Mapping Superconductivity in High-Pressure Hydrides: The Superhydra Project.

Santanu Saha,1 Simone Di Cataldo,1,2 Federico Giannessi,2 Alessio Cucciari,2 Wolfgang von der Linden,1 and Lilia Boeri2,3
1Institute of Theoretical and Computational Physics, Graz University of Technology, NAWI Graz, 8010 Graz, Austria
2Dipartimento di Fisica, Università di Roma La Sapienza, Piazzale Aldo Moro 5, I-00185 Roma, Italy
3Centro Ricerche Enrico Fermi, Via Panisperna 89 A, 00184 Rome, Italy

(Dated: May 6, 2022)

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 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 (MultiTemplate-HighThroughput 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. 139 structures of our Superhydra Database 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.

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.12

After reaching the symbolical threshold of room-temperature16,17, the research focus is rapidly shifting to the even more ambitious goal of conventional superconductivity at ambient condition.26–28 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.20,11 high-throughput (HT) screening and artificial intelligence techniques will likely play a prominent role in this search.16,17

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 literature16,23 or randomly-generated structures.11,12 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 – MultiTemplate-HighThroughput approach (MTHT) – which is explicitly designed to realize the best possible compromise between variety, reproducibility, and efficiency. By utilizing the same set of high-symmetric, representative templates for all elements of the periodic table, we 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.

As a first application, the MTHT 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 SM20 comprises a total of 880 distinct binary hydride structures, of which 139 are 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 analyse 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.

This paper is structured as follows: in the first section 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 the next section, we discuss the main features of the binary hydrides material space, using material maps of the most relevant quantities. The main computational details are summa-
Figure 1. Template structures used for multi-template-high-throughput data generation, classified by H-H bond distance into short (S), medium (M), and long (L). The label \( XN_{n_H} \) indicates the H-H distance type \( (X) \), the structure number \( (N) \), and the number of hydrogen atoms in the structure \( (subscript n_H) \) Details of the crystal structures are given in the SI.

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 sixteen 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.

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, 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 \( \mu^* = 0 \).[24] These compounds form our database of superconductors (DB2). 27 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$_2$; systems that form H$_2$-rich structures like TeH$_2$ and LiH$_6$; cage-like structures, as in sodalite-clathrate hydrides like YH$_6$. The elements were chosen to exhibit different chemical properties and high pressure behaviours. As for compositions, we restricted our search to compounds with chemical formulas AH$_n$, with $n \leq 6$, 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$_3$, and YH$_3$, 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 SM 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 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 ≤ $d_{HH} < 2$ Å); long ($d_{HH} > 2$ Å).

The final set of sixteen templates was obtained 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.

The sixteen templates are displayed in Fig. 1. In the following, they will be represented by an alphanumeric id of the form $X N_{nH}$ where $X$ stands for short(S), medium(M) or long(L) type based on the smallest $d_{HH}$; $N$ represents the serial number and the subscript $n_H$, the number of H atoms in its chemical formula. Hence, e.g. M46 would represent the 4$^{th}$ structure of medium type with composition AH$_4$. Among the templates we find well-known crystal structure types. In particular, the M73 and the M36 structures correspond to the high-$T_c$ SH$_3$ and AH$_6$ ($A = Mg, Ca, Y$) templates[12][22][23] while M13 and M53 correspond to the high-pressure phases of AlH$_3$ and FeH$_3$[33][35] respectively. L1$_4$ and L2$_1$ coincide with the CsCl and NaCl structures, and S5$_4$ with the predicted LiH$_6$ structure[22].

### 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.

1. **Geometry - space group of the primitive unit cell (SPG); number of hydrogen atoms per formula unit ($n_H$); Volume(volume); Density ($\rho_V$); Structural Template ($X N_{nH}$); nearest-neighbour 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 ($\Delta H$ 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[30].

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 ($\rho_A$); partial charge on atom H ($\rho_H$), in a.u. - average, maximum, minimum value, standard deviation (avg, max, min, dev).

5. **Elastic** - The elastic properties of the system are represented through quasi Bulk modulus estimated at 200 GPa (BM$_q$), in GPa. The definition of the BM$_q$ is provided in the SM.

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

7. **Electron-phonon ($e$ – $ph$)** - These properties have been evaluated for metallic dynamically stable structures only. We report the logarithmically-averaged phonon frequency ($\omega_{\text{avg}}$) in $K$, the total electron-phonon coupling constant ($\lambda$) 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.[13][21][37].

For example, it was early recognized that in order 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\textsuperscript{29}, and the hydrogen sublattice should be negatively charged\textsuperscript{45}. 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\textsuperscript{10}, while Belli et al. divided the known superhydrides in six classes (molecular, covalent, weak H interaction, electrode, isolated and ionic), using an inspection of interatomic distances and electronic localization function profile\textsuperscript{39}. 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\textsuperscript{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\textsuperscript{10}. We name the three clusters thus identified H\textsubscript{2}, TMs, and HTC, based on the type of hydrides they comprise. The H\textsubscript{2} cluster contains primarily structures with short H-H bonds ($d_{HH} < 0.8$ Å), i.e. molecular H\textsubscript{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 $d_{HH} > 0.8$ 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\textsubscript{2}; and 11\% (3) to TMs. The outlying high-$T_c$ structures are S\textsubscript{3}H\textsubscript{6}-NaH\textsubscript{6} in the H\textsubscript{2} cluster, and S\textsubscript{3}H\textsubscript{6}-MoH\textsubscript{6}, M\textsubscript{3}H\textsubscript{6}-NbH\textsubscript{6}, and M\textsubscript{4}H\textsubscript{6}-NbH\textsubscript{6} in TMs. Known high-$T_c$ phases, such as SH\textsubscript{3}, NaH\textsubscript{6} and YH\textsubscript{6} all belong to the HTC cluster.

Fig. 2 shows a scatter plot of the data in DB1 in the atomic number ($Z$) vs H-H distance $d_{HH}$, colored according to the assignment to one of three clusters: H\textsubscript{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 first-row $p$-, and $d$-blocks of the periodic table. Upon closer inspection, we note that also a few data points in the H\textsubscript{2} cluster have $d_{HH} > 0.8$ 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-$T_c$ behaviour. To this end, we performed a feature importance analysis after training a simple Random Forest Classifier\textsuperscript{41} 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 $F_1$ of 0.79, and an accuracy of 0.90. This result is comparable with similar studies when the dataset had a comparable size\textsuperscript{37}, 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\textsubscript{i}, $Z$, $\omega_{min}$, $\rho_H$, and $\Delta H$. 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 ($\rho_H$) and the quasi-solid modulus BM\textsubscript{q}, this correlation is less obvious. However, we will show in the next section 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.
Figure 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 a.u.; (b) Formation Enthalpy $\Delta H$ (eV/atom) estimated w.r.t stable elemental phases; (c) average quasi bulk-modulus (see SM) 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.
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 properties: from top to bottom and left to right, we show the average Bader charge on the H atom ($\rho_H$), the formation enthalpy ($\Delta H$), 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. 1-8 in SM. The choice of these four observables is motivated as follows: $\Delta H$ gives information on the degree of chemical compatibility between the structure and the template, $\rho_H$ 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 ($\Delta H$) heatmap describes the thermodynamic stability of different hydrate 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 ($\Delta H \leq 0$ eV/atom) across different templates.

These results are fairly consistent with available literature, reporting that high-pressure hydrides are mostly formed by electropositive elements w.r.t. H, as described by Pauling electronegativity. 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 $\Delta H$, chemical properties, and superconducting behavior. On the other hand, 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 $\text{Si}_4$, $\text{S}_4$, and $\text{M}_7$ are strongly favored by specific A elements. For instance, the $\text{SH}_3$-like template (M7) 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 $\rho_H$ 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 2nd 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 $\Delta H$ 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.

On top of these general trends, it is interesting to observe that in some cases the structural template also plays a crucial role. The $\text{Si}_4$, $\text{S}_4$, $\text{M}_3$, $\text{M}_4$ and $\text{M}_6$ templates exhibit a small spread in the $\rho_H$ value which is close to 0, in stark contrast with the other templates. Among the $S$-type templates, which contain H molecules, $\rho_H$ is close to 0 because of the charge neutrality of the hydrogen molecule. 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 $\rho_H$ 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 $\rho_H$ in the [−0.5: 0.0] range, but are thermodynamically unstable, with a large positive $\Delta H$.

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 $\text{SH}_3$, $\text{YH}_6$ and $\text{CaH}_6$ high-$T_c$ superconducting phases have 698, 686, 628 GPa as $BM_q$, respectively.

Similarly to $\Delta H$, $BM_q$ 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 $\rho_H$ remains close to zero ($\text{Si}_4$, $\text{S}_4$, $\text{M}_3$, $\text{M}_4$ and $\text{M}_6$) 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.

Panel (d) of Fig. 3 divides compounds in DB2 into low-$T_c$ (blue) and high-$T_c$ (red) ones. Grey-coloured boxes indicate either non-metallic or metallic dynamically unstable structures. A significant portion of compounds in the Superhydral 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: $M_73$-SiH$_6$[12], $M_7$-SeH$_6$[10], $M_6$-MgH$_6$[31], $M_6$-CaH$_6$[22,28], $M_5$-YH$_3$[29,39], $S_3$-$NaH_4$[15]. The same templates also exhibit a high-$T_c$ with other A elements, such as $M_3$-NbH$_6$ and ScH$_6$, and a few templates where high-$T_c$ superconductivity has not been reported before exhibiting a surprising high-$T_c$ behavior, such as $M_53$-SiH$_3$, $S_2$-$AsH_2$, M$_1$3-GeH$_3$. A complete list of the high-$T_c$ structures is shown in Tab. 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 SM). In particular, the M$_3$ and M$_4$ 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: $S_3$-$Na$, $S_3$-$Sc$, $S_3$-$Mo$, $S_2$-$As$, and $S_4$-$Y$, as well as $L_11$-$LiH$ and $L_21$-$BeH$. It is interesting to note that, although the starting $S_2$, $S_3$, and $S_4$ templates are S-type, the corresponding high-$T_c$ structures formed with different A elements listed in Tab. I are not classified in the H$_2$ cluster, with the exception of $S_3$-$NaH_4$. This indicates that unbiased classification based on unsupervised clustering has a clear advantage over human assumptions.

| TEMPLATE | Cluster | Formula | HDOS | $\lambda$ | $\omega_{\log}$ (K) | $T_c$ (K) |
|----------|---------|---------|------|-----------|------------------|-----------|
| S$_2$    | HTC     | AsH$_2$ | 0.28 | 1.50      | 720              | 82        |
| M$_1$    | HTC     | GaH$_3$ | 0.69 | 0.96      | 1258             | 83        |
| M$_6$    | HTC     | YH$_4$  | 0.44 | 0.98      | 1289             | 87        |
| L$_1$    | HTC     | BeH$_1$ | 0.22 | 1.68      | 724              | 91        |
| L$_11$   | HTC     | LiH$_2$ | 0.86 | 8.21      | 386              | 91        |
| S$_3$    | H$_2$   | NaH$_6$ | 0.90 | 2.95      | 516              | 92        |
| M$_6$    | TMs     | NbH$_6$ | 0.26 | 2.52      | 580              | 96        |
| M$_4$    | TMs     | NbH$_6$ | 0.26 | 2.52      | 581              | 96        |
| S$_3$    | HTC     | ScH$_5$ | 0.40 | 1.38      | 938              | 98        |
| S$_4$    | HTC     | YH$_5$  | 0.26 | 2.69      | 611              | 104       |
| M$_7$    | HTC     | ScH$_6$ | 0.87 | 1.02      | 1461             | 105       |
| S$_3$    | TMs     | MoH$_6$ | 0.48 | 2.20      | 763              | 117       |
| M$_1$    | HTC     | GeH$_3$ | 0.24 | 1.56      | 1037             | 123       |
| M$_6$    | HTC     | SrH$_4$ | 0.91 | 2.31      | 844              | 133       |
| M$_6$    | HTC     | SrH$_4$ | 0.94 | 1.42      | 1262             | 137       |
| M$_6$    | HTC     | ScH$_6$ | 0.33 | 1.93      | 1026             | 144       |
| M$_4$    | HTC     | ScH$_6$ | 0.33 | 1.92      | 1026             | 144       |
| M$_5$    | HTC     | SiH$_6$ | 0.38 | 1.41      | 1345             | 144       |
| M$_4$    | HTC     | YH$_6$  | 0.48 | 1.92      | 1323             | 185       |
| M$_3$    | HTC     | YH$_6$  | 0.48 | 1.92      | 1323             | 185       |
| M$_7$    | HTC     | SH$_3$  | 0.47 | 1.85      | 1371             | 186       |
| M$_5$    | HTC     | CaH$_6$ | 0.96 | 2.28      | 1259             | 197       |
| M$_6$    | HTC     | CaH$_6$ | 0.96 | 2.28      | 1259             | 197       |
| M$_2$    | HTC     | LiH$_2$ | 0.99 | 1.98      | 1406             | 201       |
| M$_3$    | HTC     | MgH$_6$ | 0.73 | 3.02      | 1260             | 228       |
| M$_4$    | HTC     | MgH$_6$ | 0.72 | 3.02      | 1260             | 228       |

Table I. High-$T_c$ binary hydrides in DB2 ($T_c \geq 77$ K).

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$_q$, $d_{HH}$, $\rho_H$, and HDOS, broken down into low-$T_c$ (blue), and high-$T_c$ (red), with the same meaning as the previous figure. 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, $0.8 < d_{HH} < 1.8$, $-0.4 < \rho_H < 0$ a.u., and $0.03 <$ HDOS $< 0.06$ states eV$^{-1}$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 in order 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 ($\rho_H$). Moreover, the resulting electronic structure should also ensure that electrons in H bonds are actively involved in the superconductivity pairing (HDOS).

IV. CONCLUSIONS

In summary, in this paper we introduced a high-throughput approach for the computational generation of databases of superconductors, and applied it to binary hydrides at high pressures. Our MultiTemplate-HighThroughput 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 non-hydrogen site of all structural templates. The resulting Superhydro 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.

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 ($d_{HH}$,
Figure 4. Histograms of four quantities in DB2: quasi-bulk modulus BM\textsubscript{q}, average nearest-neighbour H-H distance d\textsubscript{HH}, average residual Bader charge on the H atom (\(\rho\textsubscript{H}\)), and partial H DOS at the Fermi level HDOS (see text). Compounds are divided into low-\(T\textsubscript{c}\) (blue) and high-\(T\textsubscript{c}\) (red) structures. Inset: distribution of \(T\textsubscript{c}\)'s in DB2 the red dashed line indicates the threshold for high-\(T\textsubscript{c}\) superconductivity, given by liquid nitrogen boiling temperature (77 K).

BM\textsubscript{q}, \(\rho\textsubscript{H}\), 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\textsubscript{2}), (2) presence of \(p\)- and \(d\)-block elements (TMs) and (3) a soft lattice with BM\textsubscript{q} lower than 800 GPa (HTC). The HTC cluster, in particular, hosts the 85\% of the predicted high-\(T\textsubscript{c}\) superconductors.

We believe that the MTHT approach presents several advantages towards the generation of a database of conventional superconductors, compared to other approaches presented in literature\textsuperscript{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\textsuperscript{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 favours superconductors: 16\% of all generated structures are superconductors, and 3\% are high-\(T\textsubscript{c}\), about two orders of magnitude larger than one would have for structures generated with a structural search\textsuperscript{15}.

Overall, the MTHT 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.

V. METHODS

A. First-principles Calculations

First-principles calculations based on Density Functional Theory (DFT) was used for (i) structural predic-
tions and (ii) multi-template high-throughput runs. Both types of runs were carried out using the Perdew-Burke-Ernzerhof (PBE) exchange-correlation (xc) functions\cite{perdew1996}, with an external pressure of 200 GPa.

For the structural prediction runs, we used the plane-wave based Vienna ab-initio Simulation Package\cite{vasp}. The atoms were described by the inbuilt Projector Augmented Wave potential\cite{pawlowski}. The geometric relaxations were carried out through multistep runs with increasing kinetic energy cutoffs of 400 eV and 500 eV. Further details can be found in the SM.

For the high-throughput runs, we used the freely available plane-wave based Quantum-Espresso-6,\cite{qe} package; atoms were described by Optimized Norm Conserving Vanderbilt (ONCV) pseudopotentials\cite{oncv}, with a kinetic energy cutoff of 90 Ry. The Brillouin zone integration was performed on the whole DB1. This algorithm partitions the data into 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.

\section{Clustering}

Data analysis and clustering was performed using the pandas and keras python libraries. The number of clusters for the \textit{k}-means algorithm was decided by examining the dendrogram shown in Fig. 10 in SM, constructed using agglomerative clustering with Ward linkage\cite{ward}. 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.

\section{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 \((T_c, \lambda, \omega_{\log})\), 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. 11 in SM. The final choice of features (See Section IV B.) 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 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.

\section{ACKNOWLEDGMENTS}

S. Saha, S. Di Cataldo and W. von der Linden 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. Di Cataldo also acknowledges computational resources from CINECA, proj. IsC90. HTS-TECH. C. L. Boeri acknowledges support from Fondo Ateneo Sapienza 2018-20 and computational Resources from CINECA, proj. Hi-TSEPH.

\footnotesize
\begin{itemize}
\item \href{mailto:santanu.saha@tugraz.at}{santanu.saha@tugraz.at}
\item D. Duan, Y. Liu, F. Tian, D. Li, X. Huang, Z. Zhao, H. Yu, B. Liu, W. Tian, and T. Cui, Scientific reports 4, 6968 (2014).
\item A. Drozdov, M. Eremets, I. Troyan, V. Ksenofontov, and S. Shylin, Nature 525, 73 (2015).
\item E. Snider, N. Dasenbrock-Gammon, R. McBride, M. Debsussai, H. Vindana, K. Vencatasamy, K. V. Lawler, A. Salamat, and R. P. Dias, Nature 586, 373 (2020).
\item A. P. Drozdov, 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, M. Tkacz, and M. I. Eremets, Nature 569, 528–531 (2019).
\item M. Somayazulu, M. Ahart, A. K. Mishra, Z. M. Geballe, M. Baldini, Y. Meng, V. V. Struzhkin, and R. J. Hemley, Physical Review Letters 122, 027001 (2019).
\item S. Di Cataldo, C. Heil, W. von der Linden, and L. Boeri, Physical Review B 104 (2021).
\end{itemize}
