Mapping superconductivity in high-pressure hydrides: The Superhydra project

Santanu Saha,[1,*] Simone Di Cataldo,[1,2] Federico Giannessi,[2,3] Alessio Cucciari,[2] Wolfgang von der Linden,[1] and Lilia Boeri[2,3]

[1]*Institute of Theoretical and Computational Physics, Graz University of Technology, NAWI Graz, 8010 Graz, Austria*
[2]*Dipartimento di Fisica, Università di Roma La Sapienza, Piazzale Aldo Moro 5, I-00185 Roma, Italy*
[3]*Centro Ricerche Enrico Fermi, Via Panisperna 89 A, 00184 Rome, Italy*

(Received 6 January 2023; accepted 25 April 2023; published 31 May 2023)

The discovery of high-T_c conventional superconductivity in high-pressure hydrides has helped establish computational methods as a formidable tool to guide material discoveries in a field traditionally dominated by serendipitous experimental search. This paves the way to an ever-increasing use of data-driven approaches to the study and design of superconductors. In this work, we propose a new adaptive method to generate meaningful datasets of superconductors, based on element substitution into a small set of representative structural templates, generated by crystal structure prediction methods—adapted high-throughput approach. Our approach realizes an optimal compromise between structural variety and computational efficiency and can be easily generalized to other elements and compositions. As a first application, we apply it to binary hydrides at high pressure, realizing a database of 880 hypothetical structures, characterized with a set of electronic, vibrational, and chemical descriptors. In our `Superhydra` Database, 139 structures are superconducting according to the McMillan-Allen-Dynes approximation. Studying the distribution of T_c and other properties across the database with advanced statistical and visualization techniques, we are able to obtain comprehensive material maps of the phase space of binary hydrides. The `Superhydra` database can be thought as a first step of a generalized effort to map conventional superconductivity.

DOI: 10.1103/PhysRevMaterials.7.054806

I. INTRODUCTION

Methods for *ab initio* prediction of superconductors led to an impressive acceleration in material discoveries, following the breakthrough report of high-T_c superconductivity in high-pressure superhydrides [1,2].

After reaching the symbolical threshold of room-temperature [3–5], the research focus is rapidly shifting to the even more ambitious goal of conventional superconductivity at ambient conditions [6–9]. Meeting this challenge will require new methods which are able to explore a potentially huge material space, much broader than that of simple binary hydrides [10,11]; high-throughput (HT) screening and artificial intelligence techniques will likely play a prominent role in this search [12–19].

High-pressure superhydrides represent an ideal benchmark for such data-driven approaches, since first-principles calculations can provide accurate estimates for T_c and overcome the limitations of scarcity and inhomogeneity of data associated with other classes of superconductors. Indeed, a few works in literature have attempted to study the distribution of T_c in binary hydrides, applying empirical classification schemes to data previously published in literature [19–21] or randomly-generated structures [9,18]. Both approaches present obvious intrinsic drawbacks: While datasets based on literature are usually limited to near-ground-state structures, losing precious information on other possible chemical environments, randomly-generated sets tend to contain many unphysical or unrepresentative structures [22].

In this paper we propose an alternative strategy for the exploration of the phase space of superconductors, based on the high-throughput substitution of the elements of the periodic table in a small set of representative structural templates for different compositions—adapted high-throughput approach (adapted-HT)—which is explicitly designed to realize the best possible compromise between variety, reproducibility, and efficiency. A key aspect of our adapted HT approach against the regularly used high-throughput approaches is on the approach of identifying promising structural templates. Against conventional use of *a priori* knowledge on material systems (experiment or theory) from literature, we used crystal structure prediction methods for generation and identification of structural templates (Sec. II A) to avoid any bias or limited knowledge on the test system. By utilizing the set of high-symmetric, representative templates for all elements of the periodic table, we managed to reduce the computational time devoted to structural search and focus our resources on computing the T_c for a wider spectrum of structures, regardless of their thermodynamic stability. To maintain our study manageable, we focus on a single pressure (200 GPa), and to a finite number of compositions, as discussed in the following.

As a first application, the adapted HT approach is applied to map the phase space of binary hydrides formed by the first 55 elements of the periodic table (H and He are excluded). Our `Superhydra` database, available as a single CSV file in the Supplemental Material (SM) [23], comprises a total of 880 distinct binary hydride structures, of which 139 are

*santanu.saha@tugraz.at

superconducting and 27 are high-T_c; all structures are characterized by a set of relevant structural, electronic, vibrational and superconducting properties, computed at the same level of accuracy. Using advanced statistical methods to analyze our database, we obtain a set of *material maps* depicting the distribution of physical properties across different families of binary hydrides, and identify general trends and correlations with thermodynamical stability and superconductivity. We also identify few unreported high-T_c hydrides.

This paper is structured as follows: in Sec. II we describe the construction of the `Superhydra` database, i.e., the generation of the templates, the choice of the relevant features, and the classification of the hypothetical binary hydride structures into broad families, emerging from an advanced statistical analysis. In Sec. III, we discuss the main features of the binary hydrides material space, using material maps of the most relevant quantities. In Sec. IV we present accurate vibrational and superconducting properties for two new interesting high-T_c hydrides, unreported in the literature till date. Finally, in Sec. V, the main conclusions of our paper is discussed.

The main computational details are summarized in the methods section in the Appendix; the SM [23] contains detailed information on the database generation methodology, additional descriptions of the database, the full set of material maps, as well as additional information on the data analysis methods employed.

II. CONSTRUCTION OF THE SUPERHYDRA DATABASE

The `Superhydra` database was generated substituting the first 55 elements of the periodic table following hydrogen and helium (Li-La) into 16 binary hydride templates AH_n, obtained with fixed-compositions evolutionary crystal structure searches for selected elements. Details on the generation procedure are available in the SM [23].

The resulting structures, relaxed at a common pressure of 200 GPa, form a database of $55 \times 16 = 880$ hypothetical binary hydrides, for which we computed various electronic vibrational and thermodynamic properties as listed below (Database 1, or DB1 in the following). For the 139 metallic and dynamically stable structures, we computed electron-phonon properties in the linear response approximation, and estimated the T_c using the McMillan-Allen-Dynes formula, assuming a constant $\mu^* = 0.1$ [24]. These compounds form our database of superconductors (DB2). Twenty-seven of these compounds are high-T_c, according to the conventional definition employed throughout the manuscript, where the threshold for high-T_c is set to the N_2 boiling point (77 K).

A. Structural templates

The most critical issue in the construction of the `Superhydra` database was the choice of a set of structural templates satisfying a few crucial requirements: (i) being sufficiently *representative*, i.e., the structural templates should cover the variety of bonding environments encountered in known binary hydrides; (ii) at the same time, the number of templates should *not be too large*, and (iii) the unit cells of the individual templates should be small and possibly symmetric, to allow efficient high-throughput calculations of electronic and vibrational properties. Finally, (iv) the generation procedure should be *easily generalizable* to other systems and chemical compositions.

To meet these requirements, we proceeded as follows: we first identified a relatively small sets of elements and compositions representative of different classes of binary hydrides: low-T_c metallic hydrides; covalent hydrides like SH_3 and SeH_3 [1,25]; systems that form H_2-rich structures like TeH_4 and LiH_6 [26,27]; cage like structures, as in sodalite-clathrate hydrides like YH_6 [10,21,28]. The elements were chosen to exhibit different chemical properties and high-pressure behaviors. As for compositions, we restricted our search to compounds with chemical formulas AH_n, with $n \leqslant 6$, to ensure a relatively small unit cell and a fast computation of the physical properties. With this choice, we consciously eliminated from the `Superhydra` database a large number of known high-T_c hydrides, such as LaH_{10} [4] and YH_9 [29], under the assumption that the main features of high-n clathrates like hydrides are well represented in the database by clathrates with $n = 6$.

Once elements and compositions were decided, our generation procedure continued as follows: first, we ran several unconstrained fixed-composition evolutionary search runs for binary hydrides AH_n of selected elements (A = Li, Na, Ca, Cl, Y, Fe, S, C, Al) and compositions ($n = 1, 2, 3, 4, 6$) (see the SM [23] for details). At the end of these structural searches, we retained the 10 lowest enthalpy structures for each element and the structures with high symmetry (space group number above 75), and filtered out all the rest. After checking for duplicates, the structures were divided into three classes, based on the smallest H-H distance (d_{HH}): short ($d_{HH} < 1$ Å); medium (1 Å $\leqslant d_{HH} \leqslant 2$ Å); long ($d_{HH} > 2$ Å).

The final set of 16 templates was obtained by retaining mostly cubic structures (7), a few hexagonal (2), trigonal (3), tetragonal (3) ones, and one orthorhombic, imposing that the set should contain at least two templates for each composition, and requiring that the templates would be balanced with respect to short, medium, and long d_{HH}. Structural details for each template are provided in the SM [23].

The 16 templates are displayed in Fig. 1. In the following, they will be represented by an alphanumeric id of the form XN-SPG-A_mH_n where X stands for short (S), medium (M), or long (L) type based on the smallest d_{HH}; N represents the serial number; SPG the space group number of the structure and A_mH_n, composition of the structure. Hence, e.g., M4-166-A_1H_6 would represents the 4th structure of medium type with space group number 166 and composition AH_6. Among the templates we find well-known crystal structure types. In particular, the M7-229-A_1H_3 and the M3-229-A_1H_6 structures correspond to the high-T_c SH_3 and AH_6 (A = Mg, Ca, Y) templates [1,2,29–32], respectively; while M1-223-A_1H_3 and M5-221-A_1H_3 correspond to the high-pressure phases of AlH_3 and FeH_3 [33–35], respectively. L1-221-A_1H_1 and L2-225-A_1H_1 coincide with the CsCl and NaCl structures, and S3-166-A_1H_6 with the predicted LiH_6 structure [27]. On contrary, the S1-139-A_1H_4 template has been studied extensively in the past [36].

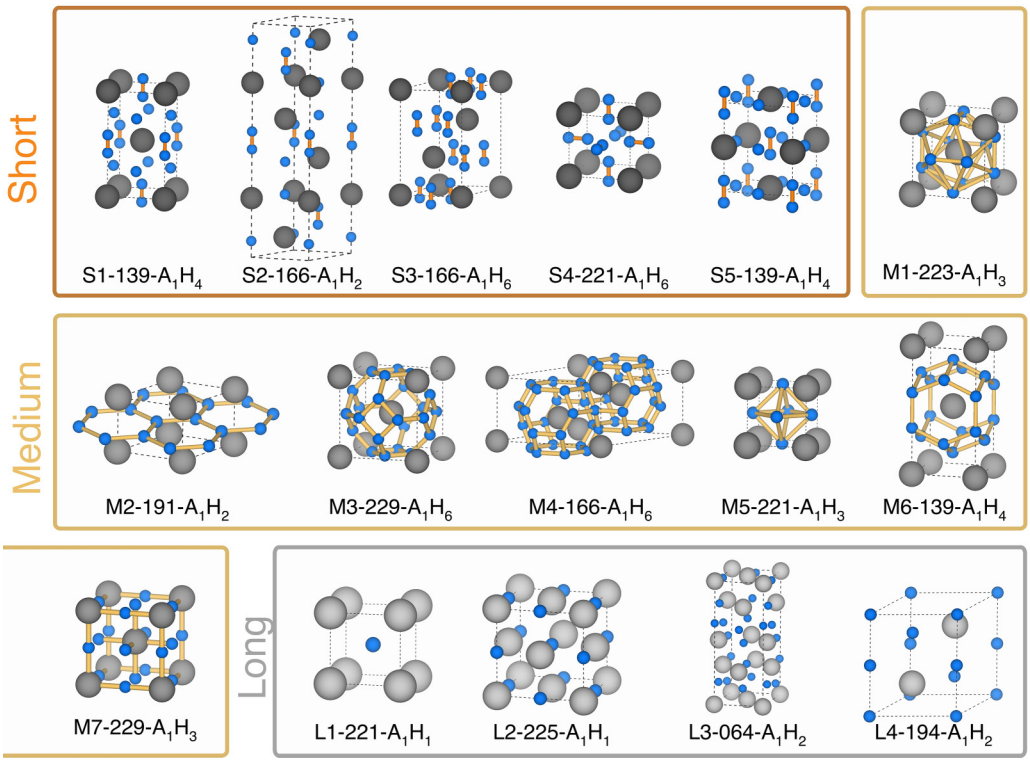

FIG. 1. Template structures used for adapted high-throughput data generation, classified by H-H bond distance into short (S), medium (M), and long (L). The label XN-SPG-A_mH_n indicates the H-H distance type (X), the structure number (N), space group number of the system (SPG), and the generic composition of the structure (A_mH_n). Details of the crystal structures are provided in the Supplemental Material [23].

B. Physical properties—Features

Here, we give a short description of the features grouped in different categories; a more detailed description and definition of each feature is given in the SM [23].

(1) Geometry—space group number of the primitive unit cell (SPG); number of hydrogen atoms per formula unit (n_H); volume (Vol); density (ρ_V); structural template (XN-SPG-A_mH_n); nearest-neighbor distance between H-H (d_{HH}); and between hydrogen and the guest atom A (d_{AH}). The volume and density is in Å^3/atom and gm/cm^3 respectively. The inter-atomic distances are in Å.

(2) Thermodynamics—Formation enthalpy per atom calculated at 200 GPa (ΔH in eV/atom). As reference for the pure elements, we employed structures from literature when available, or performed a dedicated evolutionary structural search with USPEX [37].

(3) Electronic structure—Total density of states (TDOS), partial density of states (PDOS) on atom A/H (ADOS; HDOS) in states/eV/atom, band gap (BG in eV); STATE (metal or insulator).

(4) Charge—Results of the Bader charge analysis on the unit cell; partial charge on atom A (ρ_A); partial charge on atom H (ρ_H), in arb. units—average, maximum, minimum value, standard deviation (avg, max, min, dev).

(5) Elastic—The elastic properties of the system are represented through quasi-bulk modulus BM_q (GPa), estimated at 200 GPa and 0 K. Based on standard thermodynamic relations, the isothermal compressibility (κ_0) at temperature $T = 0$ K and pressure $p = 200$ GPa can be obtained by the following three relations:

$$\begin{aligned}\frac{1}{\kappa_0} &= -\left(\frac{\partial H}{\partial V}\right)\Big|_{V=V_{200}},\\ \frac{1}{\kappa_0} &= -V\left(\frac{\partial p}{\partial V}\right)\Big|_{V=V_{200}},\\ \frac{1}{\kappa_0} &= V\left(\frac{\partial^2 U}{\partial V^2}\right)\Big|_{V=V_{200}},\end{aligned} \tag{1}$$

where H, U, and V represent enthalpy, internal energy, and volume, respectively. We evaluate these quantities at $V = V_{200}$, i.e., effective volume of the relaxed system at 200 GPa. We define the above obtained κ_0 as quasi-bulk modulus BM_q; and obtain BM_q; by averaging the three expressions. The details on the derivation of these expressions and their evaluations from ab-initio calculations is provided in the SM [23].

(6) Phonon—Maximum, minimum, and average phonon frequency at the Γ point ($\omega_\Gamma^{\mathrm{min}}$, $\omega_\Gamma^{\mathrm{max}}$, $\omega_\Gamma^{\mathrm{avg}}$) in meV.

(7) Electron-phonon (e-ph)—These properties have been evaluated for metallic dynamically stable structures only. We report the logarithmic-averaged phonon frequency ($\omega_{\log}$) in K, the total electron-phonon coupling constant (λ) and superconducting critical temperature (T_c) estimated from McMillan-Allen-Dynes formula ($\mu^* = 0.1$).

C. Statistical analysis

Several empirical classification schemes have been proposed for binary hydrides, based on structural or chemical

properties, such as interatomic distances, guest atom charge or atomic radius, electronegativity, electronic charge localization, etc. [18,19,21,38].

For example, it was early recognized that for a superhydride to exhibit high-T_c superconductivity, the interatomic H-H distance d_{HH} should lie in a sweet spot between 0.9 and 1.35 Å [21], and the hydrogen sublattice should be negatively charged [39]. Other authors recognized that high-T_c superhydrides can be divided in two classes (covalent and clathrate hydrides), depending on the existence of a covalent bond between the A–H or a weakened H–H bond [10], while Belli *et al.* divided the known superhydrides into six classes (molecular, covalent, weak H interaction, electride, isolated, and ionic), using an inspection of interatomic distances and electronic localization function profiles [21].

The `Superhydra` database is large and homogeneous enough to attempt an unbiased analysis of the properties distributions across different elements and structures, using machine learning techniques. The aim of this analysis is not to obtain accurate classifications or T_c prediction but rather to guide our study of the material maps letting correlations and trend emerge from the data [40].

We first attempt to see whether the full database can be naturally divided into families, using all calculated features, except for e-ph properties. Unsupervised learning algorithms are designed specifically for the task of separating the data into families related to similarity of some sort, with no assumptions on the physical nature of this similarity.

The k-means clustering algorithm, applied on to the whole `Superhydra` database (DB1), leads to the most physically-meaningful clusterization when $k = 3$ is employed [41]. We name the three clusters thus identified H_2, TMs, and HTC, based on the type of hydrides they comprise. The H_2 cluster contains primarily structures with short H-H bonds ($d_{\mathrm{HH}} <$ 0.8 Å), i.e., molecular H_2 hydrides, covering all possible atomic numbers; the TMs (Transition Metal) cluster contains hydrides of the first-row p elements and transition metals, covering all values of d_{HH} larger than 1.0 Å; the high-T_c (HTC) cluster comprises hydrides of the remaining elements, again with $d_{\mathrm{HH}} > 0.8$ Å. The HTC cluster contains almost all of the high-T_c structures, hence its name.

This simple division into clusters is also rather efficient in filtering out superconductors: out of the 27 hydrides with T_c higher than 77 K, 85% (23) belong to the HTC cluster; 4% (1) to H_2; and 11% (3) to TMs. The outlying high-T_c structures are NaH_6 (S3-166-A_1H_6) in the H_2 cluster, and MoH_6 (S3-166-A_1H_6), NbH_6 (M3-229-A_1H_6), and NbH_6 (M4-166-A_1H_6) in TMs. Known high-T_c phases, such as SH_3, NaH_6, and YH_6 all belong to the HTC cluster.

Figure 2 shows a scatter plot of the data in DB1 in the atomic number (Z) versus H-H distance d_{HH}, colored according to the assignment to one of three clusters: H_2, HTC, and TMs. We note that for clustering we used all the numerical features in DB1, none of which contains direct information on the chemical properties of the A atom, except for the atomic number. Despite this, the clustering algorithm did find a partitioning which bears a striking similarity with the periodic table: the threshold between data points classified in the HTC/TMs clusters coincides with the beginning of the

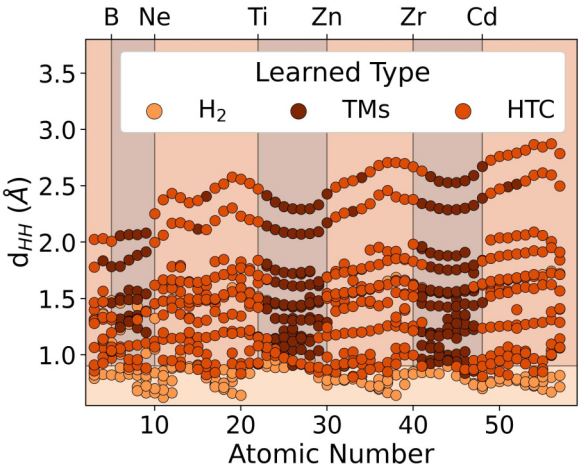

FIG. 2. Distribution of data points in DB1 with respect to minimum H-H distance d_{HH} and atomic number Z. Dark brown, orange, and light orange dots correspond to cluster TMs, HTC, and H_2, respectively.

first-row p- and d-blocks of the periodic table. Upon closer inspection, we note that also a few data points in the H_2 cluster have $d_{\mathrm{HH}} > 0.8$ Å, indicating that clustering is not based on the sole d_{HH} feature.

Having established that the data can be separated into physically-meaningful clusters, which may also correlate with T_c, we are now interested in learning which of the features in our DB may be considered effective indicators (descriptors) of high-T_c behavior. To this end, we performed a feature importance analysis after training a simple Random Forest Classifier [42] to classify the data in DB2 into high- and low-T_c.

After optimizing the model (see the Methods section for details), we obtained an average recall score of 0.79, a precision of 0.79, a $F1$ of 0.79, and an accuracy of 0.90. This result is comparable with similar studies when the dataset had a comparable size [38] and would significantly improve with a larger dataset, which could be easily obtained by systematically adding more templates. Given the relatively poor performance, we do not focus on the model itself, but only used it to obtain a qualitative identification of the most relevant features. These are: volume, ADOS, HDOS, BG, BM_q, Z, ω_{min}, ρ_H, and ΔH. For most of these features, it is easy to identify a posteriori a physical motivation for its correlation with high-T_c. For two features, i.e., the Bader charge on the H atom (ρ_H) and the quasi-bulk modulus BM_q, this correlation is less obvious. However, we will show in the next subsection that both exhibit a strong variability across elements and prototypes, and their distribution permits to draw very informative material maps of the phase space of high-pressure binary hydrides.

III. STRUCTURAL PROPERTY MAPS

In this section we discuss some general trends that emerge from an in-depth analysis of DB1.

In Fig. 3, elements of the periodic table and structural prototypes are arranged along the rows and columns of a property matrix. The resulting boxes are colored with a blue to red heatmap (from lower to higher values), for four different

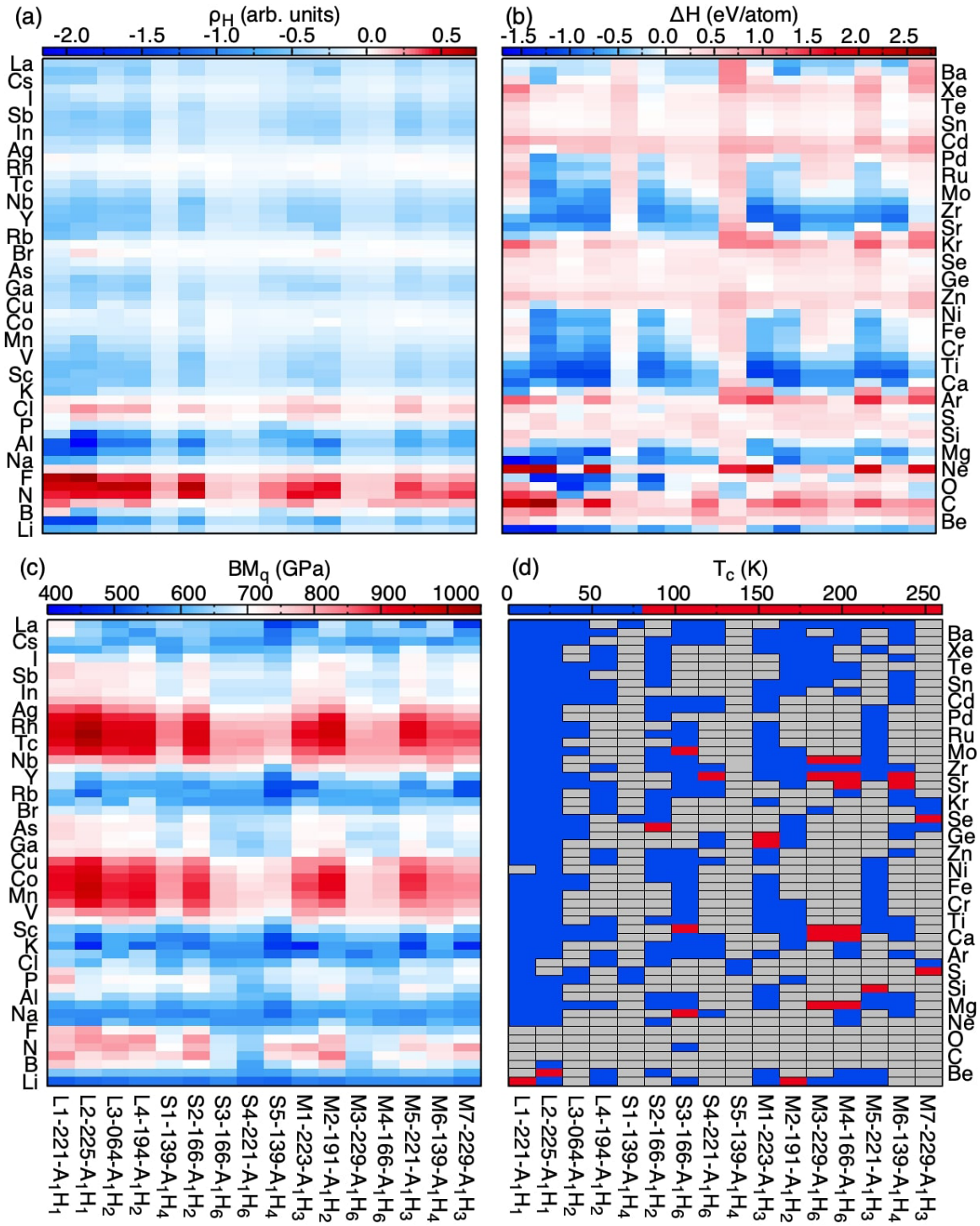

FIG. 3. Heat map of different properties (color) of all the hydride structures for different elements (y axis) and different templates (x axis). All quantities are estimated for pressure of 200 GPa. (a) Average Bader charge on the H atom in arb. units; (b) formation enthalpy ΔH (eV/atom) estimated w.r.t. stable elemental phases; (c) average quasi-bulk modulus (see the SM [23]) in GPa, and (d) critical superconducting temperature T_c in K. Red represents high-T_c and blue low-T_c; the threshold is set to $T_c = 77$ K, corresponding to N_2 boiling temperature. The methodology used to estimate different quantities are provided in the SM [23].

properties: from top to bottom and left to right, we show the average Bader charge on the H atom (ρ_H), the formation enthalpy (ΔH), the quasi-bulk modulus (BM$_q$), and the critical temperature (T_c). Analogous heatmaps for the other features of our DB1 are available in Fig. 2–9 in the SM [23]. The choice of these four observables is motivated as follows: ΔH gives information on the degree of chemical compatibility between the structure and the template, ρ_H describes the type of bonding, BM$_q$ summarizes the lattice properties and quantifies the bonding strength and, as we will discuss later on, correlates with T_c. Last, T_c is the target property, which we want to compare with the other observables.

The formation enthalpy (ΔH) heatmap describes the thermodynamic stability of different hydride templates across the periodic table: blue (red) corresponds to stable (unstable) hydrides. Rows with elements of the *p*-block and late transition metals (Si-Ar), (Cu-Kr), and (Ag-Xe) are red to white, indicating that at 200 GPa the formation of binary hydrides with these elements is less likely. *s*-block elements (except Be) and early *d*-block elements form stable binary hydrides ($\Delta H \leqslant 0$ eV/atom) across different templates.

These results are fairly consistent with available literature [18–20], reporting that high-pressure hydrides are mostly formed by electropositive elements w.r.t. H, as described by Pauling electronegativity [43]. In Fig. 3, electropositive elements exhibit a negative formation enthalpy for both S-, M-, and L-type templates, regardless of the hydrogen content. In other words, the same element can stabilize structures with different hydrogen content, hence with a different d_{HH}, chemical properties, and superconducting behavior. However, strongly electronegative elements, in particular first-row *p*-block elements, are only stable in L-type prototypes. Aside from first-row *p*-block elements, none of the other elements favors the formation of one specific template, while the opposite is true: a few specific templates, namely, S1-139-A_1H_4, S5-139-A_1H_4, and M7-229-A_1H_3 are strongly favored by specific A elements. For instance, the SH_3-like template (M7-229-A_1H_3) is mostly unstable ($\Delta H > 0$), except for a few, selected elements (N, O, S, Sc, Ti). Indeed, in these structures the A-H interaction induces orbital configurations which are energetically favorable only for specific elements.

The ρ_H heatmap in Fig. 3(a) displays the average Bader charge on the H atom w.r.t. the neutral H atom for different combinations of host element A and template.

The majority of the boxes are blue, indicating that H is generally more electronegative than most of the other elements, with the exception of late *d*-block elements (Co-Cu, Rh-Ag), halogens (Cl, Br, I), noble gases (Ar, Kr, Xe), and second-row *p*-block elements (B-Ne). The general behavior essentially follows an electronegativity trend analogous to the one at ambient pressure. However, upon closer inspection, it can be noted that the columns where ΔH switches from blue to red (and vice versa) correspond to Si, Ar, Ni, Kr, Pd, and Xe. These are precisely the elements where the calculated high-pressure electronegativity changes sign, according to a recently introduced high-pressure electronegativity scale [44,45].

On top of these general trends, it is interesting to observe that in some cases the structural template also plays a crucial role. The S1-139-A_1H_4, S3-166-A_1H_6, S4-221-A_1H_6, M3-229-A_1H_6, M4-166-A_1H_6, and M6-139-A_1H_4 templates exhibit a small spread in the ρ_H value which is close to 0, in stark contrast with the other templates. Among the S-type templates, which contain H_2 molecules, ρ_H is close to 0 because of the charge neutrality of the hydrogen molecule [46]. On the contrary, among the M-type templates it is the weak-covalent nature of the hydrogen cage which suppresses the ionic exchange between hydrogen and the A atom.

In this dataset we observe that in most of the thermodynamically stable structures the A atom is an alkali metal, alkaline earth or early transition metal, and ρ_H is in the [-0.5: 0.0] range. There are two exceptions to this trend: (i) *p*-block elements, in which structures with positively charged hydrogen exhibit lower formation enthalpy, and (ii) late transition metals, which present a value of ρ_H in the [-0.5: 0.0] range, but are thermodynamically unstable, with a large positive ΔH.

The heat map of the quasi-bulk modulus BM_q shown in Fig. 3(c) summarizes the distribution of elastic properties. BM_q varies over a range from 400 to 1000 GPa. As a representative measure, the well known SH_3, YH_6 and CaH_6 high-T_c superconducting phases have 698, 686, 628 GPa as BM_q, respectively.

Similarly to ΔH, BM_q depends more strongly on the A element than on the specific template. Elements of the *p*- and *d*-block (Ti-Zn) and (Zr-Cd) exhibit BM_q larger than 700 GPa across different templates, while *s*-block elements and $(n-1)d^1$ elements have BM_q in the 500–700 GPa range, suggesting that the type of bonds formed in their binary hydrides are softer.

The same templates for which ρ_H remains close to zero (S1-139-A_1H_4, S3-166-A_1H_6, S4-221-A_1H_6, M3-229-A_1H_6, M4-166-A_1H_6, and M6-139-A_1H_4) are an exception, as their BM_q is significantly lower than other templates with the same elements. This occurs in S- and M-type templates, but never in L-type ones, where the bulk properties are dominated by the A element. In the following, we will refer to these templates collectively as *soft* templates.

Figure 3(d) divides compounds in DB2 into low-T_c (blue) and high T_c (red) ones. Gray-colored boxes indicate either nonmetallic or metallic dynamically unstable structures. A significant portion of compounds in the `Superhydra` dataset hosts superconductivity (139 out of 880, i.e., 16%), but only a relatively small fraction of those exhibits high T_c. (27, i.e., 19%).

Reassuringly, among the high-T_c structures we find well-known superhydrides: SH_3 (M7-229-A_1H_3) [1,2], SeH_3 (M7-229-A_1H_3) [30], MgH_6 (M3-229-A_1H_6) [31], CaH_6 (M3-229-A_1H_6) [32,39], GeH_3 (M1-223-A_1H_3, A15) [47], YH_6 (M3-229-A_1H_6) [28,29], and NaH_6 (S3-166-A_1H_6) [18]. The same templates also exhibit a high-T_c with other A elements, such as NbH_6 (M3-229-A_1H_6) and ScH_6 (M3-229-A_1H_6), and a few templates where high-T_c superconductivity has not been reported before, exhibiting a surprising high-T_c behavior, such as AsH_2 (S2-166-A_1H_2) and SiH_3 (M5-221-A_1H_3). A complete list of the high-T_c structures is shown in Table I.

The majority of the high-T_c structures are formed by elements of the *s*-block, and M-type templates, where the formation enthalpy is negative. These are also regions for which the partial HDOS at the Fermi level is the largest (See Fig. 4 in the SM [23]). In particular, the M3-229-A_1H_6 and M4-166-A_1H_6 templates are the most favorable, hosting 5 and 6 high-T_c structures, respectively. In addition, although structures containing H_2 molecules are typically insulating, we do find a few high-T_c structures for templates with H_2 molecules: S3-166-A_1H_6 (Na, Sc, Mo), S2-166-A_1H_2 (As), and S4-221-A_1H_6 (Y), as well as L1-221-A_1H_1 (Li) and L2-225-A_1H_1 (Be). It is interesting to note that, although the starting S2-166-A_1H_2, S3-166-A_1H_6 and S4-221-A_1H_6 templates are S-type, the corresponding high-T_c structures formed with different A elements (listed in Table I) are not classified in the H_2 cluster, with the exception of

TABLE I. High-T_c binary hydrides in DB2 ($T_c \geqslant 77$ K). In some cases, the M3-229-A_1H_6 and M4-166-A_1H_6 relax to the same structure, as indicated by the asterisks enlisted at the bottom of the table.

Template	Cluster	Formula	HDOS-Frac	λ	$\omega_{\log}$ (K)	T_c (K)
S2-166-A_1H_2	HTC	AsH_2	0.28	1.50	720	82
M1-223-A_1H_3	HTC	GaH_3	0.69	0.96	1258	83
M6-139-A_1H_4	HTC	YH_4	0.44	0.98	1289	87
L2-225-A_1H_1	HTC	BeH_1	0.22	1.68	724	91
L1-221-A_1H_1	HTC	LiH_1	0.86	8.21	386	91
S3-166-A_1H_6	H_2	NaH_6	0.90	2.95	516	92
M3-229-A_1H_6	TMs	NbH_6	0.26	2.52	580	96
S3-166-A_1H_6	HTC	ScH_6	0.40	1.38	938	98
S4-221-A_1H_6	HTC	YH_6	0.26	2.69	611	104
M7-229-A_1H_3	HTC	SeH_3	0.47	1.02	1461	105
L3-064-A_1H_2	HTC	ClH_2	0.55	1.28	1090	105
S3-166-A_1H_6	TMs	MoH_6	0.48	2.20	763	117
M1-223-A_1H_3	HTC	GeH_3 [47]	0.24	1.56	1037	123
M6-139-A_1H_4	HTC	SrH_4	0.91	2.31	844	133
M4-166-A_1H_6	HTC	SrH_6	0.94	1.42	1262	137
M3-229-A_1H_6	HTC	ScH_6	0.33	1.93	1026	144
M5-221-A_1H_3	HTC	SiH_3	0.38	1.41	1345	144
M3-229-A_1H_6	HTC	YH_6	0.48	1.92	1323	185
M7-229-A_1H_3	HTC	SH_3	0.47	1.85	1371	186
M3-229-A_1H_6	HTC	CaH_6	0.96	2.28	1259	197
M2-191-A_1H_2	HTC	LiH_2	0.99	1.98	1406	201
M3-229-A_1H_6	HTC	MgH_6	0.73	3.02	1260	228
M4-166-A_1H_6*	TMs	NbH_6	0.26	2.52	581	96
M4-166-A_1H_6*	HTC	ScH_6	0.33	1.92	1026	144
M4-166-A_1H_6*	HTC	YH_6	0.48	1.92	1323	185
M4-166-A_1H_6*	HTC	CaH_6	0.96	2.28	1259	197
M4-166-A_1H_6*	HTC	MgH_6	0.72	3.02	1260	228

NaH_6 (S3-166-A_1H_6). This indicates that unbiased classification based on unsupervised clustering has a clear advantage over human assumptions.

To better discuss the correlation between other observables and T_c, it is useful to look at aggregate data. In Fig. 4 we show histograms with the distribution of BM_q, d_{HH}, ρ_H, and HDOS, broken down into low-T_c (blue) and high-T_c (red), with the same meaning as Fig. 3. While low-T_c structures are uniformly distributed with respect to these four variables, it clearly emerges that high-T_c ones are concentrated in a narrow range of parameters. In particular: $500 < BM_q < 700$ GPa, $0.8 < d_{HH} < 1.8$ Å, $-0.4 < \rho_H < 0$ arb. units, and $0.03 < \text{HDOS} < 0.06$ states $\text{eV}^{-1}\text{atom}^{-1}$. In other words, if one or more of these four physical observables are in the right range, this does not automatically imply a high-T_c, but to have high-T_c, the observables have to be in the right range (the conditions are necessary but not sufficient). The main consequence is that, based on these observables, it would be possible to predict whether a structure can be a high-T_c superconductor with high sensitivity, but one must be willing to accept a significant fraction of false positives. Our result, however, confirms that all high-T_c hydride superconductors share similar features: a relatively soft H lattice, where H-H interactions are much weaker than in H_2 molecules, but still strong enough to ensure H-H bonds (BM_q, d_{HH}). This type of bonding becomes possible when enough negative charge is pumped into the hydrogen orbitals from other atoms acting as reservoirs (ρ_H) [21,32]. Moreover, the resulting electronic structure should also ensure that electrons in H bonds are actively involved in the superconductivity pairing (HDOS).

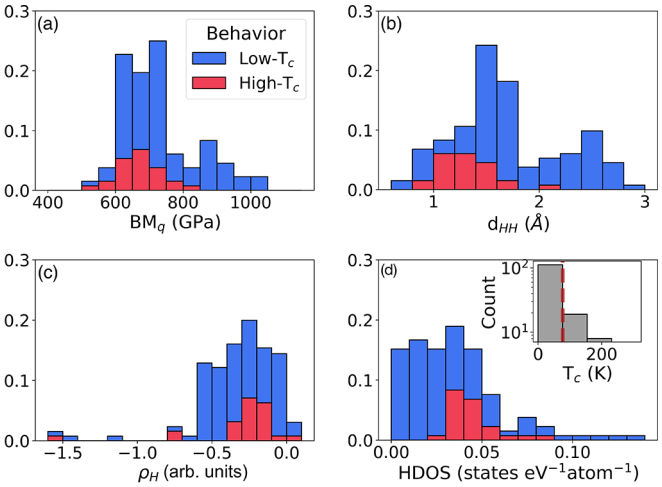

FIG. 4. Histograms of four quantities in DB2: (a) quasi-bulk modulus BM_q, (b) average nearest-neighbor H-H distance d_{HH}, (c) average residual Bader charge on the H atom (ρ_H), and (d) partial H DOS at the Fermi level HDOS (see text). Compounds are divided into low-T_c (blue) and high-T_c (red) structures. Inset: distribution of T_c's in DB2 where the red dashed line indicates the threshold for high-T_c superconductivity, given by liquid nitrogen boiling temperature (77 K).

IV. NEW SUPERCONDUCTORS

In this section, we focus on the vibrational and superconducting properties of the two high-T_c hydrides discovered by our approach that, to the best of our knowledge, are not present in the literature: AsH_2 (S2-166-A_1H_2) and SiH_3 (M5-221-A_1H_3). For these two structures we performed an accurate study of the electron-phonon properties, including a detailed convergence check. This study serve two purposes: characterize the two newfound superconductors, and allow us to estimate the accuracy of the high-throughput e-ph calculations (details on the convergence are provided in the SM [23]).

In Fig. 5 we show the total and atom-projected Éliashberg spectral function $\alpha^2F(\omega)$ at 200 GPa, along with the electron-phonon coupling $\lambda(\omega)$, for AsH_2 (S2-166-A_1H_2) and SiH_3 (M5-221-A_1H_3). In both cases the Éliashberg function is dominated by hydrogen modes, and the coupling is spread over the whole phonon spectrum. However, in SiH_3 (M5-221-A_1H_3) some phonon modes exhibit a rather strong renormalization due to the strong electron-phonon coupling (see Figs. 13 and 15 for the phonon dispersions in the SM [23]). The two phases are strongly-coupled, hydride superconductors, with metallic covalent bonds (As–H and Si–H), akin to H_3S (see Figs. 10 and 12 for the electronic structure in the SM [23]).

From the Éliashberg function we calculated λ and $\omega_{\log}$. In AsH_2 (S2-166-A_1H_2) we obtain $\lambda = 1.55$ and $\omega_{\log} = 691$ K, which results in a T_c of 81 K using the McMillan-Allen-Dynes formula, while in SiH_3 (M5-221-A_1H_3), we obtain $\lambda = 1.28$ and $\omega_{\log} = 1307$ K, and a T_c of 126 K. The converged results

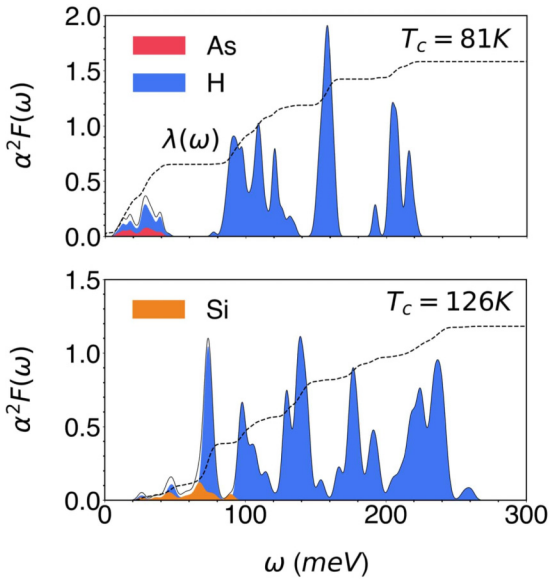

FIG. 5. Total (solid black line) and atom-projected (colored filled lines) Éliashberg function, along with the electron-phonon coupling $\lambda(\omega)$ (dashed black line) of AsH_2 (S2-166-A_1H_2) (top panel) and SiH_3 (M5-221-A_1H_3) (bottom panel) at 200 GPa. Projections onto As, H, and Si are shown in red, blue, and orange, respectively.

are in good agreement with those listed in Table I for both AsH_2 (S2-166-A_1H_2) and SiH_3 (M5-221-A_1H_3): T_c and λ are both within 10% of the converged values, which is an excellent accuracy considering the high-throughput nature of our work.

V. CONCLUSIONS

In summary, in this paper we introduced an adapted high-throughput approach for the computational generation of databases of superconductors, and applied it to binary hydrides at high pressures.

Our adapted high-throughput approach is based on the generation of template structures with crystal structure prediction methods, carefully selected to be (i) as chemically diverse as possible and (ii) high-symmetry. A database of structures is then generated by substituting elements from the whole periodic table in the nonhydrogen site of all structural templates.

The resulting `Superhydra` database contains 880 potential hydride superconductors (DB1), characterized by various electronic, vibrational, and chemical properties; the 139 dynamically stable and metallic compounds form a database of superconductors (DB2), for which T_c was computed at the McMillan-Allen-Dynes level. All data are available as a single CSV file in the SM [23].

An illustrative data-driven analyses demonstrate that both DB1 and DB2 are information-rich and provide deep physical insight into the chemistry of high-T_c superhydrides. In particular, we find that four observables, namely, the shortest H-H distance, the quasi-bulk modulus, the average residual Bader charge on hydrogen, and the hydrogen fraction of the DOS at the Fermi level (d_{HH}, BM_q, ρ_H, and HDOS, respectively) provide a set of necessary conditions for superconductivity. Based on unsupervised clustering, we propose a classification into three, physically-meaningful classes characterized by (1) short H-H bonds (H_2), (2) presence of *p*- and *d*-block elements (TMs), and (3) a soft lattice with BM_q lower than 800 GPa (HTC). The HTC cluster, in particular, hosts the 85% of the predicted high-T_c superconductors.

We believe that the adapted HT approach presents several advantages towards the generation of a database of conventional superconductors, compared to other approaches available in literature [18,21]. First, it is computationally efficient, as it avoids generating redundant data around ground-state structures that is typical of crystal structure prediction methods [22]. Second, it is easy to generalize by adding any number of templates to the basis. Third, it represents a much more complete exploration of the possible chemical interactions. Fourth, it effectively favors superconductors: 16% of all generated structures are superconductors, and 3% are high-T_c, about two orders of magnitude larger than one would have expected for structures generated with a structural search [18].

Overall, the adapted HT approach has proven to be a very effective method to map the phase space of binary hydrides and could be generalized to the more complex space of ternary hydrides, to provide much needed insight on their chemistry. As we have shown, a careful choice of the initial templates is paramount to this end. Extension to other classes of conventional superconductors is also straightforward, provided a more general definition of features. Our hope is that the `Superhydra` database will represent a much-needed first step of a generalized effort to map conventional superconductivity across the material space.

ACKNOWLEDGMENTS

S.S., S.D.C., and W.v.d.L. acknowledge computational resources from the dCluster of the Graz University of Technology, the VSC3 of the Vienna University of Technology, and support through the FWF, Austrian Science Fund, Project P 30269-N36 (Superhydra). S.D.C. also acknowledges computational resources from CINECA, proj. IsC90_HTS-TECH_C. L.B. acknowledges support from Fondo Ateneo Sapienza 2018-20 and computational resources from CINECA, Project Hi-TSEPH.

APPENDIX: COMPUTATIONAL METHODS

1. First-principles Calculations

First-principles calculations based on density functional theory (DFT) was used for (i) structural predictions and (ii) adapted high-throughput runs. Both types of runs were carried out using the Perdew-Burke-Ernzerhof (PBE) exchange-correlation (xc) functional [48], with an external pressure of 200 GPa.

For the structural prediction runs, we used the plane-wave based Vienna *ab initio* simulation package [49]. The atoms were described by the inbuilt Projector Augmented Wave potentials [50,51]. The geometric relaxations were carried out through multistep runs with increasing kinetic energy cutoffs of 400 and 500 eV. Further details can be found in the SM [23].

For the high-throughput runs, we used the freely available plane-wave-based `Quantum-Espresso`-6.5 [52] package; atoms were described by Optimized Norm Conserving Vanderbilt (ONCV) pseudopotentials [53], with a kinetic energy cutoff of 90 Ryd. The Brillouin zone integration employed Γ-centered Monkhorst-Pack (MP) $8 \times 8 \times 8$ electronic grids (**k**), with a Methfessel-Paxton smearing of 0.04 Ry. Further details are provided in the SM [23].

2. Clustering

Data analysis and clustering was performed using the pandas and keras python libraries. The number of clusters for the k-means algorithm was decided by examining the dendrogram shown in Fig. 10 in the SM [23], constructed using agglomerative clustering with Ward linkage [54]. This led us to our choice of $k = 3$. The other possibly sensible choice ($k = 6$) led to less physically meaningful results, and was discarded. The k-means clustering was then performed on the whole DB1. This algorithm partitions the data into k clusters, each centered around a mean value, as to minimize the variance within each cluster. Conceptually, the algorithm is particularly suitable when the data can be divided into groups such that each group differs significantly from the others, while data points within each group are similar.

3. Classification

To classify superconductors into high- and low-T_c, we used a random forest (RF) classifier, as implemented in sklearn. RF is a powerful ensemble method based on simple decision trees. The choice of this algorithm was motivated by its rather good robustness against overfitting, and the need for a simple model, which could be trained sufficiently well on a little over than 100 data points. We employed a randomized 80/20 split in training and test set, with stratified sampling owing to the significant class imbalance. Statistical significance of the performance metrics was ensured by averaging over 100 iterations.

As a first step, we trained an initial model including all the features in our database, except those directly related to the critical temperature by Migdal-Éliashberg theory (T_c, λ, $\omega_{\log}$), and analyzed the importance of each feature, based on the Gini impurity decrease at each split, averaged over 2000 trees. The feature importance analysis is shown in Fig. 18 in the SM [23]. The final choice of features (see Sec. II C) was motivated by a combination of this analysis and a physical meaning of the features. In addition, we weighted the minority class (high-T_c) 20 to 1, as to compensate for the class imbalance, and improve the model recall at the expense of precision.

As a second step, we optimized the hyperparameters using simple grid search. We found satisfactory results with a maximum tree depth of 5, a minimum number of samples per split of 4, and a number of estimators equal to 500, and a class weight of 20 to 1 between the high- and low-T_c class. This adjustment is intended to bias the model towards sensitivity to high-T_c structures, at the cost of precision, and compensates for the class imbalance of about 1 to 10 between high- and low-T_c structures.

[1] 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 (2021).

[2] A. Drozdov, M. Eremets, I. Troyan, V. Ksenofontov, and S. Shylin, Nature (London) **525**, 73 (2015).

[3] E. Snider, N. Dasenbrock-Gammon, R. McBride, M. Debessai, H. Vindana, K. Vencatasamy, K. V. Lawler, A. Salamat, and R. P. Dias, Nature (London) **586**, 373 (2020).

[4] A. P. Drodzov, P. Kong, V. S. Minkov, S. P. Besedin, M. A. Kuzovnikov, S. Mozaffari, L. Balicas, F. F. Balakirev, D. E. Graf, V. B. Prakapenka *et al.*, Nature (London) **569**, 528 (2019).

[5] M. Somayazulu, M. Ahart, A. K. Mishra, Z. M. Geballe, M. Baldini, Y. Meng, V. V. Struzhkin, and R. J. Hemley, Phys. Rev. Lett. **122**, 027001 (2019).

[6] S. Di Cataldo, C. Heil, W. von der Linden, and L. Boeri, Phys. Rev. B **104**, L020511 (2021).

[7] R. Lucrezi, S. Di Cataldo, W. von der Linden, L. Boeri, and C. Heil, Npj Comput. Mater. **8**, 119 (2022).

[8] S. Di Cataldo, W. von der Linden, and L. Boeri, Npj Comput. Mater. **8**, 2 (2021).

[9] S. Saha, S. D. Cataldo, M. Amsler, W. von der Linden, and L. Boeri, Phys. Rev. B **102**, 024519 (2020).

[10] J. A. Flores-Livas, L. Boeri, A. Sanna, G. Profeta, R. Arita, and M. Eremets, Phys. Rep. **856**, 1 (2020).

[11] B. Lilia, R. G. Hennig, P. J. Hirschfeld, G. Profeta, A. Sanna, E. Zurek, W. E. Pickett, M. Amsler, R. Dias, M. Eremets *et al.*, J. Phys.: Condens. Matter **34**, 183002 (2022).

[12] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nat. Mater. **12**, 191 (2013).

[13] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).

[14] H. Glawe, A. Sanna, E. Gross, and M. A. Marques, New J. Phys. **18**, 093011 (2016).

[15] J. Schmidt, L. Chen, S. Botti, and M. A. Marques, J. Chem. Phys. **148**, 241728 (2018).

[16] S. D. Griesemer, L. Ward, and C. Wolverton, Phys. Rev. Mater. **5**, 105003 (2021).

[17] H.-C. Wang, S. Botti, and M. A. Marques, Npj Comput. Mater. **7**, 1 (2021).

[18] A. M. Shipley, M. J. Hutcheon, R. J. Needs, and C. J. Pickard, Phys. Rev. B **104**, 054501 (2021).

[19] D. V. Semenok, I. A. Kruglov, I. A. Savkin, A. G. Kvashnin, and A. R. Oganov, Curr. Opin. Solid State Mater. Sci. **24**, 100808 (2020).

[20] E. Zurek, Comments Inorg. Chem. **37**, 78 (2017).

[21] F. Belli, T. Novoa, J. Contreras-García, and I. Errea, Nat. Commun. **12**, 5381 (2021).

[22] F. M. Rohrhofer, S. Saha, S. Di Cataldo, B. C. Geiger, W. von der Linden, and L. Boeri, arXiv2102.00191 (2021).

[23] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.7.054806 for all the information on DB1, DB2, a2F spectral functions and structure templates. Further a pdf file is provided, detailing all the procedure

followed to generate and analyse the database. The Supplemental Material also contains Refs. [55–60].
[24] W. L. McMillan, Phys. Rev. **167**, 331 (1968).
[25] A. Flores-Livas, José, A. Sanna, and S. Goedecker, Nov. Supercond. Mater. **3**, 6 (2017).
[26] Y. Zhang, H. Sun, and C. Chen, Phys. Rev. B **73**, 144115 (2006).
[27] E. Zurek, R. Hoffmann, N. W. Ashcroft, A. R. Oganov, and A. O. Lyakhov, Proc. Natl. Acad. Sci. USA **106**, 17640 (2009).
[28] C. Heil, S. di Cataldo, G. B. Bachelet, and L. Boeri, Phys. Rev. B **99**, 220502(R) (2019).
[29] P. Kong, V. S. Minkov, M. A. Kuzovnikov, A. P. Drodzov, S. P. Besedin, S. Mozaffari, L. Balicas, F. F. Balakirev, V. B. Prakapenka, S. Chariton, D. A. Knyazev, E. Greenberg, and M. I. Eremets, Nat. Commun. **12**, 5075 (2021).
[30] C. Heil and L. Boeri, Phys. Rev. B **92**, 060508(R) (2015).
[31] X. Feng, J. Zhang, G. Gao, H. Liu, and H. Wang, RSC Adv. **5**, 59292 (2015).
[32] F. Peng, Y. Sun, C. J. Pickard, R. J. Needs, Q. Wu, and Y. Ma, Phys. Rev. Lett. **119**, 107001 (2017).
[33] C. J. Pickard and R. J. Needs, Phys. Rev. B **76**, 144114 (2007).
[34] C. M. Pépin, A. Dewaele, G. Geneste, P. Loubeyre, and M. Mezouar, Phys. Rev. Lett. **113**, 265504(R) (2014).
[35] C. Heil, G. B. Bachelet, and L. Boeri, Phys. Rev. B **97**, 214510 (2018).
[36] T. Bi and E. Zurek, Chem. Eur. J. **27**, 14858 (2021).
[37] C. W. Glass, A. R. Oganov, and N. Hansen, Comput. Phys. Commun. **175**, 713 (2006).
[38] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, Npj Comput. Mater. **4**, 29 (2018).
[39] H. Wang, J. S. Tse, K. Tanaka, T. Iitaka, and Y. Ma, Proc. Natl. Acad. Sci. USA **109**, 6463 (2012).
[40] D. Pettifor, J. Phys. C: Solid State Phys. **19**, 285 (1986).
[41] The choice of $k = 3$ or $k = 6$ is supported by a hierarchical clustering analysis (See Fig. 10 in SM the [23]).
[42] L. Breiman, Mach. Learn. **45**, 5 (2001).
[43] L. Pauling, J. Am. Chem. Soc. **54**, 3570 (1932).
[44] M. Rahm, R. Cammi, N. W. Ashcroft, and R. Hoffman, J. Am. Chem. Soc. **141**, 10253 (2019).
[45] X. Dong, A. R. Oganov, H. Cui, X.-F. Zhou, and H.-T. Wang, Proc. Natl. Acad. Sci. USA **119**, e2117416119 (2022).
[46] The S2-166-A_1H_2 template is an exception, as H_2 molecules and isolated H atoms coexist, which is reflected in its Bader charge that differs from other S-like templates.
[47] K. Abe and N. W. Ashcroft, Phys. Rev. B **88**, 174110 (2013).
[48] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[49] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).
[50] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[51] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).
[52] bibfnamefont 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).
[53] D. R. Hamann, Phys. Rev. B **88**, 085117 (2013).
[54] J. H. Ward Jr., J. Am. Stat. Assoc. **58**, 236 (1963).
[55] A. R. Oganov and C. W. Glass, J. Chem. Phys. **124**, 244704 (2006).
[56] A. O. Lyakhov and A. R. Oganov, Phys. Rev. B **84**, 092103 (2011).
[57] A. R. Oganov, A. O. Lyakhov, and M. Valle, Acc. Chem. Res. **44**, 227 (2011).
[58] W. Tang, E. Sanville, and G. Henkelman, J. Phys.: Condens. Matter **21**, 084204 (2009).
[59] A. Dal Corso, Comput. Mater. Sci. **95**, 337 (2014).
[60] G. Eliashberg, J. Exp. Theor. Phys. **11**, 696 (1960).