Machine-learning approach for discovery of conventional superconductors

Huan Tran[1,*] and Tuoc N. Vu[2]

[1] *School of Materials Science & Engineering, Georgia Institute of Technology, 771 Ferst Dr. NW, Atlanta, GA 30332, USA*
[2] *Institute of Engineering Physics, Hanoi University of Science & Technology, 1 Dai Co Viet Rd., Hanoi 10000, Vietnam*

First-principles computations are the driving force behind numerous discoveries of hydride-based superconductors, mostly at high pressures, during the last decade. Machine-learning (ML) approaches can further accelerate the future discoveries if their reliability can be improved. The main challenge of current ML approaches, typically aiming at predicting the critical temperature T_c of a solid from its chemical composition and target pressure, is that the correlations to be learned are deeply hidden, indirect, and uncertain. In this work, we showed that predicting superconductivity at any pressure from the atomic structure is sustainable and reliable. For a demonstration, we curated a diverse dataset of 584 atomic structures for which λ and $\omega_{\log}$, two parameters of the electron-phonon interactions, were computed. We then trained some ML models to predict λ and $\omega_{\log}$, from which T_c can be computed in a post-processing manner. The models were validated and used to identify two possible superconductors whose $T_c \simeq 10-15$K at zero pressure. Interestingly, these materials have been synthesized and studied in some other contexts. In summary, the proposed ML approach enables a pathway to directly transfer what can be learned from the high-pressure atomic-level details that correlate with high-T_c superconductivity to zero pressure. Going forward, this strategy will be improved to better contribute to the discoveries of new superconductors.

I. INTRODUCTION

In the search for high critical temperature (T_c) superconductors, significant progress has been made during the last decade [1–3]. Among thousands of hydride-based superconducting materials computationally predicted [4–12], mostly at very high pressures, e.g., $P \gtrsim 100$ GPa, dozens of them, e.g., H_3S [1], LaH_{10} [2], and CSH [3], were synthesized and tested. This active research area is presumably motivated by Ashcroft, who, in 2004, predicted [13] that high-T_c superconductivity may be found in hydrogen dominant metallic alloys, probably at high P. Another driving force is the development of first-principles computational methods to predict material structures at any P [14–20] and to calculate the electron-phonon (EP) interactions [21, 22], the atomic mechanism behind the conventional superconductivity, according to the Bardeen-Cooper-Schrieffer (BCS) theory [23]. While critical debates on some discoveries [24–29] are on-going, it seems that the one-day-realized dream of superconductors at ambient conditions may be possible. Readers are referred to some reviews [5, 6, 8, 30] and a recent roadmap [9] for progresses, challenges, and future pathways of this research area.

The central role of first-principles computations in the recent discoveries of conventional superconductors stems from Éliashberg theory [31–34], of which the spectral function $\alpha^2F(\omega)$ characterizing the EP interactions could be evaluated numerically. The first inverse moment λ and logarithmic moment $\omega_{\log}$ of $\alpha^2F(\omega)$, together with an empirical Coulomb pseudopotential μ^*, are the inputs to estimate T_c by either solving the Éliashberg equations [31–34] or using the McMillan formula [35–37] (see Sec. II A for more details). In a typical workflow (Fig. 1), a search for stable atomic structures across multiple related chemical compositions is performed at a given pressure, usually with first-principles computations. Then, $\alpha^2F(\omega)$, λ, $\omega_{\log}$, and finally T_c are evaluated, identifying candidates with high estimated T_c for possible new superconducting materials. Although structure prediction [14–17] and $\alpha^2F(\omega)$ computations [21, 22] are extremely expensive and technically non-trivial, significant research efforts have been devoted to and shaped by this workflow.

Machine-learning (ML) methods have recently emerged in the discoveries of superconductors [9, 10]. As sketched in Fig. 1, existing ML efforts can be categorized into four lines, including (i) using some ML potentials to accelerate the structure prediction step [38], (ii) using some symbolic ML techniques to derive new empirical expressions for T_c [39, 40], (iii) developing some ML models to predict T_c from a chemical composition at a given pressure P [41–50], and (iv) developing some ML models to predict λ, $\omega_{\log}$, and $\alpha^2F(\omega)$ from the atomic structures [51]. While line (iii) is predominant, its role remains limited, presumably because the connections

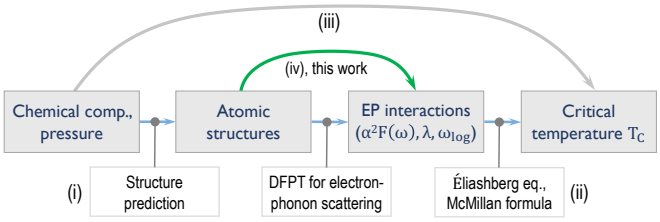

FIG. 1. A typical workflow to compute T_c. Existing ML efforts are devoted to (i) using ML potentials to accelerate the structure prediction step, (ii) deriving new formulas of T_c, and (iii) predicting T_c from chemical composition. This work is in (iv), predicting λ and $\omega_{\log}$ from atomic structures.

* huan.tran@mse.gatech.edu

from the chemical composition and the target P to T_c are deeply hidden. In fact, there are at least two "missing links" between the two ends of this approach. One of them is the atomic-level information while the other is the microscopic mechanism of the superconductivity, e.g., the EP interactions in conventional superconductors. The former is critical because for a given chemical composition, the properties of thermodynamically competing atomic structures can often be fundamentally different, e.g., one is insulating and another is conducting [18, 52]. Therefore, ignoring the atomic structure is equivalent to adding an irreducible uncertainty into the ML predictions [53]. Likewise, the latter cannot be overemphasized. In fact, bypassing $\alpha^2F(\omega)$, λ, and $\omega_{\log}$, and using an empirical value of μ^* are intractable assumptions, and thus, uncontrollable approximations. In line (iv), initialized recently by Ref. 51 during the (independent) preparation of this work, these missing links are addressed in some ways.

In this paper, we present an initial step to bring the atomic-level information into the ML-driven pathways toward new conventional (or BCS) superconductors, especially at ambient pressure. For this goal, we curated a dataset of 584 atomic structures for which more than 1,100 values of λ and $\omega_{\log}$ were computed at different values of P and reported, mostly in the last decade. The obtained dataset was visualized, validated, and standardized before being used to develop ML models for λ and $\omega_{\log}$. Then, they were used to screen over $80,000$ entries of Materials Project database [54], identifying and confirming (by first-principles computations) two thermodynamically and dynamically stable materials whose superconductivity may exist at $T_c \simeq 10-15$K and $P = 0$. We also proposed a procedure to compute λ and $\omega_{\log}$, for which convergence are generally hard to attain [51].

This scheme relies on the direct connection between the atomic structures and λ and $\omega_{\log}$, quantitatively described in Sec. II A. Pressure is an *implicit* input, i.e., P determines the atomic structures for which λ and $\omega_{\log}$ are computed/predicted. The design of this scheme has some implications. First, the ML models are trained on the atomic structures realized at high P and (computationally) proved to correlate with high-T_c superconductivity. These structures can be considered "unusual" in the sense that their high-P atomic-level details, e.g., short bond lengths and distorted bond angles, are not usually realized at zero pressure. Therefore, we hope that the ML models can identify the atomic structures realized at $P = 0$ with relevant unusual atomic-level features, and thus, they may exhibit possible high-T_c superconductivity. Second, massive material databases [55] like Materials Project [54], OQMD [56] and NOMAD with millions of atomic structures can now be screened directly with robust and reliable ML models. Given that only a small search space was explored in this demonstrative work, we expect more superconducting materials to be discovered in the next steps of our effort.

II. METHODS

A. Éliashberg theory and McMillan formula

In Éliashberg theory [31–34], $\alpha^2F(\omega)$ is a spectral function characterizing the EP scattering, which is defined as

$$\alpha^2F(\omega) = \frac{1}{N_0}\sum_{\mathbf{k}\mathbf{k}'ij\nu} |g^{ij,\nu}_{\mathbf{k},\mathbf{k}'}|^2 \delta(\varepsilon^i_{\mathbf{k}})\delta(\varepsilon^j_{\mathbf{k}'})\delta(\omega-\omega^\nu_{\mathbf{k}-\mathbf{k}'}). \quad (1)$$

Here, N_0 is the density of states at the Fermi level, $g^{ij,\nu}_{\mathbf{k},\mathbf{k}'}$ the electron-phonon matrix elements, ν the polarization index of the phonon with frequency ω, δ the delta-Dirac function, and ($\mathbf{k}$ and $\mathbf{k}'$)/($\varepsilon^i_{\mathbf{k}}$ and $\varepsilon^j_{\mathbf{k}'}$) the (electron wave vectors)/(band energies) corresponding to the band indices (i and j), respectively.

The standard method to compute $\alpha^2F(\omega)$ is density functional perturbation theory (DFPT) [21, 22], as implemented in major codes like Quantum ESPRESSO [57, 58] and ABINIT [59–61]. Having $\alpha^2F(\omega)$, T_c can be evaluated by numerically solving a set of (unfortunately, quite complicated) Éliashberg equations using, for example, Electron-Phonon Wannier (EPW) [62–64]. The much more frequent method to estimate T_c is using some empirical formulas derived from the Éliashberg equations. Perhaps the most extensively used formula is

$$T_c = \frac{\omega_{\log}}{1.2}\exp\left[-\frac{1.04(1+\lambda)}{\lambda-\mu^*(1+0.62\lambda)}\right], \quad (2)$$

which was developed by McMillan [35] and latter improved by Allen and Dynes [36, 37]. Here,

$$\lambda = 2\int_0^\infty d\omega \frac{\alpha^2F(\omega)}{\omega} \quad (3)$$

is the (averaged) isotropic EP coupling while

$$\omega_{\log} = \exp\left[\frac{2}{\lambda}\int_0^\infty d\omega \ln(\omega)\frac{\alpha^2F(\omega)}{\omega}\right]. \quad (4)$$

Following Ashcroft [13], the Coulomb pseudopotential μ^*, which appears in Eq. 2 and connects with N_0, was empirically chosen in the range between 0.10 and 0.15. Eq. (2) indicates that in general, high values of λ and/or $\omega_{\log}$ are needed for a high value of T_c. Some new empirical formulas of T_c were developed recently [39, 40] using some symbolic ML techniques. Moving forward, developing a truly *ab-initio* framework for computing T_c [65–67] is desirable and currently active.

The McMillan formula (2) is believed to be good for $\lambda \leq 1.5$ while additional empirical parameters are needed for larger λ [37]. Nevertheless, the exponential factor of Eq. 2 has a singular point at $\lambda = \mu^*/(1-0.62\mu^*)$, which could lead to unwanted/unphysical divergence. If we select $\mu^* = 0.1$ (or 0.15), $T_c \to \infty$ when λ approaches 0.1066 (or 0.1654) from below. Such values of λ have been realized in many computational works [68–71], although much larger values, e.g., $\lambda \geq 0.7$, are generally

needed for high-T_c superconductors. Given these observations, we believe that a ML approach for discovering conventional superconductor should focus on λ, $\omega_{\log}$, and perhaps $\alpha^2F(\omega)$, from which T_c can easily be estimated using, for examples, Eq. (2).

B. Basic idea and approach

The ML approach used in this work focuses on predicting λ and $\omega_{\log}$ from the atomic structure of the considered materials. As visualized in Fig. 1, the role of P is embedded in the main input of this scheme, i.e., the atomic structure, which is determined from P. The rationale of this design is two fold. First, what this ML approach will learn is a *direct and physics-inspired* correlation from an atomic structure to λ and $\omega_{\log}$ through $\alpha^2F(\omega)$, as quantitatively described in Eqs. (1), (3), and (4). Second, the training data, which include the atomic environments/structures realized at multiple values of (sometimes very high) pressure P that could lead to very high values of λ and $\omega_{\log}$, will be highly diverse and comprehensive. Consequently, the resulted ML models will thus be robust, reliable, and, more importantly, they can be used to recognize new high-T_c superconductors that resemble unusual atomic-level details *at any pressure*, specifically $P = 0$ GPa. This approach involves some challenges, one of them is how to obtain good datasets for the learning scheme. Our solution is described below.

C. Data curation

This work requires a dataset of the atomic structures for which λ, $\omega_{\log}$, and $\alpha^2F(\omega)$ were computed and reported. The curation of such a dataset is painstaking. Scientific articles published during the last $10-15$ years, reporting computed superconducting properties of new or known materials, were collected. In majority of the articles, the atomic structures were reported in some Tables while electronic files of standard formats, e.g., crystallographic information file (CIF), were given in very few cases. In some cases, important information, e.g., angle β in a monoclinic structure, was missing from the Tables. When the provided information is sufficient, we used the obtained crystal symmetry/space group, lattice parameters, Wyckoff positions, and the coordinates of the inequivalent atoms, to manually reconstruct the reported structures. All the atomic structures obtained from electronic files and/or reconstructed from data Tables were inspected visually. During this step, a good number of them were found to be clearly incorrect, largely because of typos, number overrounding, and other possible unidentified reasons, when reporting the data. Incorrect structures were discarded.

Superconducting-related properties, e.g., λ, $\omega_{\log}$, and T_c, which were computed and reported for the atomic structures at pressure P up to 800 GPa, were collected. These properties were mainly computed by some major workhorses like Quantum ESPRESSO [57, 58], ABINIT [59–61], and EPW codes [62–64], employing different pseudopotentials, XC functionals, energy cutoffs, smearing width needed to compute the δ functions appearing in the expression (1) of $\alpha^2F(\omega)$, and more. We recognize that data of λ and $\omega_{\log}$ curated from scientific literature are not entirely uniform; they rather contain a certain level of uncertainty that will inevitably be translated into the (aleatoric) uncertainty of the predictions [53]. However, the demonstrated reproducibility of advanced first-principles computations [72] suggests that data carefully produced by major codes should still be consistent and reliable.

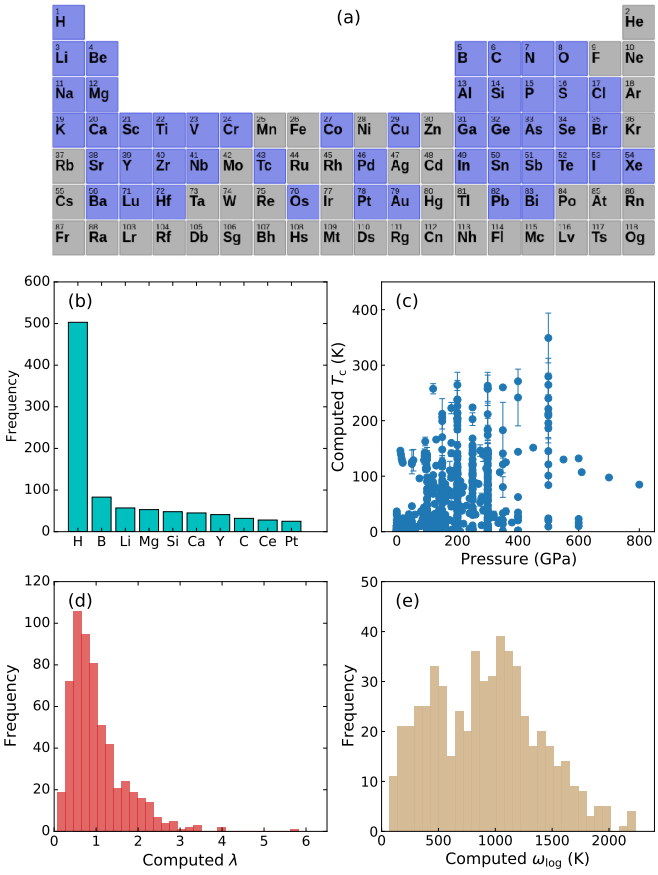

FIG. 2. A summary of computed λ, $\omega_{\log}$, and T_c dataset of 584 superconducting materials reported and curated, including (a) the Periodic Table coverage, (b) 10 most-frequent species in the dataset, (c) 567 values of T_c computed and arranged at different pressures, and the distribution of (d) 584 computed values of λ, and (e) 567 values of $\omega_{\log}$, are given. Among 53 species found in our dataset, 47 of them are shown in (a) and the other 6 species are Ac, Ce, La, Nd, Pm, and Pr. In (b), each solid circle represents a combination of a chemical composition and a pressure while errorbars are for cases T_c was computed for different atomic structures, using different methods, e.g., using McMillan formula and solving Éliashberg equations, and/or different values of μ^*.

To further improve the uniformity of the data, we used density functional theory (DFT) [73, 74] calculations to optimize the obtained atomic structures at the pressures reported, employing the same technical details used for Materials Project database. The rationale behind this step is that the predictive ML models trained on the dataset will then be used to predict λ and $\omega_{\log}$ for the atomic structures obtained from Materials Project database. Therefore, the training data should be prepared at the same level of computations with the input data for predictions. In fact, a vast majority of our DFT optimizations were terminated after about a dozen steps or below, indicating that they were already optimized very well. Details of the optimizations are given in Sec. II E. Compared with the DFPT calculations for λ and $\omega_{\log}$, the optimization step is computationally negligible.

Our dataset includes 584 atomic structures for which at least λ was computed and reported. Among them, 567 atomic structures underwent $\omega_{\log}$ and thus, T_c calculations (there is a trend in the community that computed λ is more likely to be reported than $\omega_{\log}$ when discussing the superconductivity). Our dataset, which is summarized in Figs. 2 (a), (b), (c), (d), and (e), contains 53 species and covers a substantial part of the Periodic Table. Five most frequently encountered species are H, B, Li, Mg, and Si, which were found in 505, 83, 57, 53, and 48 entries, respectively. The dominance of H in this dataset reflects the focus of the community on super hydrides when searching for high-T_c superconductors. For λ, the smallest value is 0.089, reported in Ref. 68 for the $P4/mbm$ structure of LiH_2 at $P = 150$ GPa while the largest value is 5.81 reported in Ref. 44 for the $Im\bar{3}m$ structure of CaH_6 at $P = 100$ GPa. Likewise, the smallest value of $\omega_{\log}$ is 71 K reported in Ref. 75 for the $I4/mmm$ structure of TiH at $P = 50$ GPa whereas the largest value is 2,234 K reported in Ref. 44 for the $P\bar{6}2m$ structure of CaH_{15} at $P = 500$ GPa. Figs. 2 (d) and (e) provide two histograms summarizing the λ and $\omega_{\log}$ datasets.

D. Data representation and machine-learning approaches

Materials atomic structures are not naturally ready for ML algorithms. The main reason is that they are not invariant with respect to transformations that do not change the materials in any physical and/or chemical ways, e.g., translations, rotations, and permutations of alike atoms. Therefore, we used MATMINER [76], a package that offers a rich variety of material features, to convert (or featurize) the atomic structures into numerical vectors, which meet the requirements of invariance and can be used to train ML models. Starting from several hundreds components, optimal sets of features (the vector components) were determined using the recursive feature elimination algorithm as implemented in SCIKIT-LEARN library [77]. The final version of the λ and $\omega_{\log}$ datasets have 40 and 38 features, respectively.

In principle, two featurized datasets of λ and $\omega_{\log}$ can be learned simultaneously using a multi-task learning scheme so that the underlying correlations between λ and $\omega_{\log}$ may be exploited. However, the intrinsically deep correlations in materials properties require a sufficiently big volume of data to be revealed. We have tested some multi-task learning schemes and found that with a few hundreads data points, they are not significantly better than learning λ and $\omega_{\log}$ separately. In fact, similar behaviors are commonly observed in the literature [53]. Therefore, we examined six typical ML algorithms, including support vector regression, random forest regression, kernel ridge regression, Gaussian process regression, gradient boosting regression, and artificial neural networks two develop ML models for λ and $\omega_{\log}$. For each algorithm, we created a pair of learning curves and used them to analyze the performance of the algorithm on the data we have. By carefully tuning the possible model parameters and examining the training and the validation curves, Gaussian process regression (GPR) [78, 79] was selected. Details on the learning curves and the GPR models used for predicting λ and $\omega_{\log}$ are discussed in Sec. III A.

E. First-principles calculations

First-principles calculations are needed for two purposes, i.e., to uniformly optimize the curated atomic structures and to compute $\alpha^2F(\omega)$, λ, and $\omega_{\log}$ for those identified by the ML models we developed. For the first objective, we followed the technical details used for Materials Project database, employing VASP code [80, 81], the standard PAW pseudopotentials, a basis set of plane waves with kinetic energy up to 520 eV, and the generalized gradient approximation Perdew-Burke-Ernzerhof (PBE) exchange-correlation (XC) functional. [82] Convergence in optimizing the structures was assumed when the atomic forces become $< 10^{-2}$ eV/Å after no more than 3 iterations.

In the computations of $\alpha^2F(\omega)$, λ, and $\omega_{\log}$, we used the version of DFPT implemented in ABINIT package [59–61], which also offers a rich variety of other DFT-based functionalities. Within this numerical scheme, we used the optimized norm-conserving Vanderbilt pseudopotentials (ONCVPSP-PBE-PDv0.4) [83] obtained from the PseudoDojo library [84] and the PBE XC functional [82]. The kinetic energy cutoff we used is 60 Hatree ($\simeq 1,600$eV), which is twice larger than the value suggested [83] for these norm-conserving pseudopotentials. The smearing width for computing $\alpha^2F(\omega)$ is 5×10^{-6} Ha, i.e., $\simeq 0.032$ THz. This value was selected to be $< 0.1\%$ of the entire range of frequency while covering more than 4 (numerical) spacings of the frequency grid.

Before entering the electron-phonon calculations with DFPT, the material structures under consideration were

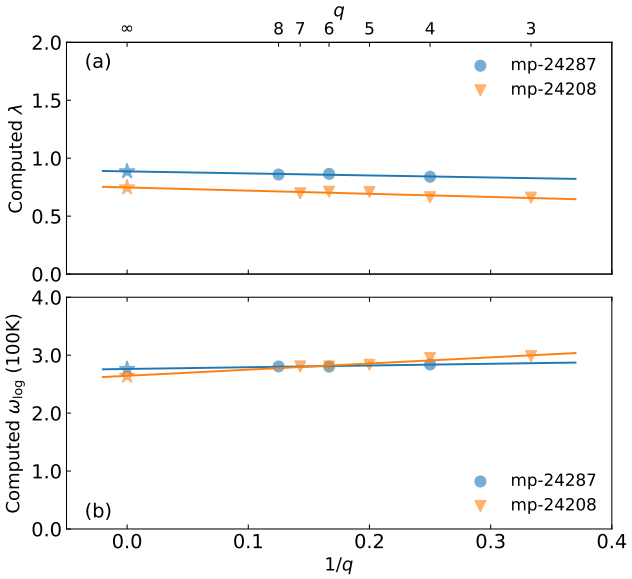

FIG. 3. Fitting procedure used to compute (a) λ and (b) $\omega_{\log}$ of mp-24287 and mp-24208, two atomic structures identified from Materials Project database. Solid symbols show λ and $\omega_{\log}$ computed with some finite **q**-point grids while stars represent the extrapolated values of λ and $\omega_{\log}$ at the limit of infinite **q**-point grid, i.e., $1/q = 0$.

repeatedly optimized until the maximum atomic force is below 10^{-5} Hatree/bohr, which is $\simeq 5.1 \times 10^{-4}$ eV/Å, after no more than 3 iterations. Because the optimizations need the simulation box to change its shape, such a small number of iterations is required to minimize the cell volume change, thereby limiting the Pulay stress, and ultimately ensuring an absolute convergence of the force calculations. This level of accuracy is generally needed for phonon-related calculations.

Eq. 1 indicates that $\alpha^2F(\omega)$ is evaluated on a **q**-point grid of $\mathbf{q} = \mathbf{k} - \mathbf{k}'$, which must be a sub-grid of the full **k**-point grid used to sample the Brillouin zone for regular DFT calculations. Therefore, calculations of $\alpha^2F(\omega)$ are extremely heavy while the convergence with respect to the **q**-point grid is critical and must be examined [44, 51]. For this goal, we first computed $\alpha^2F(\omega)$, λ, and $\omega_{\log}$ using several **q**-point grids of $q \times q \times q$ and **k**-point grids of $k \times k \times k$ where q is as large as possible depending on the structure size and $k \geq 3 \times q$. Then, the computed values of λ and $\omega_{\log}$ are fitted to a linear function of $1/q$. The values of the fitted functions at $1/q = 0$, or, equivalently, at the limit of $q \to \infty$, are the values assumed for λ and $\omega_{\log}$. This procedure is visualized in Fig. 3 when λ and $\omega_{\log}$ of two atomic structures reported in this work were computed. Details on the **q**-point and **k**-point grids and the corresponding computed data used for the fitting procedure can be found in Supplemental Material [85]. A technique of similar philosophy has been demonstrated [86] in the computations of ring-opening enthalpy, the thermodynamic quantity that controls the ring-opening polymerizations.

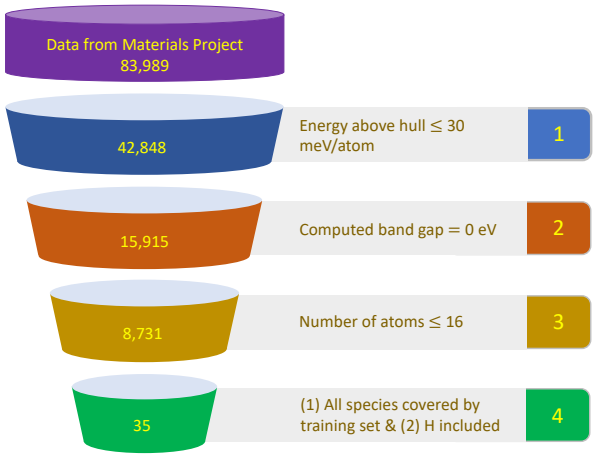

FIG. 4. Procedure to down select 35 atomic structures for predicting λ and $\omega_{\log}$ from 83,989 atomic structures of Materials Project database.

F. Candidates

We obtained the Materials Project database [54] of 83,989 atomic structures and several properties uniformly computed at $P = 0$ using VASP [80, 81]. Starting from this dataset, we selected a subset of 35 atomic structures that have *energy above hull* $E_{\text{hull}} < 0.03$ eV/atom, zero band gap ($E_g = 0$ eV), no more than 16 atoms in the primitive cell, and only the species included in the training data, specifically H (see Fig. 2). The first criterion "places" the selected atomic structures into the so-called "amorphous limit", a concept defined in an analysis of Materials Project database [87] and used to label the atomic structures that are (or nearly) thermodynamically stable and thus, they may be synthesized. In fact, some metastable ferroelectric phases of hafnia that are above the ground state of $\simeq 0.03$ eV/atom [18, 88, 89] have been stabilized and synthesized [90, 91]. Next, $E_g = 0$ eV was used to remove non-conducting materials while the third criterion aims at selecting small enough systems for which computations of λ and $\omega_{\log}$ are affordable. Finally, by considering only those having the species included in the training data, specifically H, we expect that the ML models will only be used in their domain of applicability. The procedure is summarized in Fig. 4.

The set of 35 candidates has no overlap with the training data. This set is small because the requirement of having H is very strong. In fact, removing this requirement increases the candidate set size to 2,694. Given that the ML models are extremely rapid, there is in fact no time difference between predicting λ and $\omega_{\log}$ for 35 atomic structures and predicting these properties for 2,694 atomic structures. However, the dominance of H in the training dataset strongly suggests that the smaller set of 35 candidates is more suitable for the demonstration purpose of this work. In the next step, the training dataset will be augmented with λ and $\omega_{\log}$ computed

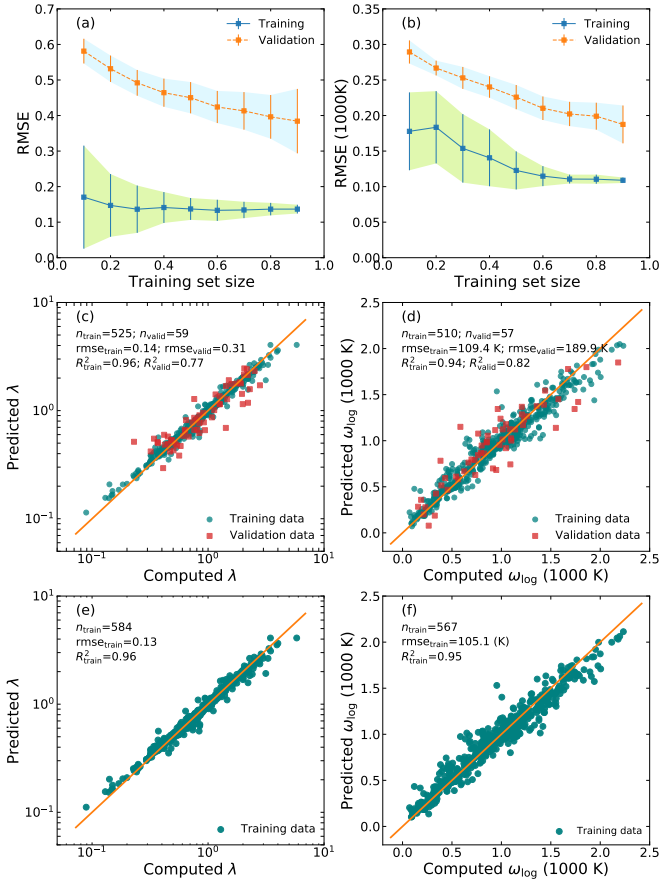

FIG. 5. Learning curves obtained by learning two datasets of (a) λ and (b) $\omega_{\log}$, a typical model trained on 90% of the data of (c) λ and (d) $\omega_{\log}$ and validated on the remaining *unseen* 10% data, and two ML models trained on 100% of the data of (e) λ and (f) $\omega_{\log}$. In (a) and (b), each data point is associated with an errorbar obtained from 100 models that were independently trained.

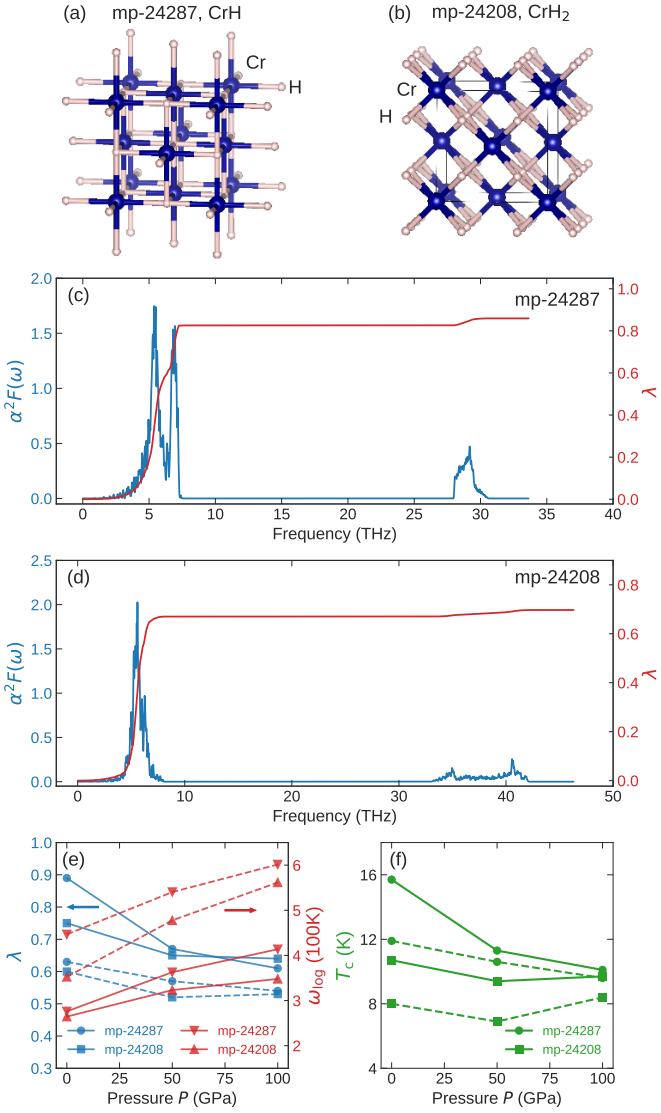

FIG. 6. Computed superconducting properties of mp-24287 and mp-24208, whose atomic structures are visualized in (a) and (b) and spectral function $\alpha^2F(\omega)$ and the accumulative $\lambda(\omega)$ are shown in (c) and (d). The **k**-point and **q**-point grids used for (c) are $24 \times 24 \times 24$ and $8 \times 8 \times 8$, respectively, while those used for (d) are $21 \times 21 \times 21$ and $7 \times 7 \times 7$, respectively. In (e) and (f), solid and dashed lines are used to show the computed (using the extrapolation procedure described in Sec. II E) and the predicted values λ, $\omega_{\log}$, and T_c (computed from λ and $\omega_{\log}$ using McMillan formula with $\mu^* = 0.1$) at $P = 0$, 50, and 100 GPa.

for materials having underrepresented species, and larger candidate sets will be examined.

III. RESULTS

A. Machine-learning models

Given a learning algorithm and a dataset that has been represented appropriately, learning curves can be created using an established procedure. In this work, each dataset was randomly split into a training set and a (holdout) validation set. Next, a ML model was trained on the training set using standard 5-fold cross-validation procedure [92] to regulate the potential overfitting. Then, the ML model was tested on the validation set, which is entirely unseen to the trained model. By repeating this procedure 100 times and varying the training set size, a training curve and a validation curve were produced from the mean and the standard deviation of the root-mean-square error (RMSE) of the predictions of the training sets and the validation sets. During the training/validating processes, randomness stems from the training/validation data splitting and the 5-fold training data splitting for internal cross validation. As such random fluctuations are suppressed statistically by averaging over 100 independent models, the learning curves could provide some useful and unbiased insights into the performance of the data, the featurize procedure,

TABLE I. Six hydrogen-containing materials that have highest predicted T_c among 35 materials in the candidate set, given in the top part. For each of them, the ID and the energy above hull E_{hull} obtained from Materials Project are given (the pressure P and computed band gap are all zero). Predicted λ, ω_{log}, and T_c were obtained from the ML models and computed using McMillan formula with $\mu^* = 0.1$. Among the 6 materials, computations were performed for 4 materials, two of them (mp-24287 and mp-24208) are dynamically stable, thus computed λ, ω_{log}, and T_c are available. In the bottom part of the Table, predicted and computed values of λ, ω_{log}, and T_c are reported for two dynamically stable materials, i.e., mp-24287 and mp-24208, at 50 GPa and 100 GPa.

MP ID	Chemical formula	Space group	P (GPa)	E_{hull} (eV/atom)	Predicted			Computation performed	Computed			
					λ	ω_{log} (K)	T_c (K)		Dyn. stable	λ	ω_{log} (K)	T_c (K)
mp-24289	PdH	$Fm\bar{3}m$	0	0.02	0.88	377.2	21.3	Yes	No	–	–	–
mp-1018133	LiHPd	$P4/mmm$	0	0	0.79	321.0	14.5	Yes	No	–	–	–
mp-24081	ScClH	$R\bar{3}m$	0	0	0.65	445.9	13.0	No	–	–	–	–
mp-24287	CrH	$Fm\bar{3}m$	0	0	0.63	446.6	11.9	Yes	Yes	0.89	276.2	15.7
mp-1008376	CeH_3	$Fm\bar{3}m$	0	0	0.60	418.5	9.5	No	–	–	–	–
mp-24208	CrH_2	$Fm\bar{3}m$	0	0	0.60	352.5	8.0	Yes	Yes	0.75	264.4	10.7
mp-24287	CrH	$Fm\bar{3}m$	50	–	0.57	540.3	10.6	Yes	Yes	0.67	362.8	11.3
	CrH	$Fm\bar{3}m$	100	–	0.54	601.4	9.6	Yes	Yes	0.61	413.7	10.1
mp-24208	CrH_2	$Fm\bar{3}m$	50	–	0.52	477.6	6.9	Yes	Yes	0.65	323.2	9.4
	CrH_2	$Fm\bar{3}m$	100	–	0.53	561.6	8.4	Yes	Yes	0.64	348.0	9.7

the learning algorithm, and ultimately the ML models that are developed.

Two learning curves obtained by using GPR to learn the (featurized) λ and ω_{log} datasets are shown in Figs. 5 (a) and (b). In both cases, the training curves saturate at $\simeq 0.15$ (for λ) and 110 K (for ω_{log}). These values are small, i.e., they are $\simeq 3-5\%$ of the data range, implying that GPR can successfully capture the behaviors of the data. On the other hand, the validation curves of λ and ω_{log} data do not saturate and keep decreasing. This behavior strongly suggests that if more data are available, the gap between the learning and the validation curves can further be reduced and the performance of the target ML models can readily be elevated.

Figs. 5 (a) and (b) reveal that an error of $\simeq 0.4$ and $\simeq 200$ K can be expected for the predictions of λ and ω_{log}, respectively. The expected errors are roughly 7 % of the whole range of λ and ω_{log} data, which are significantly small compared to the results reported in Ref. 51. Figs. 5 (c) and (d) visualize two typical ML models trained on 90% of the λ and ω_{log} datasets and validated on the remaining 10% of the datasets. Likewise, Figs. 5 (e) and (f) visualize two typical ML models, each of them was trained on the entire λ or ω_{log} dataset using the exactly same procedure. In fact, each of them is one of 100 ML models that were trained independently and used to predict λ and ω_{log} of the candidate set.

B. Discovered superconductors and validations

We used the developed ML models to predict λ and ω_{log} of 35 atomic structures in the candidate set, and then to compute the critical temperature T_c using the McMillan formula with $\mu^* = 0.1$. The predicted λ ranges from 0.31 to 0.88, and consequently, the predicted T_c ranges from 0.16 K to 21.3 K. Six candidates with highest predicted T_c are those with Materials Project ID of mp-24289, mp-1018133, mp-24081, mp-24287, mp-1008376, and mp-24208. Details of these candidates are summarized in Table I while comprehensive information of all 35 candidates can be found in Supplemental Material [85].

Examining the top six candidates, we found that mp-24081 is a trigonal structure of ScClH, whose primitive cell has 6 atoms and three very small lattice angles ($\alpha = \beta = \gamma = 21.38°$). Computations of the EP interactions in such a structure are prohibitively expensive because the required **k**-point and **q**-point grids must be extremely large. In addition, Ce, the species showing up in mp-1008376, a cubic structure of CeH_3, is not supported by ONCVPSP-PBE-PDv0.4 norm-conserving pseudopotential set [83]. Therefore, computations were performed for the remaining four candidates. Among them, mp-24289, a cubic structure of PdH and mp-1018133, a tetragonal structure of LiHPd, are dynamically unstable. In principles, each of them can be stabilized by following the imaginary phonon modes to end up at a dynamically stable structure with lower energy and symmetry [93]. Such heavy and cumbersome technical procedure was reserved for the next steps. The last two candidates are mp-24287, which is a cubic structure of CrH and mp-24208, which is also a cubic structure of CrH_2. Both of them, visualized in Figs. 6 (a) and (b), are dynamically stable and thus, their λ and ω_{log} were computed using the procedure described in Sec. II E. The phonon band structures, which prove the dynamical stability of mp-24289, mp-1018133, mp-24287, and mp-24208, can be found in the Supplemental Material [85].

Predicted and computed $\alpha^2F(\omega)$, λ, ω_{log}, and T_c (using the McMillan formula with $\mu^* = 0.1$) of mp-24287 and mp-24208 at $P = 0$ are given in Table I and Figs. 6 (c) and (d). Considering the expected errors of the ML models, it is obvious that the computed λ and ω_{log}

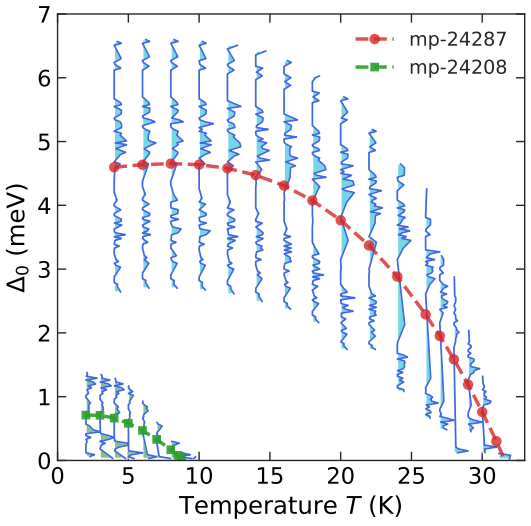

FIG. 7. Distribution of the zero-pressure superconducting gap function $\Delta_0(T)$ computed by numerically solving the Éliashberg equations for mp-24287 and mp-24208. The dashed curves, joining the middle point of the distributions, serve as the guide to the eyes. The critical temperature T_c is estimated to be at the middle point of the downward-sloping segment of the $\Delta_0(T)$ curves.

agree remarkably well with the ML predicted values. Given that magnesium diboride MgB_2 in its hexagonal $P6/mmm$ phase is the highest-T_c conventional superconductor with $T_c \simeq 39$ K [94], the examined materials have respectable (computed) critical temperature, i.e., $T_c = 15.7$ K for mp-24287 and $T_c = 10.7$ for mp-24208. By examining the electronic structures of mp-24287 and mp-24208 reported in Materials Project database, we confirmed that both of them are metallic in nature with a large density of states at the Fermi level.

We extended our validation to the high-P domain by predicting and then computing λ and $\omega_{\log}$ of mp-24287 and mp-24208 after optimizing them at $P = 50$ GPa and $P = 100$ GPa. Both of them were found to be dynamically stable at these pressures while the computed superconducting properties are shown in Table I and Figs. 6 (e) and (f). We also found that the computed and the predicted values of λ and $\omega_{\log}$ at $P = 50$ GPa and $P = 100$ GPa are remarkably consistent. For both materials, computed λ and T_c decrease while $\omega_{\log}$ increases from 0 to 100 GPa, and the ML models capture correctly these behaviors within the expected errors given from the analysis of the learning curves in Sec. III A. Specifically, predictions of λ at $P = 50$ GPa and $P = 100$ GPa are within 0.1 from the computed results, leading to a remarkably small error of $\simeq 3$ K in predicting T_c.

C. Further assessments on the predictions

We attempted to verify our predictions in a few ways. First, additional calculations for λ and $\omega_{\log}$ of mp-24287 and mp-24208 using the local-density approximation (LDA) XC functional were performed at all the pressure values examined (see Sec. III B). The obtained results, as given in the Supplemental Material [85], are highly consistent with, i.e., within $2-3\%$ of, the reported results using the PBE XC functional.

Next, we used EPW code [62–64] to numerically solve the Éliashberg equations on the imaginary axis and then approximated the real-axis superconducting gap Δ_0 of mp-24287 and mp-24208 using Páde continuation [95]. Within this scheme, the electron-phonon interactions were computed by Quantum ESPRESSO [57, 58], using the ultra-soft pseudopotentials from PS Library [96], an energy cutoff of 120 Ry (which is 60 Ha, $\simeq 1,600$ eV), a $\mathbf{k}$-point grid of $24\times24\times24$ and a $\mathbf{q}$-point grid of $6\times6\times6$. During the EPW calculations, we used a fine $\mathbf{q}$-point grid of $12\times12\times12$ and $\mu^* = 0.1$. The superconducting gap $\Delta_0(T)$ computed for mp-24287 and mp-24208 and shown in Fig. 7 projects a $T_c \simeq 22-24$ K for mp-24287 and a $T_c \simeq 7-8$ K for mp-24208. These values are in good agreement with that reported in Fig. 6, providing a confirmation of the predicted superconductivity of mp-24287 and mp-24208 at $P = 0$ GPa.

Finally, we turn our attention to the synthesizability of mp-24287 and mp-24208 by tracing their origin. Information from Materials Project database allows us to track them down to two entries numbered 191080 and 26630 of the Inorganic Crystalline Structure Database (ICSD), and finally to Refs. 97 and 98, respectively. In short, mp-24287 and mp-24208 were experimentally synthesized and resolved [97, 98] sometimes in the past. Afterwards, some experimental [99, 100] and computational [101, 102] efforts followed, examining their magnetic, electronic, and mechanical properties. Perhaps because preparing them experimentally is challenging, little more is known about these materials. Given the documented evidence of the synthesizability of both mp-24287 and mp-24208 at 0 GPa, which is in contrast with the enormous challenges of performing experiments at hundreds of GPa, we hope that these materials will be resynthesized and tested for the predicted superconductivity in the near future.

IV. REMARKS AND GOING FORWARD

Predicting λ and $\omega_{\log}$ from the atomic structures has some advantages. First, the correlation between the atomic structures and λ and $\omega_{\log}$, which will be learned, is direct, physics-inspired, and intuitive, while computing T_c from λ and $\omega_{\log}$ is trivial. Second, the obtained ML models, which are accurate and robust, can be directly used not only for extant massive material databases with millions of atomic structures but also for any structure searches in an *on-the-fly* manner. Finally, by using pressure as an implicit input, the training data can be highly diverse and comprehensive, ultimately allowing the ML models to be able to handle unusual atomic envi-

ronments, frequently encountered during unconstrained structure searches for new materials.

The accuracy demonstrated in Sec. III B for the ML models of λ and $\omega_{\log}$ is rooted from a series of factors. The list includes at least a reliable training dataset, a featurizing procedure that can capture the essential information encoded in the atomic structures, a ML algorithm that can learn the featurized data efficiently, a careful justification of the domain of applicability of the ML models, and a good candidate set. On the other hand, these stringent factors limit the number of candidates used in this work, although the ML models are already very fast to make millions of predictions.

In the next steps, we will improve the whole scheme in several ways. First, by enlarging and diversifying the dataset while maintaining its quality, the domain of applicability of the ML models will be systematically expanded. For examples, the candidate set obtained from the selection procedure described in Fig. 4 will jump to 2,694 atomic structures when we can remove the requirement of having H in the chemical composition. Coming that point, we believe that many more new superconductors can be identified and validated, at least by first-principles computations. Second, modern deep learning techniques will be used to improve and possibly to unify the featurizing and the learning steps. Third, the ML models will be integrated in an inverse design strategy to explore the practically infinite materials space in an efficient manner. Currently, (inverse) design of functional materials with targeted properties is a very active research area with many success stories [103–110]. We hope that superconducting materials discoveries can be added to this list in the near future. Finally, we will work with experimental experts to synthesize and test the superconducting materials discovered computationally, closing the loop of materials design.

V. CONCLUSIONS

We have demonstrated a ML approach for the discovery of conventional superconductors at any pressure. By exploring and learning the direct and physics-inspired correlation between the atomic structures and their possible superconducting properties, specifically λ and $\omega_{\log}$, highly accurate and reliable ML models were developed. These models were validated against the standard first-principles calculations of λ and $\omega_{\log}$, identifying two potential superconducting materials with respectable critical temperature T_c at zero pressure. Interestingly, these materials have been synthesized and studied in some other contexts. The main implication of this approach is that by learning the high-P atomic-level details that are connected to high-T_c superconductivity, the obtained ML models can be used to identify the atomic structures realized at zero pressure with possible high-T_c superconductivity. Given that the models can be used directly for massive materials databases with millions of atomic configurations, more superconductors can be expected in near future. We plan to improve this strategy in multiple ways, hoping that it can better contribute to the search of high-T_c superconductors that has been highly active during the last decade.

ACKNOWLEDGEMENTS

Work by T.N.V. was supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA03. The authors thank Chris Pickard, Guochun Yang, Bin Li, Samuel Poncé, and Kamal Choudhary for useful communications. Computations were performed at the San Diego Supercomputer Center (Expanse) within the XSEDE/ACCESS allocation number DMR170031.

[1] A. Drozdov, M. Eremets, I. Troyan, V. Ksenofontov, and S. Shylin, Nature **525**, 73 (2015).
[2] A. P. Drozdov, P. P. Kong, V. S. Minkov, S. P. Besedin, M. A. Kuzovnikov, S. Mozaffari, L. Balicas, F. F. Balakirev, D. E. Graf, V. B. Prakapenka, E. Greenberg, D. A. Knyazev, T. M., and M. I. Eremets, Nature **569**, 528 (2019).
[3] E. Snider, N. Dasenbrock-Gammon, R. McBride, M. Debessai, H. Vindana, K. Vencatasamy, K. V. Lawler, A. Salamat, and R. P. Dias, Nature **586**, 373 (2020).
[4] D. Duan, Y. Liu, F. Tian, D. Li, X. Huang, Z. Zhao, H. Yu, B. Liu, W. Tian, and T. Cui, Sci. Rep. **4**, 6968 (2014).
[5] E. Zurek and T. Bi, J. Chem. Phys. **150**, 050901 (2019).
[6] K. P. Hilleke and E. Zurek, J. Appl. Phys. **131**, 070901 (2022).
[7] I. Errea, F. Belli, L. Monacelli, A. Sanna, T. Koretsune, T. Tadano, R. Bianco, M. Calandra, R. Arita, F. Mauri, *et al.*, Nature **578**, 66 (2020).
[8] G. Gao, L. Wang, M. Li, J. Zhang, R. T. Howie, E. Gregoryanz, V. V. Struzhkin, L. Wang, and S. T. John, Mater. Today Phys. **21**, 100546 (2021).
[9] L. Boeri, R. G. Hennig, P. J. Hirschfeld, G. Profeta, A. Sanna, E. Zurek, W. E. Pickett, M. Amsler, R. Dias, M. Eremets, C. Heil, R. Hemley, H. Liu, Y. Ma, C. Pierleoni, A. Kolmogorov, N. Rybin, D. Novoselov, V. I. Anisimov, A. R. Oganov, C. J. Pickard, T. Bi, R. Arita, I. Errea, C. Pellegrini, R. Requist, E. Gross, E. R. Margine, S. R. Xie, Y. Quan, A. Hire, L. Fanfarillo, G. R. Stewart, J. J. Hamlin, V. Stanev, R. S. Gonnelli, E. Piatti, D. Romanin, D. Daghero, and R. Valenti, J. Phys. Condens. Matter **34**, 183002 (2021).
[10] M. Yazdani-Asrami, A. Sadeghi, W. Song, A. Madureira, J. Pina, A. Morandi, and M. Parizh,

Supercond. Sci. Technol. **35**, 123001 (2022).
[11] S. Shah and A. N. Kolmogorov, Phys. Rev. B **88**, 014107 (2013).
[12] Z. Zhang, T. Cui, M. J. Hutcheon, A. M. Shipley, H. Song, M. Du, V. Z. Kresin, D. Duan, C. J. Pickard, and Y. Yao, Phys. Rev. Lett. **128**, 047001 (2022).
[13] N. Ashcroft, Phys. Rev. Lett. **92**, 187002 (2004).
[14] A. R. Oganov, ed., *Modern Methods of Crystal Structure Prediction* (Wiley-VCH, Weinheim, Germany, 2011).
[15] A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs, Nat. Rev. Mater. **4**, 331 (2019).
[16] C. J. Pickard and R. J. Needs, J. Phys. Condens. Matter. **23**, 053201 (2011).
[17] R. J. Needs and C. J. Pickard, APL Materials **4**, 053210 (2016).
[18] T. D. Huan, V. Sharma, G. A. Rossetti, and R. Ramprasad, Phys. Rev. B **90**, 064111 (2014).
[19] T. D. Huan, Phys. Rev. Mater. **2**, 023803 (2018).
[20] T. D. Huan, V. N. Tuoc, and N. V. Minh, Phys. Rev. B **93**, 094105 (2016).
[21] S. Baroni, S. de Gironcoli, and A. Dal Corso, Rev. Mod. Phys. **73**, 515 (2001).
[22] F. Giustino, Rev. Mod. Phys. **89**, 015003 (2017).
[23] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **106**, 162 (1957).
[24] J. Hirsch and F. Marsiglio, Nature **596**, E9 (2021).
[25] T. Wang, M. Hirayama, T. Nomoto, T. Koretsune, R. Arita, and J. A. Flores-Livas, Phys. Rev. B **104**, 064510 (2021).
[26] J. Hirsch and F. Marsiglio, Phys. Rev. B **103**, 134505 (2021).
[27] M. Gubler, J. A. Flores-Livas, A. Kozhevnikov, and S. Goedecker, Phys. Rev. Mater. **6**, 014801 (2022).
[28] M. Eremets, V. Minkov, A. Drozdov, P. Kong, V. Ksenofontov, S. Shylin, R. Prozorov, F. Balakirev, D. Sun, S. Mozaffari, and L. Balicas, J. Supercond. Nov. Magn. **35**, 965 (2022).
[29] J. Hirsch, Appl. Phys. Lett. **121**, 080501 (2022).
[30] W. E. Pickett, arXiv preprint arXiv:2204.05930 (2022).
[31] G. Eliashberg, Sov. Phys. J. Exp. Theor. Phys. **11**, 696 (1960).
[32] P. B. Allen and B. Mitrović, Solid State Phys. **37**, 1 (1983).
[33] F. Marsiglio, Ann. Phys. **417**, 168102 (2020).
[34] A. V. Chubukov, A. Abanov, I. Esterlis, and S. A. Kivelson, Ann. Phys. **417**, 168190 (2020).
[35] W. L. McMillan, Phys. Rev. **167**, 331 (1968).
[36] R. Dynes, Solid State Commun. **10**, 615 (1972).
[37] P. B. Allen and R. C. Dynes, Phys. Rev. B **12**, 905 (1975).
[38] Q. Yang, J. Lv, Q. Tong, X. Du, Y. Wang, S. Zhang, G. Yang, A. Bergara, and Y. Ma, Phys. Rev. B **103**, 024505 (2021).
[39] S. Xie, G. Stewart, J. Hamlin, P. Hirschfeld, and R. Hennig, Phys. Rev. B **100**, 174513 (2019).
[40] S. Xie, Y. Quan, A. Hire, B. Deng, J. DeStefano, I. Salinas, U. Shah, L. Fanfarillo, J. Lim, J. Kim, *et al.*, npj Comput. Mater. **8**, 14 (2022).
[41] K. Hamidieh, Comput. Mater. Sci. **154**, 346 (2018).
[42] K. Matsumoto and T. Horide, Appl. Phys. Express **12**, 073003 (2019).
[43] T. Ishikawa, T. Miyake, and K. Shimizu, Phys. Rev. B **100**, 174506 (2019).
[44] A. M. Shipley, M. J. Hutcheon, R. J. Needs, and C. J. Pickard, Phys. Rev. B **104**, 054501 (2021).
[45] P. Song, Z. Hou, P. B. de Castro, K. Nakano, K. Hongo, Y. Takano, and R. Maezono, arXiv preprint arXiv:2103.00193 (2021).
[46] T. D. Le, R. Noumeir, H. L. Quach, J. H. Kim, J. H. Kim, and H. M. Kim, IEEE Trans. Appl. Supercond. **30**, 1 (2020).
[47] P. J. García-Nieto, E. Garcia-Gonzalo, and J. P. Paredes-Sánchez, Neural. Comput. Appl. **33**, 17131 (2021).
[48] V. Stanev, K. Choudhary, A. G. Kusne, J. Paglione, and I. Takeuchi, Commun. Mater. **2**, 1 (2021).
[49] S. Raviprasad, N. A. Angadi, and M. Kothari, in *2022 3rd International Conference for Emerging Technology (INCET)* (IEEE, 2022) pp. 1–5.
[50] G. Revathy, V. Rajendran, B. Rashmika, P. S. Kumar, P. Parkavi, and J. Shynisha, Materials Today: Proceedings (2022).
[51] K. Choudhary and K. Garrity, npj Comput. Mater. **8**, 244 (2022).
[52] T. N. Vu, S. K. Nayak, N. T. T. Nguyen, S. P. Alpay, and H. Tran, AIP Adv. **11**, 045120 (2021).
[53] V. N. Tuoc, N. T. Nguyen, V. Sharma, and T. D. Huan, Phys. Rev. Mater. **5**, 125402 (2021).
[54] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Materials **1**, 011002 (2013).
[55] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. Billinge, *et al.*, npj Comput. Mater. **8**, 59 (2022).
[56] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).
[57] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. d. Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, J. Phys.: Condens. Matter **21**, 395502 (2009).
[58] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, *et al.*, J. Phys.: Condens. Matter **29**, 465901 (2017).
[59] X. Gonze, B. Amadon, P.-M. Anglade, J.-M. Beuken, F. Bottin, P. Boulanger, F. Bruneval, D. Caliste, R. Caracas, M. Côté, T. Deutsch, L. Genovese, P. Ghosez, M. Giantomassi, S. Goedecker, D. Hamann, P. Hermet, F. Jollet, G. Jomard, S. Leroux, M. Mancini, S. Mazevet, M. Oliveira, G. Onida, Y. Pouillon, T. Rangel, G.-M. Rignanese, D. Sangalli, R. Shaltaf, M. Torrent, M. Verstraete, G. Zerah, and J. Zwanziger, Comput. Phys. Commun. **180**, 2582 (2009).
[60] X. Gonze, G. M. Rignanese, M. Verstraete, J.-M. Beuken, Y. Pouillon, R. Caracas, F. Jollet, M. Torrent, G. Zerah, M. Mikami, P. Ghosez, M. Veithen, J.-Y. Raty, V. Olevano, F. Bruneval, L. Reining, R. Godby, G. Onida, D. R. Hamann, and D. C. Allan, Zeit. Kristallogr. **220**, 558 (2005).

[61] X. Gonze, F. Jollet, F. A. Araujo, D. Adams, B. Amadon, T. Applencourt, C. Audouze, J.-M. Beuken, J. Bieder, A. Bokhanchuk, E. Bousquet, F. Bruneval, D. Caliste, M. Côté, F. Dahm, F. D. Pieve, M. Delaveau, M. D. Gennaro, B. Dorado, C. Espejo, G. Geneste, L. Genovese, A. Gerossier, M. Giantomassi, Y. Gillet, D. Hamann, L. He, G. Jomard, J. L. Janssen, S. L. Roux, A. Levitt, A. Lherbier, F. Liu, I. Lukačević, A. Martin, C. Martins, M. Oliveira, S. Poncé, Y. Pouillon, T. Rangel, G.-M. Rignanese, A. Romero, B. Rousseau, O. Rubel, A. Shukri, M. Stankovski, M. Torrent, M. V. Setten, B. V. Troeye, M. Verstraete, D. Waroquiers, J. Wiktor, B. Xu, A. Zhou, and J. Zwanziger, Comput. Phys. Commun. **205**, 106 (2016).

[62] F. Giustino, M. L. Cohen, and S. G. Louie, Phys. Rev. B **76**, 165108 (2007).

[63] E. R. Margine and F. Giustino, Phys. Rev. B **87**, 024505 (2013).

[64] S. Poncé, E. R. Margine, C. Verdi, and F. Giustino, Comput. Phys. Commun. **209**, 116 (2016).

[65] M. Lüders, M. Marques, N. Lathiotakis, A. Floris, G. Profeta, L. Fast, A. Continenza, S. Massidda, and E. Gross, Phys. Rev. B **72**, 024545 (2005).

[66] M. Marques, M. Lüders, N. Lathiotakis, G. Profeta, A. Floris, L. Fast, A. Continenza, E. Gross, and S. Massidda, Phys. Rev. B **72**, 024546 (2005).

[67] A. Sanna, C. Pellegrini, and E. Gross, Phys. Rev. Lett. **125**, 057001 (2020).

[68] Y. Xie, Q. Li, A. R. Oganov, and H. Wang, Acta Crystallogr. C Struct. Chem. **70**, 104 (2014).

[69] D. Y. Kim, R. H. Scheicher, and R. Ahuja, Phys. Rev. Lett. **103**, 077002 (2009).

[70] S. Di Cataldo, W. Von Der Linden, and L. Boeri, Phys. Rev. B **102**, 014516 (2020).

[71] H. Xie, Y. Yao, X. Feng, D. Duan, H. Song, Z. Zhang, S. Jiang, S. A. Redfern, V. Z. Kresin, C. J. Pickard, *et al.*, Phys. Rev. Lett. **125**, 217001 (2020).

[72] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. Dal Corso, *et al.*, Science **351**, aad3000 (2016).

[73] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[74] W. Kohn and L. Sham, Phys. Rev. **140**, A1133 (1965).

[75] J. Zhang, J. M. McMahon, A. R. Oganov, X. Li, X. Dong, H. Dong, and S. Wang, Phys. Rev. B **101**, 134108 (2020).

[76] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, Comput. Mater. Sci. **152**, 60 (2018).

[77] 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, J. Mach. Learn. Res. **12**, 2825 (2011).

[78] C. K. I. Williams and C. E. Rasmussen, in *Advances in Neural Information Processing Systems 8*, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (MIT Press, 1995).

[79] C. E. Rasmussen and C. K. I. Williams, eds., *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA, 2006).

[80] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).

[81] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).

[82] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[83] D. Hamann, Phys. Rev. B **88**, 085117 (2013).

[84] M. J. van Setten, M. Giantomassi, E. Bousquet, M. J. Verstraete, D. R. Hamann, X. Gonze, and G.-M. Rignanese, Comput. Phys. Commun. **226**, 39 (2018).

[85] See Supplemental Material for more information reported in this paper.

[86] H. Tran, A. Toland, K. Stellmach, M. K. Paul, W. Gutekunst, and R. Ramprasad, J. Phys. Chem. Lett. **13**, 4778 (2022).

[87] M. Aykol, S. S. Dwaraknath, W. Sun, and K. A. Persson, Sci. Adv. **4**, eaaq0148 (2018).

[88] R. Batra, H. D. Tran, and R. Ramprasad, Appl. Phys. Lett. **108**, 172902 (2016).

[89] R. Batra, T. D. Huan, J. Jones, G. A. Rossetti, and R. Ramprasad, J. Phys. Chem. C **121**, 4139 (2017).

[90] X. Sang, E. D. Grimley, T. Schenk, U. Schroeder, and J. M. LeBeau, Appl. Phys. Lett. **106**, 162905 (2015).

[91] T. S. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, Appl. Phys. Lett. **99**, 102903 (2011).

[92] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, Vol. 112 (Springer, 2013).

[93] H. D. Tran, M. Amsler, S. Botti, M. A. L. Marques, and S. Goedecker, J. Chem. Phys. **140**, 124708 (2014).

[94] J. Nagamatsu, N. Nakagawa, T. Muranaka, Y. Zenitani, and J. Akimitsu, Nature **410**, 63 (2001).

[95] F. Marsiglio, M. Schossmann, and J. Carbotte, Phys. Rev. B **37**, 4965 (1988).

[96] A. Dal Corso, Comput. Mater. Sci. **95**, 337 (2014).

[97] V. Antonov, A. Beskrovnyy, V. Fedotov, A. Ivanov, S. Khasanov, A. Kolesnikov, M. Sakharov, I. Sashin, and M. Tkacz, J. Alloys Compd. **430**, 22 (2007).

[98] C. A. Snavely and D. A. Vaughan, J. Am. Chem. Soc. **71**, 313 (1949).

[99] J. Poźniak-Fabrowska, B. Nowak, and M. Tkacz, J. Alloys Compd. **322**, 82 (2001).

[100] V. E. Antonov, V. K. Fedotov, A. S. Ivanov, A. I. Kolesnikov, M. A. Kuzovnikov, M. Tkacz, and V. A. Yartys, J. Alloys Compd. **905**, 164208 (2022).

[101] K. Miwa and A. Fukumoto, Phys. Rev. B **65**, 155114 (2002).

[102] S. Kanagaprabha, R. Rajeswarapalanichamy, G. Sudhapriyanga, A. Murugan, M. Santhosh, and K. Iyakutti, in *AIP Conf. Proc.*, Vol. 1665 (AIP Publishing LLC, 2015) p. 030010.

[103] T. D. Huan, A. Mannodi-Kanakkithodi, and R. Ramprasad, Phys. Rev. B **92**, 014106 (2015).

[104] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, Sci. Rep. **6**, 20952 (2016).

[105] Y.-Y. Zhang, W. Gao, S. Chen, H. Xiang, and X.-G. Gong, Comput. Mater. Sci. **98**, 51 (2015).

[106] H. J. Xiang, B. Huang, E. Kan, S.-H. Wei, and X. G. Gong, Phys. Rev. Lett. **110**, 118702 (2013).

[107] V. Fung, J. Zhang, G. Hu, P. Ganesh, and B. G. Sumpter, npj Comput. Mater. **7**, 1 (2021).

[108] G. M. Coli, E. Boattini, L. Filion, and M. Dijkstra, Sci. Adv. **8**, eabj6731 (2022).

[109] A. Lininger, M. Hinczewski, and G. Strangi, ACS Photonics **8**, 3641 (2021).
[110] C. J. Court, A. Jain, and J. M. Cole, Chem. Mater. **33**, 7217 (2021).