The Eliashberg theory of superconductivity accounts for the fundamental physics of conventional superconductors, including the retardation of the interaction and the Coulomb pseudopotential, to predict the critical temperature $T_c$. McMillan, Allen, and Dynes derived approximate closed-form expressions for the critical temperature within this theory, which depends on the electron–phonon spectral function $\alpha^2 F(\omega)$. Here we show that modern machine-learning techniques can substantially improve these formulae, accounting for more general shapes of the $\alpha^2 F$ function. Using symbolic regression and the SISSO framework, together with a database of artificially generated $\alpha^2 F$ functions and numerical solutions of the Eliashberg equations, we derive a formula for $T_c$ that performs as well as Allen–Dynes for low-$T_c$ superconductors and substantially better for higher-$T_c$ ones. This corrects the systematic underestimation of $T_c$ while reproducing the physical constraints originally outlined by Allen and Dynes. This equation should replace the Allen–Dynes formula for the prediction of higher-temperature superconductors.

### INTRODUCTION

Although the theory of electron–phonon superconductivity due to Bardeen–Cooper–Schrieffer, Gor’kov, Eliashberg, Migdal, and others is well-established, it has not historically aided in the discovery of new superconductors. The materials space to search for new superconductors is vast, and it is, therefore, desirable to find a practical way to use theory as a guide. Recent computational developments may allow a new approach to superconducting materials discovery based on ab-initio and materials-genome type methods.

One approach to this problem, pioneered by McMillan and Allen and Dynes, is to search for a formula for $T_c$ based on materials-specific parameters derived from the Eliashberg equations of superconductivity. These parameters, mostly moments of the electron–phonon spectral function $\alpha^2 F(\omega)$, can be determined by experiment or, more recently, calculated within ab initio approaches. In principle, this allows one also to deduce how to optimize $T_c$ if one can optimize one or more of these parameters.

The Allen–Dynes equation has played a crucial role in debates on how to achieve high-temperature superconductivity by both theorists, who use it to predict $T_c$, and by experimentalists, who extract $\lambda$ from measured $T_c$ and $\omega_0$. Nevertheless, it is important to recall that the Allen–Dynes equation has been derived from Eliashberg theory within an approximation where the momentum dependence of the Eliashberg function is neglected. It is based on 217 Eliashberg solutions of three types of $\alpha^2 F(\omega)$ shapes (those obtained from tunneling data on Hg and Pb, and those obtained for a single Einstein model).

There have been several important advances in providing more detailed solutions to the Eliashberg equations since the work of Allen and Dynes. Combescot solved the Eliashberg equations on the weak coupling side and obtained an expression for $T_c$ that depends on $\omega_0$ and a shape-dependent integral. Recently, Marsiglio et al. solved the Eliashberg equations numerically at small $\lambda$, exhibiting some deviations of the theory from the BCS result in this limit, in particular the correction $\frac{1}{2} \xi$ to the BCS $T_c$. And of course the full equations can be solved numerically for any coupling, including the momentum dependence of $\alpha^2 F$ if desired. The validity of the theory as $\lambda$ increases is a subtle question, and has been the subject of a number of recent studies.

In this paper, we solve the Eliashberg equations using different types of electron–phonon spectral functions, including multimodal Einstein-like spectra and a set of $\alpha^2 F$ obtained from first-principles calculations. We find that, while the Allen–Dynes formula accurately predicts the Eliashberg $T_c$ for $\lambda$ values near 1.6 (the coupling constant for Hg and Pb), it nevertheless deviates from the Eliashberg $T_c$ when $\lambda$ is significantly larger or smaller than 1.6 and when the shape of $\alpha^2 F(\omega)$ differs from the simple unimodal Einstein model. This deficiency highlights the need to improve on Allen–Dynes to investigate the high-pressure, high-temperature hydrides of great current interest.

In a previous paper, we used an analytical machine learning approach to try to improve on the Allen–Dynes formula, testing and training on tiny databases from the Allen–Dynes table of 29 superconducting materials. This proof of principle work showed that the SISSO framework, properly constrained by physical law, could substantially improve the performance of the Allen–Dynes equation with a smaller number of parameters. Clearly, it is necessary to apply this approach to a more extensive and diverse database.

Here, we proceed more systematically and show how we can teach the machine Eliashberg theory by generating large databases of $\alpha^2 F$ functions from both real materials and single- and multimodal artificial ones and learning the results of $T_c$ from solutions to the Eliashberg equations. We additionally include in our study $\alpha^2 F$ functions for superhydrides, extending training and testing to the higher $\lambda$ range. We show that the Allen–Dynes
equation fails in this region particularly badly, since it was designed to fit materials with the ratio of the Allen–Dynes parameters \( \omega_d / \omega_{\log} \approx 1 \), which is strongly violated in some of the higher-\( T_c \) materials. Here \( \lambda \) is the integral \( 2 \int_0^\infty \alpha^2 F(\omega) / \omega \, d\omega \), the frequencies \( \omega_n \) are the \( n \)th root of the \( n \)th moment of the normalized distribution \( g(\omega) = 2/(\lambda \omega) \alpha^2 F(\omega) \), and \( \omega_{\log} \equiv \exp \left( \int_0^\infty \omega \log(\omega) \, d\omega \right) \).

We begin by introducing the McMillan and Allen–Dynes equations, against which we will compare our results. McMillan\(^4\), in an attempt to improve on the BCS weak-coupling \( T_c \) theory, incorporated elements of Eliashberg theory\(^22\) into a phenomenological expression, relating \( T_c \) to physical parameters that could in principle be extracted from tunneling data\(^23\),

\[
T_c \simeq \frac{\omega_d}{1.45} \exp \left( \frac{1.04(1 + \lambda)}{\lambda - \mu^*(1 + 0.62\lambda)} \right),
\]

where \( \mu^* \) is the Coulomb pseudopotential and \( \omega_d \) is the Debye frequency. Note that the McMillan formula predicts a saturation of \( T_c \) in the strong-coupling limit, \( \lambda \to \infty \), for fixed \( \omega_d \).

Allen and Dynes\(^2\) showed that the true Eliashberg \( T_c \) did not obey such a bound in this limit but rather grew as \( \sqrt{\lambda} \). They proposed an alternate approximate fit to Eliashberg theory based on data on a few low-\( T_c \) superconductors known in 1975

\[
T_c = \frac{f_1 f_2 \omega_{\log}}{1.20} \exp \left( -\frac{1.04(1 + \lambda)}{\lambda - \mu^*(1 + 0.62\lambda)} \right),
\]

where \( f_1 \) and \( f_2 \) are factors depending on \( \lambda, \mu^*, \omega_{\log} \), and \( \omega_d \).

### RESULTS AND DISCUSSION

Figure 1 outlines our methods and computational workflow. We begin by collecting \( \alpha^2 F(\omega) \) spectral functions from ab initio calculations and augmenting the dataset with artificial spectral functions based on generated Gaussian functions. The Coulomb pseudopotential \( \mu^* \) is sampled as a free parameter and used, alongside the spectral functions, as an input to the Eliashberg equations. Eliashberg theory yields the superconducting gap function \( \Delta \), from which we extract \( T_c \). At the same time, we extract the quantities \( \lambda, \omega_{\log}, \) and \( \omega_d \) from \( \alpha^2 F \). Next, we use machine learning techniques to learn the relationship between the four model inputs, or features, and the critical temperature from Eliashberg theory \( T_c \) for \( \mu^* \). Finally, we compare the predictive models for \( T_c \) and discuss the feature-\( T_c \) relationships.

#### Computational details

We compile a set of 2874 electron–phonon spectral functions \( \alpha^2 F(\omega) \), summarized in Table 1. Of these, 13 are conventional phonon mediated superconductors, where we calculate \( \alpha^2 F \) using the electron–phonon Wannier package (EPW)\(^12,24\) of the Quantum Espresso (QE) code\(^25,26\). An additional 42 (29 classic and 13 hydride superconductors) are obtained from the computational superconductivity literature. We augment the dataset by generating 2819 artificial multimodal \( \alpha^2 F(\omega) \) functions and calculating the corresponding \( T_c \)’s with the EPW code. The superconducting transition temperatures are estimated by using both the Allen–Dynes equation and by solving the isotropic Eliashberg equations. The raw data are available upon request.

The artificially generated \( \alpha^2 F(\omega) \) consist of three Gaussian peaks with randomly selected peak location and height

\[
\alpha^2 F(\omega) = \sum_{i=1}^3 \lambda_i \omega \exp\left( -\frac{\omega^2 - \omega_i^2}{2(\omega_d)^2} \right),
\]

where \( G(\omega) \) is a normalized Gaussian with width of 1/8 of the peak frequency \( \omega_d \).

The total \( \lambda \) is then equal to the sum of the \( \lambda_i \) which simplifies sampling of the space of spectral functions. The artificial trinodal \( \alpha^2 F \) spectral functions resemble those of many realistic materials, see Fig. 2 for the example of LaAl. The Allen–Dynes and Eliashberg \( T_c \) for the hydrides are obtained from published work (see refs. in Table 1).

To ensure efficient sampling of the input spaces, we select values of \( \lambda \) and \( \mu^* \) with pseudorandom Sobol sequences. As shown in Fig. 3, our uniform sampling scheme results in a set of artificially generated \( \alpha^2 F \) corresponding to an approximately uniform distribution of \( T_c \). Next, we removed artificial entries with \( T_c > 400 \) K to better reflect the distribution of realistic materials. While the histogram of \( \mu^* \) remains approximately uniform after this truncation, the histograms of \( \lambda, \omega_{\log}, \) and \( \omega_d \) become skewed towards lower values.

#### Data

In the Allen–Dynes formula, the “arbitrarily chosen” shape-dependent factor \( f_2 \) is based on the numerical solutions using the spectral functions of Hg, Pb, and the Einstein model\(^2\). Because the

---

**Table 1.** Summary of the datasets used for training and validation of the machine learning model.

| Name           | Entries\(^b\) | Training | Validation |
|----------------|---------------|----------|------------|
| Calculated\(^b\) | 13/30         | Y        |            |
| Gaussian\(^b\)  | 2819/-        | Y        |            |
| Literature\(^c\) | 29/149       | Y        |            |
| Hydrides\(^c\)  | 13/19         | Y        |            |

\(^a\)Unique/resampled with varying \( \mu^* \).

\(^b\)This work.

\(^c\)Published papers\(^15–78\).

---

**Fig. 1** Workflow for identifying new machine learning models for \( T_c \) from \( \alpha^2 F(\omega) \) spectra and derived quantities. The workflow is organized into four computational modules: data collection, preprocessing, machine learning, and application.
So far, we discussed the differences between the Allen–Dynes and Eliashberg spectral functions. We name the two learned prefactors a posteriori based on their functional forms and the mechanisms by which they reduce the systematic underprediction of $T_c$ at higher temperatures. Like the Allen–Dynes prefactor $f_\omega$, $f_\nu$ includes the ratio $\omega_\log/\omega_2$, modifying the prediction based on the shape of $\alpha^2F(\omega)$. Moreover, $f_\omega$ also scales with $\sqrt{\lambda}$, like the Allen–Dynes prefactor $f_\omega$. This is in agreement with the correct large-$\lambda$ behavior of Eliashberg theory, unlike our earlier work and the modified $T_c$ equation with linear correction proposed recently by Shipley et al. The manifestation of both behaviors in $\omega$ even for some simple cases. The root mean square error in the Allen–Dynes paper is 5.6%. When the ratio $\omega_\log/\omega_2$ is 1, the shape of $\alpha^2F$ is that of the unimodal Einstein model and the number of $\alpha^2F(\omega)$ shapes is small, it is expected that the Allen–Dynes $T_c(T_{\text{AD}})$ would have significant errors in some instances. Figure 4 illustrates such deviations for bimodal Gaussian spectral functions. So far, we discussed the $\alpha^2F(\omega)$ shapes in an abstract sense because there is no single parameter that uniquely determines their shape. Allen and Dynes proposed using the ratio $\omega_\log/\omega_2$ as an indicator of the shape of $\alpha^2F(\omega)$. In Fig. 4, the ratio $T_c(T_{\text{AD}})/T_c$ is plotted against $\omega_\log/\omega_2$ for $\lambda=0.6, 1, 2, 3$ and 4. The results demonstrate that there can be significant differences between the Allen–Dynes $T_{\text{AD}}$ and Eliashberg $T_c$ even for some simple cases. The root mean square error in the Allen–Dynes paper is 5.6%. When the ratio $\omega_\log/\omega_2$ is 1, the shape of $\alpha^2F$ is that of the unimodal Einstein model and the data are obtained from refs. 35–37. For the bimodal spectral functions, we select $\lambda_1=\lambda_2$ for simplicity and vary the total $\lambda$ from 1 to 4.

Allen–Dynes $T_c$ accurately predict the Eliashberg $T_c$ regardless of the coupling strength. When the ratio $\omega_\log/\omega_2$ decreases, i.e. the shape of $\alpha^2F$ has more structure; whether the Allen–Dynes formula can then still reasonably predict the Eliashberg $T_c$ depends on the electron–phonon coupling strength.

In this work, we train and test machine-learning models using the datasets listed in Table 1. Two sizes are reported for each non-Gaussian dataset, indicating the number of unique materials compared to the total number of datapoints. We sample $\mu'$ between [0.1, 0.16] which covers a wide range of possible $\mu'$ values. The calculated, artificial Gaussian, and literature-derived $\alpha^2F$ datasets are used for training all machine learning models. We left the hydride materials out of the training in order to validate the extrapolative capacity of each model.

Correction factors for $T_c$ from symbolic regression

As in our previous symbolic regression effort, we use the SISIO framework to generate millions of candidate expressions by recursively combining the input variables with mathematical operators such as addition and exponentiation. We performed symbolic regression twice, sequentially, to obtain two dimensionless prefactors of the McMillan exponential, yielding a machine learned critical temperature

$$T_c^\text{ML} = f_{\omega\mu} \frac{\omega_\log}{\omega_2} \exp \left( \frac{1.04(1+\lambda)}{\lambda - \mu'(1+0.62\lambda)} \right).$$

(6)

We name the two learned prefactors a posteriori based on their functional forms and the mechanisms by which they reduce the error in predicting $T_c$. The first factor

$$f_{\omega} = 1.92 \left( \frac{\lambda + \omega_\log/\omega_2 - \sqrt{\lambda}}{\lambda \exp(\omega_\log/\omega_2)} \right) - 0.08$$

(7)

is obtained from the fit to the ratio $T_c^\text{ML}/T_c^{\text{McMillan}}$ and eliminates the systematic underprediction of $T_c$ at higher temperatures. Like the Allen–Dynes prefactor $f_\omega$, $f_\omega$ includes the ratio $\omega_\log/\omega_2$, modifying the prediction based on the shape of $\alpha^2F(\omega)$. Moreover, $f_\omega$ also scales with $\sqrt{\lambda}$, like the Allen–Dynes prefactor $f_\omega$. This is in agreement with the correct large-$\lambda$ behavior of Eliashberg theory, unlike our earlier work and the modified $T_c$ equation with linear correction proposed recently by Shipley et al. The manifestation of both behaviors in $f_\omega$ gives credence to our symbolic regression approach because it incorporates the primary effects of the Allen–Dynes equation with fewer parameters. Applying the
correction $T_c = f_{\omega T_c}^{\text{McMillan}}$ achieves a percent RMSE of 15.2% across the materials (non-Gaussian model) data, compared to 48.6% when using the Allen–Dynes equation.

The second correction factor

$$f_\mu = \frac{6.86 \exp\left(\frac{\mu}{\mu - \mu_0}\right)}{1 - \mu - \frac{\mu_0}{\mu}} + 1$$

is obtained from the fit to the ratio $T_c^M(f_{\omega T_c}^{\text{McMillan}})$, effectively correcting the residual error from the fit of $f_\omega$ and thus cannot be used independently. Applying the correction $T_c = f_{\omega f_\mu T_c}^{\text{McMillan}}$ achieves a percent RMSE of 15.1% across the materials datasets, compared to 15.2% when using $f_\omega$ alone. The influence of $f_\mu$ is more apparent when examining clusters of points corresponding to resampled $\mu$ values for a single material, where the systematic error in $T_c^{\text{ML}}/T_c^M$ is reduced.

Note that $f_\mu \rightarrow 1$ in both of the limits $\lambda \rightarrow 0$ and $\lambda \rightarrow \infty$, and in fact does not vary by more than $\sim$10% from 1 over the data set.

Figure 5 shows that, apart from the low-$T_c$ non-hydride materials for which the difference is smaller than 0.1 K, the corrections $f_\omega$ and $f_\mu$ dramatically improve predictions compared to using the Allen–Dynes equation. Since we excluded the hydrides from the training, these results successfully validate our data-driven symbolic regression approach by demonstrating the extrapolative capacity of the learned equations.

To further quantify the similarity between the existing Allen–Dynes prefactors and the machine-learned prefactors, we employ two statistical measures, the Spearman and distance correlation. The Spearman correlation is a measure of monotonicity in the relationship between rankings of two variables. Like the Pearson correlation coefficient for linear correlation, the Spearman correlation varies between $-1$ and $1$, where extrema imply high correlation and zero implies no correlation. Unlike the Pearson correlation, the Spearman correlation does not assume normally distributed datasets. By construction, all four prefactors tend to unify for many materials, resulting in asymmetric distributions that are unsuitable for analysis with parametric measures like the Pearson correlation.

In addition to the Spearman correlation, we compute the distance correlation, another nonparametric measure of the dependence between two variables. The distance correlation is defined as the ratio of the distance covariance and the product of the distance standard deviations, where distance covariance is the weighted Euclidean distance between the joint characteristic function of the two variables and the product of their marginal characteristic functions. Unlike the Pearson and Spearman correlation coefficients, the distance correlation varies between 0 and 1, where 0 indicates that the variables are independent, measuring both linear and nonlinear association.

Table 2 shows a strong relationship between $f_1$, $f_2$, and $f_\omega$ according to both Spearman and distance correlation metrics, with values close to one. This numerical analysis reinforces the conclusion that $f_\omega$ reproduces characteristics of both $f_1$ and $f_2$, as illustrated earlier in the comparison of functional forms. On the other hand, both Spearman correlation and distance correlation measures indicate slightly weaker relationships between $f_\mu$ and the other three prefactors. The relative independence of $f_\mu$ compared to $f_\omega$, $f_1$, and $f_2$ stems from the sequential nature of the fitting process.

Comparing predictive models for $T_c$

To compare existing equations for $T_c$ with the corrections identified in this work, we benchmarked the RMSE across non-hydride materials, artificial Gaussians, and hydrides as tabulated in Table 3. Additionally, we compute the %RMSE by normalizing each RMSE by the mean value across the corresponding dataset.

| Prefactors | Correlation | Distance |
|------------|-------------|----------|
| $f_1$      | $f_2$       | 0.932    | 0.964    |
| $f_1$      | $f_\omega$  | 0.943    | 0.973    |
| $f_1$      | $f_\mu$     | $-0.870$ | 0.693    |
| $f_1$      | $f_2$       | 0.850    | 0.965    |
| $f_1$      | $f_\omega$  | $-0.861$ | 0.662    |
| $f_1$      | $f_\mu$     | $-0.887$ | 0.731    |

The Spearman correlation measures the rank correlation between two variables while the distance correlation measures the dependence between two variables, including both linear and nonlinear association. Values closer to zero indicate weaker relationships. While $f_1$, $f_\omega$, and the new prefactor $f_\mu$ are strongly correlated to one another, $f_\mu$ shows a weaker relationship. The relative independence of $f_\mu$ reflects its origin as a second correction fit to the residual error of $f_\omega$. 

---

**Table 2. Correlations between Allen–Dynes and machine-learned prefactors.**
To assess the behavior of each model with increasing \( \lambda \), we plot \( T_c / \omega_{\log} \) for each model in Fig. 6.

As expected, the Allen–Dynes equation improves on the McMillan equation across all three groups. On the other hand, the equation identified by Xie et al. \(^{21}\) in an earlier symbolic regression work performs slightly worse on the low-\( T_c \) non-hydride dataset but achieves lower RMSE across the artificial Gaussian and hydride materials despite being trained on a small set of 29 low-\( T_c \) materials.

Applying the new \( f_\omega \) prefactor to the McMillan equation reduces \%RMSE in non-hydride materials from 14.4% to 8.4%, in artificial Gaussian models from 45.1% to 9.2%, and in hydrides from 36.6% to 5.8%. Moreover, applying both \( f_\omega \) and \( f_\lambda \) results in a further, modest improvement to the RMSE. In Fig. 6, our machine-learned correction (blue) is nearly equal to the Allen–Dynes equation (gray) for values of \( \lambda \) up to 1 but rapidly increases at larger \( \lambda \). Both bounds, for higher and lower values of \( \omega_{\log} / \omega_{c,\log} \), exceed the bounded region of the Allen–Dynes equation, indicating that at least part of the new model’s success is due to an improvement in capturing the behavior of \( T_c \) with increasing \( \lambda \).

We additionally fit a random forest (RF) model and an artificial neural network (ANN) model using the same training data to compare against our symbolic regression method. Hyperparameters for RF and ANN models were selected using 10-fold leave-cluster-out cross-validation and the same clusters identified for symbolic regression. On the other hand, the model error was estimated using nested cross-validation, where the inner loop was performed using a conventional 5-fold cross-validation scheme. Production models used in Fig. 6 were fit with the selected hyperparameters using the entire training set.

The RF is an ensemble model comprised of decision trees, each fit to random subsets of the data and queried to yield an independent prediction. Each decision tree uses a flow-chart-like series of decisions (branches) to yield predictions (leaves) and is optimized by varying decision thresholds. While individual decision trees are prone to overfitting, a RF produces robust predictions by averaging the predictions of its members. The optimized RF model, consisting of 100 decision trees with a maximum depth of eight splits per tree, achieved the lowest RMSE across all three models, with 4.7% RMSE in the testing set of hydride materials. This success may be attributed to both the flexibility of the method and the relative complexity compared to other methods. With up to 128 nodes per tree, the RF evaluates tens of thousands of binary decisions per prediction. On the other hand, as illustrated in Fig. 6, the resulting output (green) is discontinuous. Furthermore, the RF does not have the ability to extrapolate outside of regions of the input spaces included in the training data, resulting in constant-value outputs. This deficiency is evident in both upper- and lower-bound curves above \( \lambda = 3.8 \), where the RF correction results in a simple rescaling of the McMillan curve.

The ANN models in this work are feedforward neural networks, also known as multi-layer perceptrons, designed to learn highly non-linear function approximators to map multiple inputs to a target output. The feedforward architecture involves an input layer consisting of one neuron per input, one or more hidden layers, and an output layer consisting of one neuron per target. The value at each non-input neuron is a weighted, linear summation of the values in the preceding layer followed by a non-linear activation function. The optimized ANN includes three hidden layers with forty neurons each, totaling 3521 trainable parameters of multiplicative weights and additive biases. Despite the increased model complexity, the ANN performs similarly to the symbolic regression model, with slightly lower training RMSE and slightly higher testing RMSE. With increasing \( \lambda \), the ANN model yields similar values of \( T_c \) as indicated by the overlap between the shaded regions of the symbolic regression model (blue) and the ANN (yellow).

For low to moderate values of \( \lambda \), such as those originally studied by Allen and Dynes, all models behave similarly and the...
dimensionless corrections \( f_1, f_2, f_3, f_4 \) for ANN, RF are close to unity. However, as \( \lambda \) increases, the ANN, RF, and symbolic regression corrections deviate significantly from the Allen–Dynes equation as well as the previous symbolic regression equation. The corrections introduced in this work successfully correct the systematic underprediction of \( T_c \) with the symbolic regression solution offering simplicity and accuracy. Moreover, the monotonicity constraint in the symbolic regression search guarantees invertibility, allowing experimentals to extract \( \lambda \) from measured \( T_c \) and the electron–phonon spectral function. This characteristic is not guaranteed for the RF and ANN models.

Summary

The present work demonstrates the application of symbolic regression to a curated dataset of \( \alpha^2 F(\omega) \) spectral functions, yielding an improved analytical correction to the McMillan equation for the critical temperature of a superconductor. We showed that the well-known Allen–Dynes equation, an early improvement based on fitting to a very limited set of spectral functions, exhibits systematic error when predicting the Eliashberg compounds with density-functional theory using the QE code. We showed that the well-known Allen equation for the critical temperature of a superconductor. We used the optimized norm-conserving pseudopotential and the PBE version of the generalized gradient exchange-correlation functional. We sample the Brillouin zone for the electron orbitals using a \( 24 \times 24 \times 24 \) k-point mesh. To obtain the critical temperatures \( T_c \) of the compounds, we solve the isotropic Eliashberg equations with the EPW code.

Symbolic regression

We performed symbolic regression using the sure independence screening (SIS) and sparsifying operator (SISSO) framework, generating millions of candidate expressions. Based on memory constraints, the subspace of expressions was limited to those generated within four iterations. This limitation precludes the appearance of expressions of the complexity of the Allen–Dynes equation, motivating our search for a dimensionless correction to the McMillan equation rather than directly learning models for \( T_c \).

The initial quantities for generating expressions were the three dimensionless quantities \( \alpha, \lambda, \mu \), and the ratio \( \omega_\text{ph}/\omega_\text{ph} \). Candidates were generated using the set of operators \{ +, -, \times, \exp, \log, \sqrt{\cdot}, \sqrt[3]{\cdot}, \cdots \}. During the SIS step, these expressions were ranked based on their correlation to the ratio \( T^*_c/T^\text{McMillan}_c \), rather than \( T^*_c \) to identify dimensionless, multiplicative corrections to \( T^\text{McMillan}_c \).

To facilitate generalizability, we employ leave-cluster-out cross-validation during the generation of expressions using k-means-clustering with \( k = 10 \) on the combined set of 179 non-hyride and 2819 artificial-Gaussian entries. For each round of cross-validation, we generate candidate equations using a different subset of nine clusters and used the remaining cluster to evaluate performance using the root-mean-square error metric. As such, each training sample was left out of training and used for testing during one round. The top 10,000 models, ranked by root-mean-square error (RMSE) across the training set, were returned from each round. Models that did not appear in all ten rounds, corresponding to those with poor performance in one or more clusters, were eliminated.

Following the same principle, we ranked the remaining equations by the average RMSE across all ten rounds.

We note that the sparsifying operator (SO) step of the SISSO framework offers increased model complexity, as we explored in our previous work, but is limited in functional form to linear combinations of expressions generated from the preceding step. The linear combination of expressions from the initial subspace, by extension, also excludes equations as complex as the Allen–Dynes correction. Therefore, we did not consider linear combinations of expressions, meaning the SO simply selected the first-ranked expression from the SIS step in each run.

DATA AVAILABILITY

The database of critical temperatures, descriptors derived from the computed spectral functions \( \alpha^2 F(\omega) \), and symbolic regression logs are freely available at MaterialsWeb.org and https://www.materialscloud.org. Open Source Code

The symbolic regression workflow software we developed is freely available on Github (https://github.com/henniggroup/).

REFERENCES

1. Norman, M. R. Materials design for new superconductors. Rep. Prog. Phys. 79, 074502 (2016).
2. Boeri, L. et al. The 2021 Room-Temperature Superconductivity Roadmap. J. Condens. Matter Phys. (2021). https://doi.org/10.1088/1361-648X/ac2864.
3. Duan, D., Yu, H., Xie, H. & Cui, T. Ab Initio Approach and Its Impact on Superconductivity. J. Supercond. Nov. Mag. 32, 53 (2019).
4. McMillan, W. L. Transition temperature of strong-coupled superconductors. Phys. Rev. Lett. 167, 331–344 (1968).
5. Allen, P. B. & Dynes, R. C. Transition temperature of strong-coupled superconductors reanalyzed. Phys. Rev. B 12, 905–922 (1975).
6. Combescot, R. Critical temperature of superconductors: Exact solution from Eliashberg equations on the weak-coupling side. Phys. Rev. B 42, 7810–7824 (1990).
7. Marsiglio, F. Eliashberg theory in the weak-coupling limit. Phys. Rev. B 98, 024523 (2018).
68. Tütüncü, H. M. & Srivastava, G. P. Phonons and superconductivity in the cubic perovskite Cr$_7$RhN. *J. Appl. Phys.* **112**, 093914 (2012).
69. Yue, S.-Y., Cheng, L., Liao, B. & Hu, M. Electron–phonon interaction and superconductivity in the high-pressure c16 phase of lithium from first principles. *Phys. Chem. Chem. Phys.* **20**, 27125–27130 (2018).
70. Ortigoza, M. A. et al. Ab initio lattice dynamics and electron-phonon coupling of Bi(111). *Phys. Rev. B* **90**, 195438 (2014).
71. Cuamba, A. S., Lu, H.-Y. & Ting, C. S. Electronic structure and phonon-mediated superconductivity in SchiP compound: First-principles calculations. *Phys. Rev. B* **94**, 094513 (2016).
72. Chen, J. Phonons in bulk and monolayer HfS$_2$ and possibility of phonon-mediated superconductivity: A first-principles study. *Solid State Commun.* **237-238**, 14–18 (2016).
73. Tütüncü, H. M., Karaca, E. & Srivastava, G. P. Electron-phonon superconductivity in the filled skutterudites LaRu$_3$P$_{12}$, LaRu$_3$As$_{12}$, and LaPt$_3$Ge$_{12}$. *Phys. Rev. B* **95**, 214514 (2017).
74. Karaca, E., Tütüncü, H. M., Srivastava, G. P. & Uğur, S. Electron-phonon superconductivity in the ternary phosphides BaM$_2$P$_2$ (M = Ni, Rh, and Ir). *Phys. Rev. B* **94**, 054507 (2016).
75. Tütüncü, H. M., Karaca, E., Uzunok, H. Y. & Srivastava, G. P. Physical properties of hexagonal BaPtAs with noncentrosymmetric SrPbSb-type and centrosymmetric YPtAs-type crystal structures: Effects of spin-orbit coupling. *Phys. Rev. B* **100**, 174507 (2019).
76. Singh, S. & Kumar, R. Variation of superconducting transition temperature of YSn$_3$ under negative pressure. *J. Supercond. Nov. Magn.* **32**, 1157–1162 (2018).
77. Tütüncü, H. M., Karaca, E., Uzunok, H. Y. & Srivastava, G. P. Role of spin-orbit coupling in the physical properties of LaX$_3$ (X = In, P, Bi) superconductors. *Phys. Rev. B* **97**, 174512 (2018).
78. Ono, S. Low-temperature electron-phonon relaxation in Cu and Ag thin films. *Phys. Rev. B* **101**, 201404 (2020).

**ACKNOWLEDGEMENTS**

We are grateful to L. Boeri, P. Allen, and W. Pickett for valuable discussions. We thank D. Semenok for providing data related to Actinium hydrides. The work presented here was performed under the auspice of Basic Energy Sciences, United States Department of Energy, contract number DE-SC0020385. Partial funding was also provided by the University of Florida Informatics Institute.

**AUTHOR CONTRIBUTIONS**

All authors contributed extensively to the work presented in this paper. S.R.X., Y.Q., A.C.H., L.F., J.L., J.K., G.R.S., J.J.H., and R.G.H. conceived the overall methodology of data assembly, augmentation, Eliashberg calculations, and symbolic regression. Y.Q., B.D., J.M.D., I.S., and U.S.H. performed the literature search and collected the spectral function data for superconductors. S.R.X. and Y.Q. implemented the algorithm and performed the calculations and analysis. S.R.X., Y.Q., P.J.H., and R.G.H. contributed to the writing of the manuscript.

**COMPETING INTERESTS**

The authors declare no competing interests.

**ADDITIONAL INFORMATION**

Correspondence and requests for materials should be addressed to R. G. Hennig.

Reprints and permission information is available at [https://www.nature.com/reprints](https://www.nature.com/reprints)

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

© The Author(s) 2022