, 015005 (2025).

[21] K. Liu, B. Ge, F. Liu, M. Feng, Y. Ji, Y. Li, W. Lu, X. Jiang, and Y. Liu, Machine learning assisted development of Heusler alloys for high magnetic moment, Comput. Mater. Sci. **250**, 113692 (2025).

[22] J. F. Belot, V. Taufour, S. Sanvito, and G. L. W. Hart, Machine learning predictions of high-Curie-temperature materials, Appl. Phys. Lett. **123**, 042405 (2023).

[23] J. Nelson and S. Sanvito, Predicting the Curie temperature of ferromagnets using machine learning, Phys. Rev. Mater. **3**, 104405 (2019).

[24] D. Baigutlin, V. Sokolovskiy, V. Buchelnikov, and S. Taskaev, Machine learning algorithms for optimization of magnetocaloric effect in all-d-metal Heusler alloys, J. Appl. Phys. **136**, 183903 (2024).

[25] Y.-C. Tang, K.-Y. Cao, R.-N. Ma, J.-B. Wang, Y. Zhang, D.-Y. Zhang, C. Zhou, F.-H. Tian, M.-X. Fang, and S. Yang, Accurate prediction of magnetocaloric effect in NiMn-based Heusler alloys by prioritizing phase transitions through explainable machine learning, Rare Metals **44**, 639 (2025).

[26] F. Heusler, W. Starck, and E. Haupt, Magnetisch-chemische studien, Verh. Dtsch. Phys. Ges **5**, 219 (1903).

[27] F. Heusler and E. Take, The nature of the Heusler alloys, Trans. Faraday Soc. **8**, 169 (1912).

[28] H. Uzunok, E. Karaca, S. Bağcı, and H. Tütüncü, Physical properties and superconductivity of Heusler compound $LiGa_2Rh$: A first-principles calculation, Solid State Commun. **311**, 113859 (2020).

[29] A. Roy, J. W. Bennett, K. M. Rabe, and D. Vanderbilt, Half-Heusler semiconductors as piezoelectrics, Phys. Rev. Lett. **109**, 037602 (2012).

[30] Q. Gao, I. Opahle, O. Gutfleisch, and H. Zhang, Designing rare-earth free permanent magnets in Heusler alloys via interstitial doping, Acta Mater. **186**, 355 (2020).

[31] S. Idrissi, S. Ziti, H. Labrim, and L. Bahmad, Half-metallicity and magnetism in the full Heusler alloy Fe_2MnSn with L21 and XA stability ordering phases, J. Low Temp. Phys. **202**, 343 (2021).

[32] A. Davidson, V. P. Amin, W. S. Aljuaid, P. M. Haney, and X. Fan, Perspectives of electrically generated spin currents in ferromagnetic materials, Phys. Lett. A **384**, 126228 (2020).

[33] A. Hirohata and D. C. Lloyd, Heusler alloys for metal spintronics, MRS Bull. **47**, 593 (2022).

[34] S. Sanvito, C. Oses, J. Xue, A. Tiwari, M. Zic, T. Archer, P. Tozman, M. Venkatesan, M. Coey, and S. Curtarolo, Accelerated discovery of new magnets in the Heusler alloy family, Sci. Adv. **3**, e1602241 (2017).

[35] P. Rüßmann, Be-Zimmermann, G. Géranton, C. Oran, and E. Rabel, Judftteam/jukkr: v3.6, Zenodo (2022).

[36] X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, and T. Y.-J. Han, Explainable machine learning in materials science, npj Comput. Mater. **8**, 204 (2022).

[37] R. Kováčik, P. Mavropoulos, and S. Blügel, The juhemd (jülich-Heusler-magnetic-database) of the Monte Carlo simulated critical temperatures of the magnetic phase transition for experimentally reported Heusler and Heusler-like materials, Materials Cloud Archive (2022).

[38] J. P. Perdew, K. Burke, and Y. Wang, Generalized gradient approximation for the exchange-correlation hole of a many-electron system, Phys. Rev. B **54**, 16533 (1996).

[39] B. Fricke, W.-D. Sepp, T. Bastug, S. Varga, K. Schulze, J. Anton, and V. Pershina, Use of the DV xα-method in the field of superheavy atoms, in *Advances in Quantum Chemistry* (Elsevier, San Diego, California, 1998), pp. 109–121.

[40] The critical temperature itself is named `resval` within JuHemd.

[41] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed. (Springer, New York, NY, USA, 2009).

[42] R. Hilgers, D. Wortmann, and S. Blügel, Data processing for the juhemd database and ml-training and evaluation scripts, Zenodo (2022).

[43] C. Felser, L. Wollmann, S. Chadov, G. H. Fecher, and S. S. P. Parkin, Basics and prospective of magnetic Heusler compounds, APL Mater. **3**, 041518 (2015).

[44] Beyond binary classification in classes above and below a given threshold we did experiments with more than two classes (e.g. a desired interval of values as well as classes above and below this interval) observed a significantly decreased performance. Such a more detailed classification requires significantly more training data to perform similarly with regards to performance metrics.

[45] M. Feurer and F. Hutter, *Automatic Machine Learning: Methods, Systems, Challenges*, edited by F. Hutter, L. Kotthoff, and J. Vanschoren, Hyperparameter Optimization (Springer, Cham, Switzerland, 2019).

[46] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. (Springer, New York, NY, USA, 2021).

[47] R. Shwartz-Ziv and A. Armon, Tabular data: Deep learning is not all you need, Inf. Fusion **81**, 84 (2022).

[48] L. Grinsztajn, E. Oyallon, and G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Info. Proc. Syst. **35**, 507 (2022).

[49] D. H. Wolpert and W. G. Macready, No free lunch theorems for optimization. IEEE transactions on evolutionary computation, IEEE Trans. Evol. Computat. **1**, 67 (1997).

[50] R. Hilgers, D. Wortmann, and S. Blügel, Ml-ready Curie temperatures and descriptors extracted from the juhemd database, Materials Cloud Archive (2022).

[51] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, USA, 2017), pp. 4765–4774.

[52] L. S. Shapley, A value for n-person games, in *Contributions to the Theory of Games II*, edited by H. W. Kuhn and A. W. Tucker (Princeton University Press, Princeton, 1953), pp. 307–317.

[53] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. **2**, 56 (2020).

[54] L. Li and Y. S. Abu-Mostafa, Data complexity in machine learning, Comput. Sci. Tech. Rep. **54** (2006).

[55] D. Gatteschi and L. Bogani, Complexity in molecular magnetism, in *Complexity in Chemistry and Beyond: Interplay Theory and Experiment*, edited by C. Hill and D. G. Musaev (Springer Netherlands, Dordrecht, 2012), pp. 49–72.

[56] X. Ying, An overview of overfitting and its solutions, J. Phys.: Conf. Ser. **1168**, 022022 (2019).

[57] P. Geurts, D. Ernst, and L. Wehenkel, Extremely randomized trees, Mach. Learn. **63**, 3 (2006).

[58] B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Stat. **7**, 1 (1979).

[59] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant *et al.*, Array programming with NumPy, Nature (London) **585**, 357 (2020).

[60] Łukasz. Mentel, Mendeleev–a python resource for properties of chemical elements, ions and isotopes, Zenodo (2014).

[61] T. Tantau, The TikZ and PGF Packages (2013), https://github.com/pgf-tikz/pgf.

[62] J. D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. **9**, 90 (2007).

[63] T. A. Caswell, M. Droettboom, A. Lee, E. S. De Andrade, T. Hoffmann, J. Hunter, J. Klymak, E. Firing, D. Stansby, N. Varoquaux, J. H. Nielsen, B. Root, R. May, P. Elson, J. K. Seppänen, D. Dale, J.-J. Lee, D. McDougall, A. Straw, P. Hobson *et al.*, matplotlib/matplotlib: Rel: v3.4.3, Zenodo (2021).

[64] M. L. Waskom, Seaborn: Statistical data visualization, J. Open Source Software **6**, 3021 (2021).

[65] W. S. Cleveland, Robust locally weighted regression and smoothing scatterplots, J. Am. Stat. Assoc. **74**, 829 (1979).

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).

[67] S. Seabold and J. Perktold, Statsmodels: Econometric and statistical modeling with python, in *9th Python in Science Conference* (2010).

[68] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, and N. Marzari, Materials cloud, a platform for open computational science, Scientific Data **7**, 299 (2020).