Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints

Olexandr Isayev,[†] Denis Fourches,[†] Eugene N. Muratov,[†] Corey Oses,[‡] Kevin Rasch,[‡] Alexander Tropsha,*[,†] and Stefano Curtarolo*[,‡,¶]

[†]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, United States

[‡]Center for Materials Genomics, and [¶]Materials Science, Electrical Engineering, Physics and Chemistry, Duke University, Durham, North Carolina 27708, United States

S Supporting Information

ABSTRACT: As the proliferation of high-throughput approaches in materials science is increasing the wealth of data in the field, the gap between accumulated-information and derived-knowledge widens. We address the issue of scientific discovery in materials databases by introducing novel analytical approaches based on structural and electronic materials fingerprints. The framework is employed to (i) query large databases of materials using similarity concepts, (ii) map the connectivity of materials space (i.e., as a materials cartograms) for rapidly identifying regions with unique organizations/properties, and (iii) develop predictive Quantitative Materials Structure−Property Relationship models for guiding materials design. In this study, we test these fingerprints by seeking target material properties. As a quantitative example, we model the critical temperatures of known superconductors. Our novel materials fingerprinting and materials cartography approaches contribute to the emerging field of materials informatics by enabling effective computational tools to analyze, visualize, model, and design new materials.

Materials Cartography
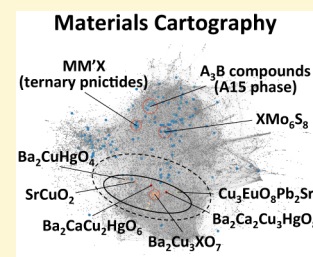
INTRODUCTION

Designing materials with desired physical and chemical properties is recognized as an outstanding challenge in materials research.[1−3] Material properties directly depend on a large number of key variables, often making the property prediction complex. These variables include constitutive elements, crystal forms, and geometrical and electronic characteristics; among others. The rapid growth of materials research has led to the accumulation of vast amounts of data. For example, the Inorganic Crystal Structure Database (ICSD) includes more than 170 000 entries.[4] Experimental data are also included in other databases, such as MatWeb[5] and MatBase.[6] In addition, there are several large databases such as AFLOWL-IB,[7,8] Materials Project,[9] and the Harvard Clean Energy Project[10,11] that contain thousands of unique materials and their theoretically calculated properties. These properties include electronic structure profiles estimated with quantum mechanical methods. The latter databases have great potential to serve as a source of novel functional materials. Promising candidates from these databases may in turn be selected for experimental confirmation using rational design approaches.[12]

The rapidly growing compendium of experimental and theoretical materials data offers a unique opportunity for scientific discovery in materials databases. Specialized data mining and data visualization methods are being developed within the nascent field of materials informatics.[1−3,13−16] Similar approaches have been extensively used in cheminformatics with resounding success. For example, in many cases, these approaches have served to help identify and design small organic molecules with desired biological activity and acceptable environmental/human-health safety profiles.[17−20] Application of cheminformatics approaches to materials science would allow researchers to (i) define, visualize, and navigate through the materials space, (ii) analyze and model structural and electronic characteristics of materials with regard to a particular physical or chemical property, and (iii) employ predictive materials informatics models to forecast the experimental properties of *de novo* designed or untested materials. Thus, rational design approaches in materials science constitute a rapidly growing field.[14−16,21−26]

Herein, we introduce a novel materials fingerprinting approach. We combine this with graph theory, similarity searches, and machine learning algorithms. This enables the unique characterization, comparison, visualization, and design of materials. We introduce the concept and describe the development of materials fingerprints that encode materials' band structures, density of states (DOS), crystallographic, and constitutional information. We employ materials fingerprints to visualize this territory via advancing the new concept of *materials cartography*. We show this technology identifies clusters of specific groups of materials with similar properties. Finally, we develop Quantitative Materials Structure−Property Relationship (QMSPR) models that rely on materials fingerprints. We then employ these models to discover novel

Band Structure Fingerprints
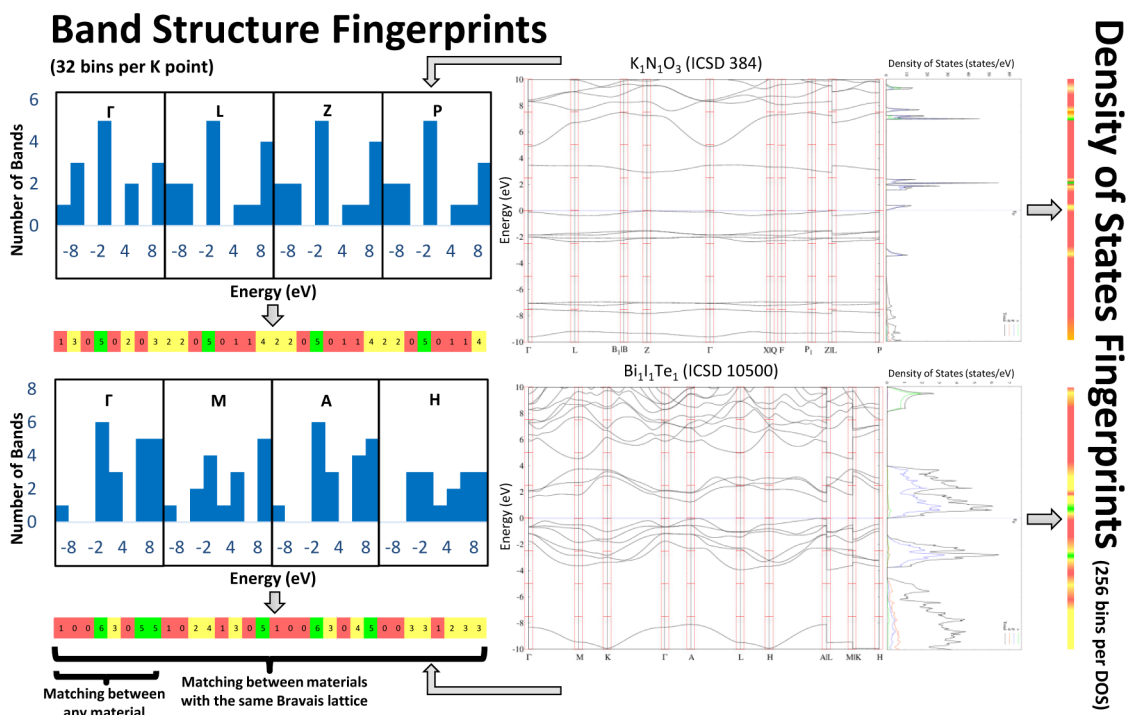
(32 bins per K point)

Figure 1. Construction of materials fingerprints from the band structure and the density of states. For simplicity, we illustrate the idea of B-fingerprints with only 8 bins.

materials with desired properties that lurk within the materials databases.

METHODS

AFLOWLIB Library and Data. AFLOWLIB is a database of density functional theory (DFT) calculations managed by the software package AFLOW.[27,28] At the time of the study, the AFLOWLIB database included the results of calculations characterizing over 20000 crystals, but has since grown to include 50000 entries—representing about a third of the contents of the ICSD.[4] Of the characterized systems, roughly half are metallic and half are insulating. AFLOW leverages the VASP Package[29] to calculate the total energy of a given crystal structure with PAW pseudopotentials[30] and PBE[31] exchange-correlation functional. The entries of the repositories have been described previously.[7,28,32]

Data set of Superconducting Materials. We have compiled experimental data for superconductivity critical temperatures T_c for more than 700 records from the Handbook of Superconductivity[33] and the CRC Handbook of Chemistry and Physics,[34] as well as the SuperCon Database.[35] As we have shown recently,[36] data curation is a necessary step for any Quantitative Structure-Property Relationship (QSPR) modeling. In the compiled data set, several T_c values have been measured under strained conditions, such as different pressures and magnetic fields. We have only kept records taken under standard pressure and with no external magnetic fields. For materials with variations in reported T_c values in excess of 4 K, original references were revisited and records have been discarded when no reliable information was available. T_c values with a variation of less than 3 K have been averaged. Of the remaining 465 materials (T_c range of 0.1–133 K), most records show a variability in T_c of ± 1 K between different sources. Such a level of variability would be extremely influential in materials with low T_c ($T_c < 1$ K) because we have used the decimal logarithm of experimentally measured critical temperature ($\log(T_c)$) as our target property.

To appropriately capture information inherent to materials over the full range of T_c, we have constructed two data sets for the development of three models. The **continuous model** serves to predict T_c and utilizes records excluding materials with T_c values less than 2 K. This

data set consists of 295 unique materials with a $\log(T_c)$ range of 0.30–2.12. The **classification model** serves to predict the position of T_c (above/below) with respect to the threshold T_{thr} (unbiasedly set to 20 K as observed in Figure 4e, see Results and Discussion section). It utilizes records incorporating the aforementioned excluded materials, as well as lanthanum cuprate (La_2CuO_4, ICSD No. 19003). Lanthanum cuprate had been previously discarded for high variability ($T_c = 21-39$ K), but now satisfies the classification criteria. This data set consists of 464 materials (29 with $T_c > T_{thr}$ and 435 with $T_c \leq T_{thr}$). Finally, the **geometrical model** serves to identify structural components that most influence T_c. It utilizes the same data set as the continuous model.

Materials Fingerprints. Following the central paradigms of structure–property relationships, we assume that (i) properties of materials are a direct function of their structure and (ii) materials with similar structures (as determined by constitutional, topological, spatial, and electronic characteristics) are likely to have similar physical and chemical properties.

Thus, encoding material characteristics in the form of numerical arrays of descriptors,[2,15] or *fingerprints*,[37] enables the use of classical cheminformatics and machine-learning approaches to mine, visualize, and model any set of materials. We have encoded the electronic structure diagram for each material as two distinct types of arrays (Figure 1): a *symmetry-dependent fingerprint* (band structure based B-fingerprint) and a *symmetry-independent fingerprint* (density of state based D-fingerprint).

B-Fingerprint. Along every special high-symmetry point of the Brillouin zone (BZ), the energy diagram has been discretized into 32 bins to serve as our fingerprint array. Each BZ has a unique set of high-symmetry points.[7] The comparison set of high-symmetry points belonging to a single BZ type is considered *symmetry-dependent*. To name a few examples, the Brillouin zone path of a cubic lattice ($\Gamma X M \Gamma R X \mid M R$) is encoded with just four points (Γ, M, R, X), giving rise to a fingerprint array of length 128. The body-centered orthorhombic lattice is more complex[7,8] ($\Gamma X L T W R X_1 Z \Gamma Y S W \mid L_1 Y \mid Y_1 Z$) and is represented by 13 points (Γ, L, L_1, L_2, R, S, T, W, X, X_1, Y, Y_1, Z), giving a fingerprint array of length 416. Conversely, the comparison of identical **k**-points not specifically belonging to any BZ is
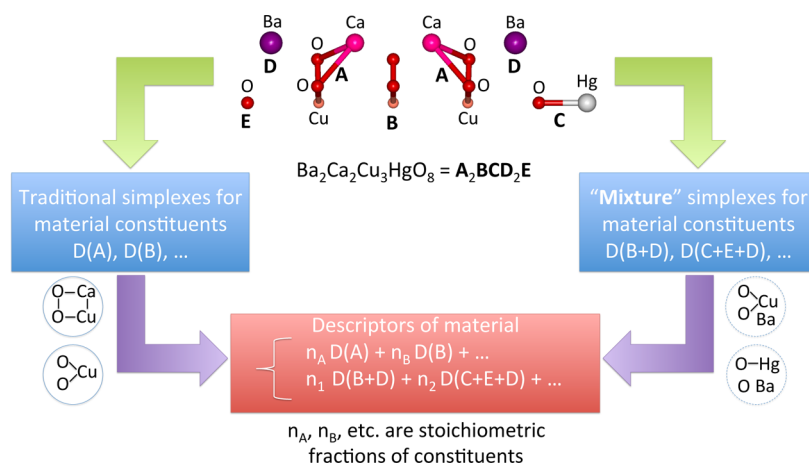
Figure 2. Generation of SiRMS descriptors for materials.

always possible when only restricted to Γ. Consequently, we limit our models to Γ point B-fingerprint in the present work.

D-Fingerprint. A similar approach can be taken for the DOS diagrams, which are sampled in 256 bins (from min to max) and the magnitude of each bin is discretized in 32 bits. Therefore, the D-fingerprint is a total of 1024 bytes. Owing to the complexity and limitations of the *symmetry-dependent* B-fingerprints, we have only generated *symmetry-independent* D-fingerprints. The length of these fingerprints is tunable depending on the objects, applications, and other factors. We have carefully designed the domain space and length of these fingerprints to avoid the issues of enhancing boundary effects or discarding important features.

SiRMS Descriptors for Materials. To characterize the structure of materials from several different perspectives, we have developed descriptors similar to those used for small organic molecules that can reflect their compositional, topological, and spatial (stereochemical) characteristics. Classical cheminformatics tools can only handle small organic molecules. Therefore, we have modified the Simplex (SiRMS) approach[38] based on our experience with mixtures[39,40] in order to make this method suitable for computing descriptors for materials.

The SiRMS approach[38] characterizes small organic molecules by splitting them into multiple molecular fragments called simplexes. Simplexes are tetratomic fragments of fixed composition (1D), topology (2D), and chirality and symmetry (3D). The occurrences of each of these fragments in a given compound are then counted. As a result, each molecule of a given data set can be characterized by its SiRMS fragment profiles. These profiles take into account atom types, connectivity, etc.[38] Here, we have adapted the SiRMS approach to describe materials with their fragmental compositions.

Every material is represented according to the structure of its crystal unit cell (Figure 2). Computing SiRMS descriptors for materials is equivalent to the computation of SiRMS fragments for nonbonded molecular mixtures. Bounded simplexes describe only a single component of the mixture. Unbounded simplexes could either belong to a single component, or could span up to four components of the unit cell. A special label is used during descriptor generation to distinguish "mixture" (belonging to different molecular moieties) simplexes from those incorporating elements from a single compound.[40]

Thus, the structure of every material is characterized by both bounded and unbounded SiRMS descriptors as illustrated in Figure 2. The descriptor value of a given simplex fragment is equal to the number of its occurrences in the system. In the case of materials, this value has been summed throughout all the constituents of a system; taking into account their stoichiometric ratios and crystal lattice (see Figure 2). "Mixture" descriptors are weighted according to the smallest stoichiometric ratio of constituents within this mixture, and added throughout all the mixtures in a system. Atoms in simplexes are differentiated according to their type (element) and partial charge. For the latter, atoms are divided into six groups corresponding to their

partial charge: $A \leq -2 < B \leq -1 < C \leq 0 < D \leq 1 < E \leq 2 < F$. In addition, we have developed a special differentiation of atoms in simplexes to account for their groups on the periodic table. That is, all elements belonging to the same group are encoded by the same symbol.

Network Representation (Material Cartograms). To represent the library of materials as a network, we considered each material, encoded by its fingerprint, as a node. Edges exist between nodes with similarities above certain thresholds. In this study, we use fingerprint-based Tanimoto similarity and a threshold $S = 0.7$. This network representation of materials is defined as the graph $G(V,E)$, where $V = \{\nu_1 | \nu_2 \in L\}$ and $E = \{(\nu_1,\nu_2) | \text{sim}(\nu_1,\nu_2) \geq T\}$. Here, L denotes a materials library, $\text{sim}(\nu_1,\nu_2)$ denotes a similarity between materials ν_1 and ν_2, and T denotes a similarity threshold.

To examine if the materials networks are scale-free, we analyzed the degree distributions of the networks. Networks are considered scale-free if the distribution of vertex degrees of the nodes follows the power law: $p(x) = kx^{-\alpha}$ where k is the normalization constant, and α is the exponent. The materials networks have been visualized using the Gephi package.[41] The ForceAtlas 2 algorithm,[42] a type of force-directed layout algorithm, has been used for the graph layout. A force-directed layout algorithm considers a force between any two nodes, and minimizes the "energy" of the system by moving the nodes and changing the forces between them. The algorithm guarantees that the topological similarity among nodes determines their vicinity, leading to accurate and visually informative representations of the materials space.

■ RESULTS AND DISCUSSION

Similarity Search in the Materials Space. In the first phase of this study, the optimized geometries, symmetries, band structures, and densities of states available in the AFLOWLIB consortium databases were converted into fingerprints, or arrays of numbers.

We encoded the electronic structure diagram for each material as two distinct types of fingerprints (Figure 1): band structure *symmetry-dependent* fingerprints (B-fingerprints), and density of states *symmetry-independent* fingerprints (D-fingerprints). The B-fingerprint is defined as a collated digitalized histogram of energy eigenvalues sampled at the high-symmetry reciprocal points with 32 bins. The D-fingerprint is a string containing 256 4-byte real numbers, each characterizing the strength of the DOS in one of the 256 bins dividing the [−10, 10] eV interval. More details are in the Methods section.

This unique, condensed representation of materials enabled the use of cheminformatics methods, such as similarity searches, to retrieve materials with similar properties but
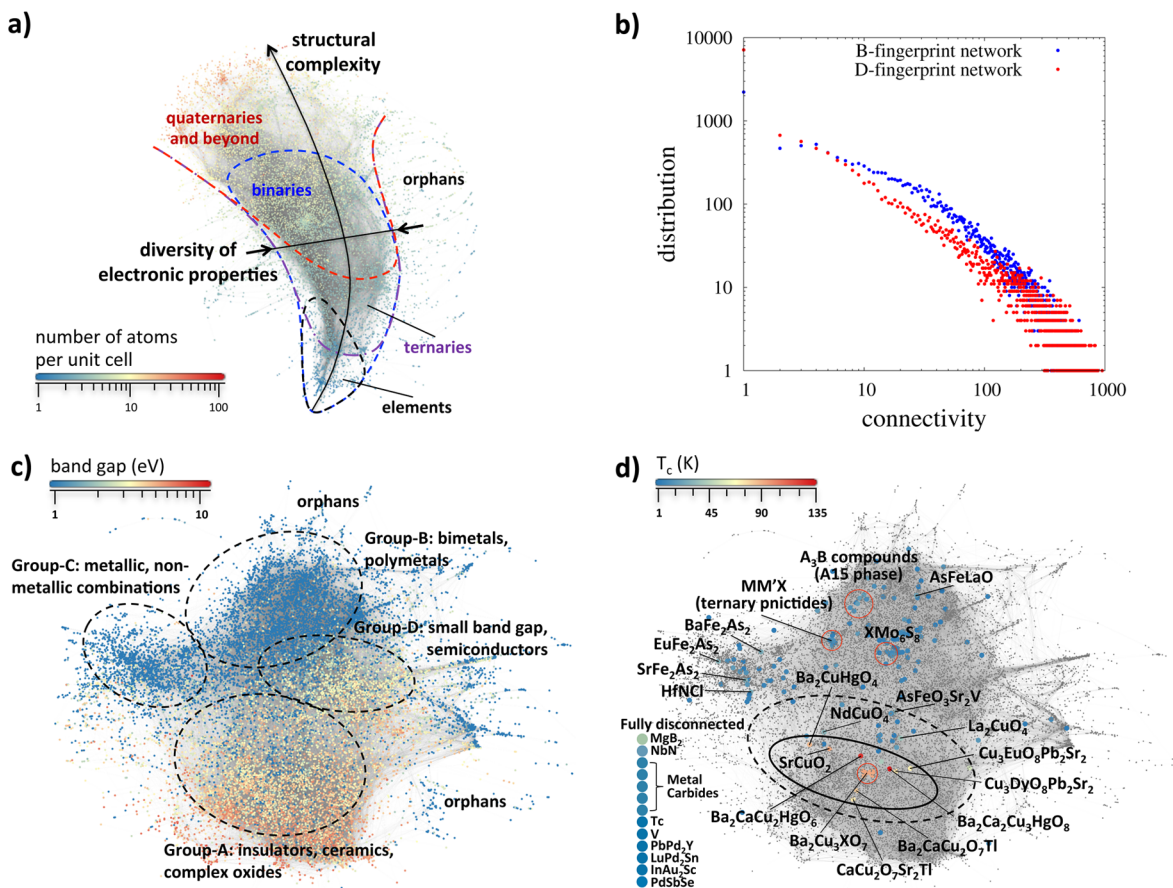
Figure 3. Materials cartograms with D- (top) and B-fingerprint network representations (bottom). (a) D-fingerprint network representation of materials. Materials are colored according to the number of atoms per unit cell. Regions corresponding to pure elements, binary, ternary, and quaternary compounds are outlined. (b) Distribution of connectivity within the network. (c) Mapping band gaps of materials. Points colored in deep blue are metals; insulators are colored according to the band gap value. Four large communities are outlined. (d) Mapping the superconductivity critical temperature, T_c, with relevant regions outlined.

different compositions from the AFLOWLIB database. As an added benefit, our similarity search can also quickly find duplicate records. For example, we have identified several $BaTiO_3$ records with identical fingerprints (ICSD Nos. 15453, 27970, 6102, and 27965 in the AFLOWLIB database). Thus, fingerprint representation afforded rapid identification of duplicates, which is the standard first step in our chem-informatics data curation workflow.[36] It is well-known that standard DFT has severe limitations in the description of excited states and needs to be substituted with more advanced approaches to characterize semiconductors and insulators.[43−47] However, there is a general trend of DFT errors being comparable in similar classes of systems. These errors may thus be considered "systematic", and are irrelevant when one seeks only similarities between materials.

The first test case is gallium arsenide, GaAs (ICSD No. 41674), a very important material for electronics[48] in the AFLOWLIB database. GaAs is taken as the reference material, and the remaining 20000+ materials from the AFLOWLIB database are taken as the virtual screening library. The pairwise similarity between GaAs and any of the materials represented by our D-fingerprints is computed using the Tanimoto similarity coefficient (S).[49] The top five materials (GaP, Si, SnP, GeAs, InTe) retrieved show very high similarity $(S > 0.8)$ to GaAs, and all five are known to be semiconductor materials.[34,50,51]

In addition, we have searched the AFLOWLIB database for materials similar to barium titanate ($BaTiO_3$) with the perovskite structure (ICSD No. 15453) using B-fingerprints. $BaTiO_3$ is widely used as a ferroelectric ceramic or piezo-electric.[52] Out of the six most similar materials with $S > 0.8$, five (BiOBr, $SrZrO_3$, $BaZrO_3$, $KTaO_3$ and $KNbO_3$) are well-known for their optical properties.[53] The remaining material, cubic YbSe (ICSD No. 33675), is largely unexplored. One can therefore formulate a testable hypothesis suggesting that this material may be ferroelectric or piezoelectric.

We also investigated the challenging case of topological insulators. They form a rare group of insulating materials with conducting surface-segregated states (or interfaces)[54] arising from a combination of spin−orbit coupling and time-reversal symmetry.[55] Although DFT calculations conducted for materials in AFLOWLIB do not incorporate spin−orbit coupling for the most part,[54] various topological insulators showed exceptionally high band-structure similarities, a manifesto for B-fingerprints. The two materials most similar to Sb_2Te_3[55] (based on B-fingerprints) with $S > 0.9$ were Bi_2Te_3[56,57] and Sb_2Te_2Se.[58] Five out of six materials most similar to Bi_2Te_2Se[58,59] are also known topological insulators: Bi_2Te_2S, Bi_2Te_3, Sb_2Te_2Se, $GeBi_2Te_4$,[58] and Sb_2Se_2Te.[54,60]

These examples demonstrate proof of concept and illustrate the power of simple yet uncommon fingerprint-based similarity searches for rapid and effective identification of materials with
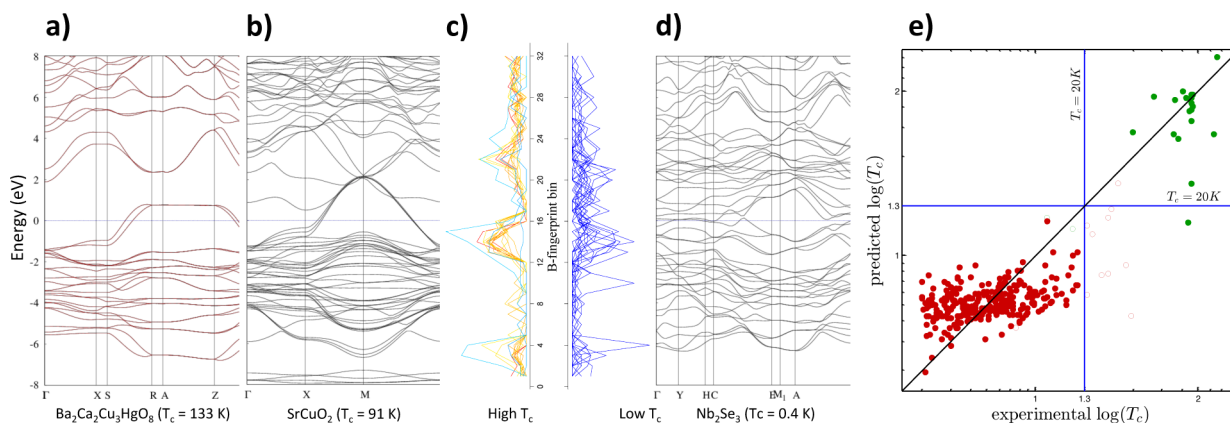
Figure 4. Comparison high-low T_c aligned band structures and T_c predictions. (a) Band structure for $Ba_2Ca_2Cu_3HgO_8$, T_c = 133 K. (b) Band structure of $SrCuO_2$ (ICSD No. 16217, T_c = 91 K).[74] (c) Aligned B-fingerprints for the 15 materials with the highest and lowest T_c. (d) Band structure of Nb_2Se_3 (ICSD No. 42981, T_c = 0.4 K). (e) Plot of the predicted vs experimental critical temperatures for the *continuous model*. Materials are color-coded according to the *classification model*: solid/open green (red) circles indicate correct/incorrect predictions in $T_c > T_{thr}$ ($T_c \leq T_{thr}$), respectively.

similar properties in large databases. They also illuminate the intricate link between the structure and properties of materials by demonstrating that similar materials (as defined by their fingerprint similarity) have similar properties (such as being ferroelectric or insulators). This observation sets the stage for building and exploring QMSPR models, as discussed in the following sections.

Visualizing and Exploring the Materials Space. The use of fingerprint representation and similarity concepts led us to develop the materials network. Compounds are mapped as nodes. We use the *force directed graph drawing* algorithm[61] in which positions of the compounds are initially taken randomly. There is a force between the nodes: a repulsive Coulomb component and an optional attractive contribution with a spring constant equal to the Tanimoto coefficient between D-fingerprints (effective when $S \geq 0.7$). Two nodes are connected only when the coefficient is at or above the threshold. The model is equilibrated through a series of heating and quenching steps. Figure 3a shows the result in which we add Bezier-curved lines depicting regions of accumulation. We shall refer to this approach to visualizing and analyzing materials and their properties as *materials cartography*.

The network shown in Figure 3a is colored according to overall complexity. Pure systems, 79% of the total 246 unary nodes, are confined in a small, enclosed region. Binary nodes cover more configurational space, and 82% of the 3700+ binaries lie in a compact region. Ternaries are scattered. They mostly populate the center of the space (91% of the 5300+ ternaries). Quaternaries and beyond are located at the top part of the network (92% of the 1080 nodes). This region is the most distant from that of unary nodes, which tends to be disconnected from the others. Indeed, overlap between binaries and ternaries is substantial. The diversification of electronic properties and thickness of the compact envelope grows with structural complexity. Orphans are defined as nodes with a very low degree of connectivity [only the vertices (materials) connected by edges are shown (∼39% of the database)]. Interestingly, of the 200 materials with connectivity smaller than 12, most are La-based (36 bimetallic and 126 polimetallic) or Ce-based (10 nodes).

The degree of connectivity is illustrated in Figure 3b. The panel indicates the log−log distribution of connectivity across

the sample set. The red and blue points measure the D-fingerprints (Figure 3a) and B-fingerprints connectivity (Figure 3c), respectively. Table 1 in the Supporting Information contains relevant statistical information about the cartograms. Although the power law distribution of Figure 3b is typical of scale-free networks and similar to many networks examined in cheminformatics and bioinformatics,[62−64] in our case, connectivity differs. In previous examples,[62−64] most of the nodes have only a few connections; with a small minority being highly connected to a small set of "hubs".[65,66] In contrast, the AFLOWLIB network is highly heterogeneous: most of the hubs' materials are concentrated along the long, narrow belt along the middle of the network. The top 200 nodes (ranked by connectivity) are represented by 83 polymetallics (CoCrSi, $Al_2Fe_3Si_3$, Al_8Cr_4Y, etc.), 102 bimetallics (Al_3Mo, As_3W_2, $FeZn_{13}$, etc.), 14 common binary compounds (GeS, AsIn, etc.), and boron (ICSD No. 165132). This is not entirely surprising, since these materials are well studied and represent the lion's share of the ICSD database. Al_3FeSi_2 (ICSD No. 79710), an uncommonly used material, has the highest connectivity of 946. Meanwhile, complex ceramics and exotic materials are relatively disconnected.

A second network, built with B-fingerprints, is illustrated in Figure 3c. While this network preserves most of the topological features described in the D-fingerprint case (Figure 3a), critical distinctions appear. The B-fingerprint network separates metals from insulators. Clustering and subsequent community analyses show four large groups of materials. Group-A (∼3000 materials) consists predominantly of insulating compounds (63%) and semiconductors (10%). Group-B distinctly consists of compounds with polymetallic character (70% of ∼2500 materials). In contrast, Group-C includes ∼500 zero band gap materials with nonmetal atoms, including halogenides, carbides, silicides, etc. Lastly, Group-D has a mixed character with ∼300 small band gap (below 1.5 eV) materials; and ∼500 semimetals and semiconductors.

Lithium scandium diphosphate, $LiScP_2O_7$ (ICSD No. 91496), has the highest connectivity of 746 for the B-fingerprint network. Very highly connected materials are nearly evenly distributed between Groups-A and -B, forming dense clusters within their centers. As in the case of the D-fingerprint network, the connectivity distribution follows a power law

(Figure 3b, see Table 1 in the Supporting Information for additional statistics); indicating that this is a scale-free network.

To illustrate one possible application of the materials networks, we chose superconductivity—one of the most elusive challenges in solid-state physics. We have compiled experimental data for 295 stoichiometric superconductors that were also available in AFLOWLIB. All materials in the data set are characterized with the fingerprints specified in the Methods section. The data set includes both prominently high temperature superconducting materials such as layered cuprates, ferropnictides, iron arsenides-122, MgB_2; as well as more conventional compounds such as A15, ternary pnictides, etc. Our model does not consider the effect of phonons, which play a dominant role in many superconductors.[67] High-throughput parametrization of phonon spectra is still in its infancy,[68] and only recently have vibrational descriptors been adapted to large databases.[69] We envision that future development of vibrational fingerprints following these guidelines will capture similarities between known, predicted, and verified superconductors (i.e., MgB_2 vs LiB_2[70,71] and MgB_2 vs Fe−B compounds[72,73]).

All materials are identified and marked on the B-fingerprint network, and are color-coded according to their critical temperature, T_c (Figure 3d). All high-T_c superconductors are localized in a relatively compact region. The distribution is centered on a tight group of $Ba_2Cu_3X\,O_7$ compounds (the so-called Y123, where X = lanthanides). The materials with the two highest T_c values in our set are $Ba_2Ca_2Cu_3HgO_8$ (ICSD No. 75730, T_c = 133 K) and $Ba_2CaCu_2HgO_6$ (ICSD No. 75725, T_c = 125 K). Their close grouping manifested a significant superconductivity hot-spot of materials with similar fingerprints. We aligned the B-fingerprints for 15 superconductors with the highest T_c values in Figure 4c.

All of the top 15 high T_c superconductors are layered cuprates, which have dominated high T_c superconductor research since 1986.[75] These compounds are categorized as Charge-Transfer Mott Insulators (CTMI).[76] There are three distinct bands that are conserved for these structures around −6, −1, and 4 eV relative to the Fermi energy at Γ (within the simple DFT+U description available in AFLOWLIB, Figure 4c). These features are consistent with the three-band Hubbard-like picture characteristic of CTMIs.[77,78]

Meanwhile, the fingerprint distribution for 15 materials with the lowest T_c was random (Figure 4c). The importance of band structure features in superconductivity has long been recognized.[79−81] Thus, materials cartography based on the B-fingerprint network allows us to visualize this phenomenon concisely.

Predictive QMSPR Modeling. We developed QMSPR models (continuous,[82] classification, and structural) to compute superconducting properties of materials from their structural characteristics. To achieve this objective, we compiled two superconductivity data sets consisting of (i) 295 materials with continuous T_c values ranging from 2 to 133 K and (ii) 464 materials with binary T_c values. The models were generated with Random Forest (RF)[83] and Partial Least Squares (PLS)[84] techniques. These used both B- and D-fingerprints, as well as the Simplex (SiRMS)[38] descriptors. These fingerprints were adapted for materials modeling for the first time in this study (see Methods section). Additionally, we incorporated atomic descriptors that differentiate by element, charge, and group within the periodic table. Statistical characteristics for all 464 materials used for the QMSPR analysis are reported in the Supporting Information (Tables 2−4).

Attempts to develop QMSPR models using B- and D-fingerprints for both data sets were not satisfactory, indicating that our fingerprints, while effective in qualitative clustering, do not contain enough information for quantitatively predicting target properties (QMSPR model acceptance criteria has been discussed previously[85]). Thus, we employed more sophisticated chemical fragment descriptors, such as SiRMS,[38] and adapted them for materials modeling (see Methods section).

Continuous Model. We constructed a *continuous model* which serves to predict the value of T_c with a consensus RF- and PLS-SiRMS approach. It has a cross-validation determination coefficient of $Q^2 = 0.66$ (5-fold external CV; see Table 2 of the Supporting Information). Figure 4e shows predicted versus experimental T_c values for the continuous model: all materials having $\log(T_c) \leq 1.3$ were scattered, but within the correct range. Interestingly, we notice that systems with $\log(T_c) \geq 1.3$ received higher accuracy, with the exceptions of MgB_2 (ICSD No. 26675), Nb_3Ge (ICSD No. 26573), $Cu_1Nd_2O_4$ (ICSD No. 4203), As_2Fe_2Sr (ICSD No. 163208), Ba_2CuHgO_4 (ICSD No. 75720), and ClHfN (ICSD No. 87795) (all highly underestimated). Not surprisingly MgB_2[86] is an outlier in our statistics. This is in agreement with the fact that to date no superconductor with an electronic structure similar to MgB_2 has been found.

Classification Model. By observing the existence of the threshold T_{thr} = 20 K ($\log(T_{thr})$ = 1.3), we developed a *classification model.* It is based on the same RF-SiRMS technique, but it is strictly used to predict the position of T_c with respect to the threshold, above or below. The classification model has a balanced accuracy (BA) of 0.97 with 5-fold external CV analysis. The type of points in Figure 4e illustrates the classification model outcome: solid/open green (red) circles for correct/incorrect predictions in $T_c > T_{thr}$ ($T_c \leq T_{thr}$), respectively.

For $T_c \leq T_{thr}$ and $T_c > T_{thr}$, accuracies of prediction were 98% and 90% (cumulative 94%). (Figure 4e, see Table 3 in the Supporting Information for additional statistics). Among the 464 materials, 10 systems with experimental $T_c > T_{thr}$ were predicted to have $T_c \leq T_{thr}$) [FeLaAsO (ICSD No. 163496), $AsFeO_3Sr_2V$ (ICSD No. 165984), As_2EuFe_2 (ICSD No. 163210), As_2Fe_2Sr, $CuNd_2O_4$ (ICSD No. 86754), As_2BaFe_2 (ICSD No. 166018), MgB_2, ClHfN, La_2CuO_4, and Nb_3Ge]. Only two with experimental $T_c \leq T_{thr}$ were predicted with $T_c > T_{thr}$ (AsFeLi (ICSD No. 168206), As_2CaFe_2 (ICSD No. 166016)). Owing to the spread around the threshold, additional information about borates and Fe−As compounds is required for proper training of the learning algorithm.

In the past, it has been shown that QSAR approaches can be used for the detection of mis-annotated chemical compounds, a critical step in data curation.[36] We have employed a similar approach here. In our models, three materials, ReB_2 (ICSD No. 23871), Li_2Pd_3B (ICSD No. 84931), and La_2CuO_4, were significantly mispredicted. More careful examination of the data revealed that T_c's of ReB_2 and Li_2Pd_3B were incorrectly extracted from literature. We also found that La_2CuO_4 had the largest variation of reported values within the data set. Therefore, it was excluded from the regression. This approach illustrates that QMSPR modeling should be automatically implemented to reduce and correct erroneous entries.

Structural Model. We also developed a *structural model* meant to capture the geometrical features that most influence
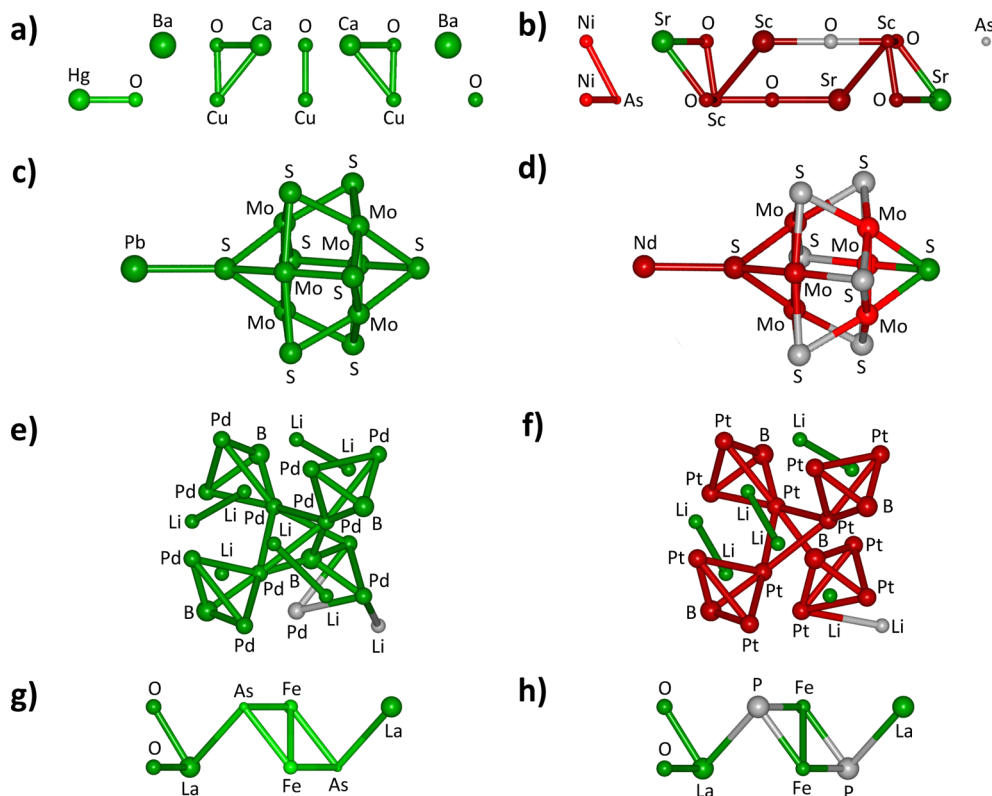
Figure 5. Materials color-coded according to atom contributions to $\log(T_c)$. Atoms and structural fragments that decrease superconductivity critical temperatures are colored in red and those enhancing T_c are shown in green. Uninfluential fragments are in gray. (a) $Ba_2Ca_2Cu_3HgO_8$; (b) $As_2Ni_2O_6Sc_2Sr_4$; (c) Mo_6PbS_8; (d) Mo_6NdS_8; (e) Li_2Pd_3B; (f) Li_2Pt_3B; (g) FeLaAsO; (h) FeLaPO.

T_c. It employs SiRMS descriptors, PLS approaches, and 5-fold external cross-validation. The predictive performance of this model ($Q^2 = 0.61$) is comparable to that of the SiRMS-based RF model (see Table 2 in Supporting Information for additional statistics). The top 10 statistically significant geometrical fragments and their contributions to T_c variations are shown in Table 4 of the Supporting Information. All descriptor contributions were converted to atomic contributions (details discussed previously[87]) and related to material structures. Examples of unit cell structures for pairs of similar materials with different T_c values were color-coded according to atomic contributions to T_c and are shown in Figure 5 (green for $T_c\uparrow$, red for $T_c\downarrow$, and gray for neutral).

Examples of fragments for materials having $T_c > T_{thr}$ [$Ba_2Ca_2Cu_3HgO_8$, ICSD No. 75730, $\log(T_c) = 2.12$] and $T_c \leq T_{thr}$ [$As_2Ni_2O_6Sc_2Sr_4$, ICSD No. 180270, $\log(T_c) = 0.44$] are shown in Figure 5 parts a and b, respectively. They indicate that individual atom contributions are nonlocal as they strongly depend upon the atomic environment (Figure 5c–h), for example, Mo_6PbS_8 [ICSD No. 644102, $\log(T_c) = 1.13$] and Mo_6NdS_8 [ICSD No. 603458, $\log(T_c) = 0.54$] differ by a substitution, yet the difference in T_c is substantial. Furthermore, substitution of Nd for Pb affects contributions to the target property from all the remaining atoms in the unit cell (Figure 5c,d). The same observation holds for Li_2Pd_3B [ICSD No. 84931, $\log(T_c) = 0.89$] and Li_2Pt_3B [ICSD No. 84932, $\log(T_c) = 0.49$] Figure 5e,f; as well as FeLaAsO [ICSD No. 163496, $\log(T_c) = 1.32$] and FeLaPO [ICSD No. 162724, $\log(T_c) = 0.82$] Figure 5g,h.

CONCLUSION

With high-throughput approaches in materials science increasing the data-driven content of the field, the gap between accumulated-information and derived knowledge widens. The issue can be overcome by adapting the data-analysis approaches developed during the past decade for chem- and bioinformatics.

Our study gives an example of this. We introduce novel materials fingerprint descriptors that lead to the generation of networks called *materials cartograms*: nodes represent compounds; connections represent similarities. The representation can identify regions with distinct physical and chemical properties, the key step in searching for interesting, yet unknown compounds.

Starting from atomic-compositions, bond-topologies, structure-geometries, and electronic properties of materials publicly available in the AFLOWLIB repository, we have introduced cheminformatics models leveraging novel materials fingerprints. Within our formalism, simple band-structure and DOS fingerprints are adequate to locate metals, semiconductors, topological insulators, piezoelectrics, and superconductors. More complex QMSPR modeling[38] are used to tackle qualitative and quantitative values of superconducting critical temperature and geometrical features helping/hindering criticality.[38]

In summary, the fingerprinting cartography introduced in this work has demonstrated its utility in an initial set of problems. This shows the possibility of designing new materials and gaining insight into the relationship between the structure and physical properties of materials. Further advances in the analysis and exploration of databases may become the

foundation for rationally designing novel compounds with desired properties.

◼ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional statistical information relevant to the materials cartograms and all three models (continuous, classification, and structural). This material is available free of charge via the Internet at http://pubs.acs.org.

◼ AUTHOR INFORMATION

Corresponding Authors

*E-mail: alex_tropsha@unc.edu.
*E-mail: stefano@duke.edu.

Notes

The authors declare no competing financial interest.

◼ ACKNOWLEDGMENTS

◼ REFERENCES

(1) Rajan, K. *Mater. Today* **2005**, *8*, 38−45.

(2) Curtarolo, S.; Hart, G. L. W.; Buongiorno Nardelli, M.; Mingo, N.; Sanvito, S.; Levy, O. *Nat. Mater.* **2013**, *12*, 191−201.

(3) Potyrailo, R.; Rajan, K.; Stoewe, K.; Takeuchi, I.; Chisholm, B.; Lam, H. *ACS Comb. Sci.* **2011**, *13*, 579−633.

(4) Bergerhoff, G.; Hundt, R.; Sievers, R.; Brown, I. D. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 66−69.

(5) MatWeb, LLC. *MatWeb Material Property Data*. http://www.matweb.com. 2011.

(6) *Idemat Database*. Delft University of Technology, Matbase. http://www.matbase.com. 2003.

(7) Setyawan, W.; Curtarolo, S. *Comput. Mater. Sci.* **2010**, *49*, 299−312.

(8) Setyawan, W.; Gaume, R. M.; Lam, S.; Feigelson, R. S.; Curtarolo, S. *ACS Comb. Sci.* **2011**, *13*, 382−390.

(9) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. *Appl. Phys. Lett. Mater.* **2013**, *1*, 011002.

(10) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.

(11) Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Román-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; Shrestha, S.; Mondal, R.; Sokolov, A.; Bao, Z.; Aspuru-Guzik, A. *Energy Environ. Sci.* **2014**, *7*, 698−704.

(12) Materials Genome Initiative. Office of Science and Technology Policy, White House, http://www.whitehouse.gov/mgi 2011.

(13) Suh, C.; Rajan, K. *Mater. Sci. Technol.* **2009**, *25*, 466−471.

(14) Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sanchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. *Energy Environ. Sci.* **2011**, *4*, 4849−4861.

(15) Schuett, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Mueller, K. R.; Gross, E. K. U. *Phys. Rev. B* **2014**, *89*, 205118.

(16) Seko, A.; Maekawa, T.; Tsuda, K.; Tanaka, I. *Phys. Rev. B* **2014**, *89*, 054303.

(17) Laggner, C.; Kokel, D.; Setola, V.; Tolia, A.; Lin, H.; Irwin, J. J.; Keiser, M. J.; Cheung, C. Y. J.; M, D. L., Jr.; Roth, B. L.; Peterson, R. T.; Shoichet, B. K. *Nat. Chem. Biol.* **2012**, *8*, 144−146.

(18) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriquiz, R. M.; Huang, X. P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; et al. *Nature* **2012**, *492*, 215−220.

(19) Cherkasov, A.; Muratove, E. N.; Fourches, D.; Varnexk, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. *J. Med. Chem.* **2013**, *57*, 4977.

(20) Lusci, A.; Pollastri, G.; Baldi, P. *J. Chem. Inf. Model.* **2013**, *53*, 1563−1575.

(21) Balachandran, P. V.; Broderick, S. R.; Rajan, K. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* **2011**, *467*, 2271−2290.

(22) Kong, C. S.; Luo, W.; Arapan, S.; Villars, P.; Iwata, S.; Ahuja, R.; Rajan, K. *J. Chem. Inf. Model.* **2012**, *52*, 1812−1820.

(23) Balachandran, P. V.; Rajan, K. *Acta Crystallogr. Sect. B, Struct. Sci.* **2012**, *68*, 24−33.

(24) Srinivasan, S.; Rajan, K. *Materials* **2013**, *6*, 279−290.

(25) Broderick, S.; Ray, U.; Srinivasan, S.; Rajan, K.; Balasubramanian, G. *Appl. Phys. Lett.* **2014**, *104*, 243110.

(26) Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. *Comput. Mater. Sci.* **2014**, *83*, 185−195.

(27) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. *Comput. Mater. Sci.* **2012**, *58*, 218−226.

(28) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno Nardelli, M.; Mingo, N.; Levy, O. *Comput. Mater. Sci.* **2012**, *58*, 227−235.

(29) Kresse, G.; Furthmüller, J. *Comput. Mater. Sci.* **1996**, *6*, 15.

(30) Blöchl, P. E. *Phys. Rev. B* **1994**, *50*, 17953−17979.

(31) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(32) Taylor, R. H.; Rose, F.; Toher, C.; Levy, O.; Yang, K.; Buongiorno Nardelli, M.; Curtarolo, S. *Comput. Mater. Sci.* **2014**, *93*, 178−192.

(33) Poole, C. P. *Handbook of Superconductivity*; Academic Press: San Diego, CA, 2000.

(34) Lide, D. R. *CRC Handbook of Chemistry and Physics*; Taylor & Francis: Oxford, U.K., 2004.

(35) National Institute of Materials Science, Materials Information Station. *SuperCon.* http://supercon.nims.go.jp/index_en.html. 2011.

(36) Fourches, D.; Muratov, E.; Tropsha, A. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.

(37) Valle, M.; Oganov, A. R. *Acta Crystallogr., Sect. A* **2010**, *66*, 507−517.

(38) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. *J. Comp.-Aided Mol. Des.* **2008**, *22*, 403−421.

(39) Muratov, E. N.; Varlamova, E. V.; Artemenko, A. G.; Polishchuk, P. G.; Nikolaeva-Glomb, L.; Galabov, A. S.; Kuz'min, V. E. *Struct Chem.* **2013**, *24*, 1665−1679.

(40) Muratov, E. N.; Varlamova, E. V.; Artemenko, A. G.; Polishchuk, P. G.; Kuz'min, V. E. *Mol. Inf.* **2012**, *31*, 202−221.

(41) Bastian, M.; Heymann, S.; Jacomy, M. *Int. Conf. Weblogs Social Media* **2009**, *8*, 361−362.

(42) Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. *PLoS One* **2014**, *9*, e98679.

(43) Hedin, L. *Phys. Rev.* **1965**, *139*, A796−A823.

(44) Aryasetiawan, F.; Gunnarsson, O. *Rep. Prog. Phys.* **1998**, *61*, 237.

(45) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. *J. Chem. Phys.* **2003**, *118*, 8207−8215.

(46) Liechtenstein, A. I.; Anisimov, V. I.; Zaanen, J. *Phys. Rev. B* **1995**, *52*, R5467−R5470.

(47) Himmetoglu, B.; Floris, A.; de Gironcoli, S.; Cococcioni, M. *Int. J. Quantum Chem.* **2014**, *114*, 14−49.

(48) INSPEC. *Properties of Gallium Arsenide*; INSPEC, Institution of Electrical Engineers: MI, 2007.

(49) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. *J. Med. Chem.* **2014**, *57*, 3186−3204.

(50) Littlewood, P. B. *Crit. Rev. Solid. State. Mater. Sci.* **1983**, *11*, 229−285.

(51) Madelung, O. *Semiconductors: Data Handbook*; Springer: Berlin, 2004.

(52) Bhalla, A. S.; Guo, R.; Roy, R. *Mater. Res. Innovat.* **2000**, *4*, 3−26.

(53) Rabe, K. M.; Ahn, C. H.; Triscone, J. M. *Physics of Ferroelectrics: A Modern Perspective*; Springer: New York, 2010.

(54) Yang, K.; Setyawan, W.; Wang, S.; Buongiorno Nardelli, M.; Curtarolo, S. *Nat. Mater.* **2012**, *11*, 614−619.

(55) Hasan, M. Z.; Kane, C. L. *Rev. Mod. Phys.* **2010**, *82*, 3045−3067.

(56) Chen, Y. L.; Analytis, J. G.; Chu, J.-H.; Liu, Z. K.; Mo, S.-K.; Qi, X.-L.; Zhang, H.-J.; Lu, D. H.; Dai, X.; Fang, Z.; Zhang, S.-C.; Fisher, I. R.; Hussain, Z.; Shen, Z.-X. *Science* **2009**, *325*, 178−181.

(57) Zhang, T.; Cheng, P.; Chen, X.; Jia, J.-F.; Ma, X.; He, K.; Wang, L.; Zhang, H.; Dai, X.; Fang, Z.; Xie, X.; Xue, Q.-K. *Phys. Rev. Lett.* **2009**, *103*, 266803.

(58) Xu, S.-Y.; Wray, L. A.; Xia, Y.; Shankar, R., Petersen, A.; Fedorov, A.; Lin, H.; Bansil, A.; Hor, Y. S.; Grauer, D.; Cava, R. J.; Hasan, M. Z.*Condens. Matters*, **2010**; arXiv:1007.5111v1.

(59) Arakane, T.; Sato, T.; Souma, S.; Kosaka, K.; Nakayama, K.; Komatsu, M.; Takahashi, T.; Ren, Z.; Segawa, K.; Ando, Y. *Nat. Commun.* **2012**, *3*, 636.

(60) Zhang, H.-J.; Liu, C.-X.; Qi, X.-L.; Dai, X.; Fang, Z.; Zhang, S.-C. *Nat. Phys.* **2009**, *5*, 438−442.

(61) Herman, I.; Melançon, G.; Marshall, M. S. *IEEE Trans Visual Comput. Graphics* **2000**, *6*, 24−43.

(62) Girvan, M.; Newman, M. E. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7821−7826.

(63) Newman, M. E. J. *SIAM Rev.* **2003**, *45*, 167−256.

(64) Yildirim, M. A.; Goh, K.-I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M. *Nat. Biotechnol.* **2007**, *25*, 1119−1126.

(65) Jeong, H.; Tombor, B.; Albert, R.; Oltavai, Z. N.; Barabási, A.-L.; Dandekar, T.; Schuster, S. *Nature* **2000**, *407*, 651−654.

(66) Barabási, A.-L.; Albert, R. *Science* **1999**, *286*, 509−512.

(67) Tinkham, M. *Introduction to Superconductivity*; Dover Publications: Mineaola, NY, 1996.

(68) Jahnatek, M.; Levy, O.; Hart, G. L. W.; Nelson, L. J.; Chepulskii, R. V.; Xue, J.; Curtarolo, S. *Phys. Rev. B* **2011**, *84*, 214110.

(69) Toher, C.; Plata, J. J.; Levy, O.; de Jong, M.; Asta, M. D.; Buongiorno Nardelli, M.; Curtarolo, S. *Phys. Rev. B* **2014**, *90*, 174107.

(70) Kolmogorov, A. N.; Curtarolo, S. *Phys. Rev. B* **2006**, *73*, 180501(R).

(71) Kolmogorov, A. N.; Curtarolo, S. *Phys. Rev. B* **2006**, *74*, 224507.

(72) Kolmogorov, A. N.; Shah, S.; Margine, E. R.; Bialon, A. F.; Hammerschmidt, T.; Drautz, R. *Phys. Rev. Lett.* **2010**, *105*, 217003.

(73) Gou, H.; Dubrovinskaia, N.; Bykova, E.; Tsirlin, A.; Kasinathan, D.; Schnelle, W.; Richter, A.; Merlini, M.; Hanfland, M.; Abakumov, A.; Batuk, D.; van Tendeloo, G.; Nakajima, Y.; Kolmogorov, A.; Dubrovinsky, L. *Phys. Rev. Lett.* **2013**, *111*, 157002.

(74) Takahashi, H.; Môri, N.; Azuma, M.; Hiroi, Z.; Takano, M. *Physica C* **1994**, *227*, 395−398.

(75) Bednorz, J. G.; Müller, K. A. Z. *Physik B−Condens. Matter* **1986**, *64*, 189−193.

(76) Zaanen, J.; Sawatzky, G. A.; Allen, J. W. *Phys. Rev. Lett.* **1985**, *55*, 418−421.

(77) Manske, D.; Eremin, I.; Bennemann, K. H. In *The Physics of Superconductors*; Bennemann, K. H., Ketterson, J. B., Eds.; Springer: Berlin Heidelberg, 2004.

(78) Emery, V. J. *Phys. Rev. Lett.* **1987**, *58*, 2794−2797.

(79) Zaanen, J.; Chakravarty, S.; Senthil, T.; Anderson, P.; Lee, P.; Schmalian, J.; Imada, M.; Pines, D.; Randeria, M.; Varma, C.; Vojta, M.; Rice, M. *Nat. Phys.* **2006**, *2*, 138−143.

(80) Micnas, R.; Ranninger, J.; Robaszkiewicz, S. *Rev. Mod. Phys.* **1990**, *62*, 113−171.

(81) Orenstein, J.; Millis, A. J. *Science* **2000**, *288*, 468−474.

(82) Bramer, M. *Principles of Data Mining*; Springer-Verlag: London, 2007.

(83) Breiman, L. *Mach. Learn.* **2001**, *45*, 5−32.

(84) Wold, S.; Sjöström, M.; Eriksson, L. *Chemometr. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(85) Tropsha, A. *Mol. Inf.* **2010**, *29*, 476−488.

(86) Buzea, C.; Yamashita, T. *Supercond. Sci. Technol.* **2001**, *14*, R115.

(87) Muratov, E. N.; Artemenko, A. G.; Varlamova, E. V.; Polischuk, P. G.; Lozitsky, V. P.; Fedchuk, A. S.; Lozitska, R. L.; Gridina, T. L.; Koroleva, L. S.; Sil'nikov, V. N.; et al. *Future Med. Chem.* **2010**, *2*, 1205−1226.