This is the accepted manuscript made available via CHORUS. The article has been published as:

Accelerated materials property predictions and design using motif-based fingerprints

Tran Doan Huan, Arun Mannodi-Kanakkithodi, and Rampi Ramprasad

Phys. Rev. B **92**, 014106 — Published 8 July 2015

DOI: 10.1103/PhysRevB.92.014106

Accelerated materials property predictions and design using motif-based fingerprints

Tran Doan Huan,[1] Arun Mannodi-Kanakkithodi,[1] and Rampi Ramprasad[1,*]

[1] *Institute of Materials Science, University of Connecticut, 97 North Eagleville Rd., Unit 3136, Storrs, CT 06269-3136, USA*

(Dated: June 15, 2015)

Data-driven approaches are particularly useful for computational materials discovery and design as they can be used for rapidly screening over a very large number of materials, thus suggesting lead candidates for further in-depth investigations. A central challenge of such approaches is to develop a numerical representation, often referred to as a fingerprint, of the materials. Inspired by recent developments in chem-informatics, we propose a class of hierarchical motif-based topological fingerprints for materials composed of elements such as C, O, H, N, F, etc., whose coordination preferences are well understood. We show that these fingerprints, when representing either molecules or crystals, may be effectively mapped onto a variety of properties using a similarity-based learning model and hence can be used to predict relevant properties of a material, given that its fingerprint can be defined. Two simple machine-learning based procedures are introduced to demonstrate that the learning model can be inverted to identify the desired fingerprints and then, to reconstruct molecules which possess a set of targeted properties.

I. INTRODUCTION

Data-driven approaches towards materials design and discovery are rapidly increasing in popularity, demand and potency.[1–15] This emerging trend is fueled by the availability and emergence of large materials databases,[16–18] as well as our ability to progressively accumulate materials data via high-throughput computations[19,20] and experiments.[16–18] Data-driven strategies aimed at rapid property predictions, and ultimately at rational or informed materials design, rely on exploiting the information content of past data, and using such information within heuristic or statistical interpolative learning models to provide estimates of properties of a new material. This approach is entirely analogous to similar pursuits undertaken within chem- and bio-informatics wherein lead candidates worthy of further in-depth investigations are identified rapidly in a first-level of screening.[4,5,14]

Data-driven property prediction strategies have two steps. The first involves representing materials numerically via descriptors, attribute vectors, or fingerprints. In the second step, using available "training" data sets, a mapping is established between the numerical representation of materials and their properties, thus leading to a prediction model. Subsequently, the properties of a new material are estimated using this model after reducing the material to its numerical representation.

One of the central challenges in this whole process is deciding on an appropriate and acceptable numerical representation of materials. The specific choice of this representation is entirely application dependent, and can range from high level descriptors (e.g., *d*-band center, atomic electronegativities)[21,22] to topological features (e.g., substructural motifs)[20,23,24] to microscopic fingerprints that may capture chemical and configurational degrees of freedom (e.g., coulomb matrix, symmetry functions).[25–28] Regardless of the specific choice, the representations are expected to satisfy certain basic requirements. These include invariance of the representation with respect to transformations of the material such as translation, rotation, and permutation of like elements. Moreover, it is desired that the representation be intuitive, elegant and physically and chemically meaningful.

In this contribution, inspired by developments in chem-informatics,[14,15] we propose a class of hierarchical motif-based topological fingerprints. This choice, in which the motifs are molecular fragments of varying sizes, is particularly suited to representing molecules and solids composed of elements such as H, C, N, O, F, etc., whose coordination preferences are well understood. Large datasets of molecules and solids are considered, and it is shown that the fingerprints may be effectively mapped to a variety of properties using a similarity based learning algorithm. Moreover, it is demonstrated that the learning model may be inverted to identify fingerprints, and subsequently, to reconstruct actual molecules that possess a desired set of target properties.

II. DATASETS

In the present work, we restrict ourselves to systems composed of C, O and H. We used two datasets, one for molecules and one for crystals, to demonstrate the applicability of the proposed fingerprints. Of these two datasets, the former was taken from Ref. 19 while the latter was prepared by us.

A. Molecule dataset

A dataset of more than 134,000 small molecules made up of C, O, H, N, and F was reported in Ref. 19. This reliable dataset, which contains the optimized geometries, and energetic, electronic, and thermodynamic properties calculated using the B3LYP hybrid exchange-correlation

(XC) functional and the 6-31G(2df,p) basis set with the Gaussian 09 software, sets up the stage for many interesting data-mining works.[29,30] A subset of this dataset, containing 45,708 molecules composed of C, O, and H was used in this work. Five properties were considered, including the atomization energy $\mathcal{E}_{\rm at}$, the energy gap $E_{\rm HL}$ between highest occupied and lowest unoccupied molecular orbitals (HOMO-LUMO gap), the isotropic polarizability α, the heat capacity $C_{\rm v}$, and the zero-point vibration energy $\mathcal{E}_{\rm ZP}$.

B. Crystal dataset

In addition to the molecules dataset, we prepared another dataset of 215 organic crystals comprising of C, O, and H. This includes

1. 12 existing polymers composed of C, O, and H,
2. 16 new polymer structures predicted by the minima-hopping method[31–33] and USPEX[34] for 16 quasi-one-dimensional polymer chain models reported in Ref. 3,
3. 34 organic crystals composed of C and H and 153 organic crystals composed of C, O, and H obtained from *Crystallography Open Database*.[18]

The obtained structures were optimized by first-principles calculations within the DFT formalism as implemented in Vienna *Ab initio* Simulation Package (VASP),[35–38] utilizing the semi-local rPW86 XC functional[39] and a plane wave energy cutoff of 400 eV. A Monkhorst-Pack **k**-point mesh[40] with the spacing of no more than 0.15Å^{-1} in the reciprocal space were used for sampling the Brillouin zone, while the van der Waals interactions were estimated with the non-local density functional vdW-DF2.[41] Convergence was assumed when the atomic forces exerting on the atomic sites are smaller than 0.01 eV/Å. The entire crystals dataset, which includes the optimized structures, the atomization energies $\mathcal{E}_{\rm at}$, the band gaps $E_{\rm g}$, and the electronic and ionic parts of the dielectric constants, $\epsilon_{\rm elec}$ and $\epsilon_{\rm ion}$, can be found in the Supplemental Material.[42]

III. FINGERPRINTS

A hierarchy of equilibrium structure fingerprints of the same family with increasing levels of sophistication are proposed here. The construction of fingerprints was guided by two simple chemical concepts, i.e., chemical bonds and coordination number. The former intuitively characterizes the short-range interatomic interactions[43] while the latter is the number of bonds involving a given atom. In major classes of materials composed of light elements like C, H, O, N, and F, these concepts are well-defined. In particular, the length of a given bond involving these elements falls in a narrow range (see Refs. 44 and 45 for a comprehensive bond length statistics). For instance, the equilibrium length of a single bond between two C atoms is $\simeq 1.50$Å, the length of a double bond between two C atoms is $\simeq 1.45$Å, and the length of a double bond between a C atom and an O atom is $\simeq 1.20$Å.[44,45] The coordination number is also well-defined, i.e., for a C atom, it can only be 2, 3, or 4 while each O atom can generally bond with 1 or 2 other atoms. Therefore, atoms in a structure can be unambiguously classified (or labeled) by $\mathcal{A}i$ where $\mathcal{A}$ is the type of the element ($\mathcal{A} \in \{\mathrm{C}, \mathrm{O}, \mathrm{H}\}$) and i is its coordination number. Likewise, bonds can be specified by the types of its two ends, e.g., $\mathcal{A}i$-$\mathcal{B}j$. For the datasets of C, O, and H, the six possible atom types are C2, C3, C4, O1, O2, and H1 while there are sixteen chemically permissible types of bonds, namely C2-C2, C2-C3, C2-C4, C2-O1, C2-O2, C2-H1, C3-C3, C3-C4, C3-O1, C3-O2, C3-H1, C4-C4, C4-O2, C4-H1, O2-O2, and O2-H1. Except C2-O1, C2-O2, and O2-O2, thirteen of them are present in our molecules and crystals datasets. The atom and bond types belong to a family of related structural building units (subsequently described) that can be used to numerically represent the materials structures and hence, are used to define the fingerprints. In particular, the $i^{\rm th}-$order fingerprint $\mathbf{f}^{(i)}$ is defined in terms of its components as

$$f_\kappa^{(i)} = \frac{n_\kappa^{(i)}}{N_{\rm at}}. \qquad (1)$$

Here, $n_\kappa^{(i)}$ is the number of building units (or fragments or motifs) of type κ and $N_{\rm at}$ is the number of atoms either in the molecule or in the unit cell of a crystal. Four types of fingerprints, namely $\mathbf{f}^{(0)}$, $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, and $\mathbf{f}^{(3)}$, are discussed in the following subsections.

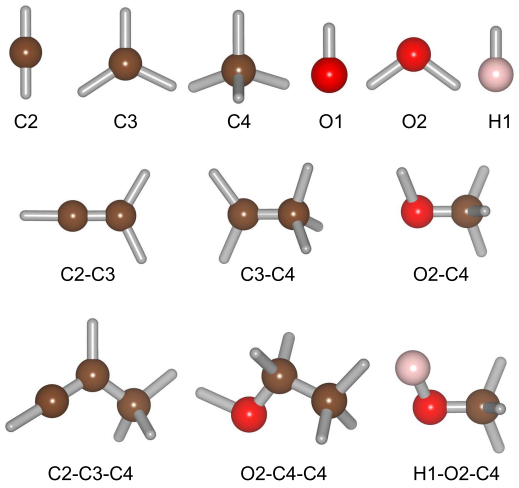

FIG. 1. (Color online) Illustration of motifs of several types, including the atom types ($\mathcal{A}i$, top row), some of the bond types ($\mathcal{A}i$-$\mathcal{B}j$, middle row) and two-bond catenations ($\mathcal{A}i$-$\mathcal{B}j$-$\mathcal{C}k$, bottom row) of materials composed by carbon, oxygen, and hydrogen.

A. 0^{th}-order fingerprint, $\mathbf{f}^{(0)}$

The simplest (0^{th}-order) fingerprint $\mathbf{f}^{(0)}$ represents the fractions of all the element types $\mathcal{A}$ existing in the structures, i.e., $\kappa \equiv \mathcal{A}$. Therefore, in the definition (1) of $\mathbf{f}^{(0)}$, $n^{(0)}_{\kappa\equiv\mathcal{A}}$ is the number of atoms of element $\mathcal{A}$. This fingerprint is a three-dimensional vector whose components satisfy a simple normalization condition $\sum_{\mathcal{A}\in\{\text{C,O,H}\}} f^{(i)}_{\mathcal{A}} = 1$.

B. 1^{st}-order fingerprint, $\mathbf{f}^{(1)}$

Next in the hierarchy is the case $\kappa \equiv \mathcal{A}i$ in which $n^{(1)}_{\kappa\equiv\mathcal{A}i}$ is the number of $\mathcal{A}$ atoms which are $i-$fold coordinated. $\mathbf{f}^{(1)}$ is a 6-dimensional vector, satisfying several constraints established from the definition or from the chemistry. The first one is the normalization condition, given as

$$\sum_{\mathcal{A}i} f^{(1)}_{\mathcal{A}i} = 1. \tag{2}$$

Within the two datasets, all the C2 atoms should be grouped by pairs, forming triple C≡C bonds. Therefore, the number of C2 atoms, which is $N_{\text{at}} \times f^{(1)}_{\text{C2}}$, must be an even integer. Moreover, since each C3 atom only make a double bond with either an O1 atom or another C3 atom, one must have $f^{(1)}_{\text{C3}} \geq f^{(1)}_{\text{O1}}$ while $N_{\text{at}} \times \left[f^{(1)}_{\text{C3}} - f^{(1)}_{\text{O1}}\right]$ is an even number. By examining the connectivity of a structure, another constraint reads

$$f^{(1)}_{\text{H1}} - 2f^{(1)}_{\text{C4}} - f^{(1)}_{\text{C3}} + f^{(1)}_{\text{O1}} = \frac{2}{N_{\text{at}}}\left(1 - N_{\bigcirc} - d\right) \tag{3}$$

where $N_{\bigcirc}$ is the number of closed loops of bonds and d is a structure-dependent parameter. For molecules and crystals composed of isolated substructures (or molecules), $d = 0$ while for crystals composed of connected substructures, $d > 0$. The derivation of this constraint is given in Appendix A. The last constraint of $\mathbf{f}^{(1)}$ is written in the form of a recursion relation, i.e.,

$$\sum_{i} f^{(1)}_{\mathcal{A}i} = f^{(0)}_{\mathcal{A}}. \tag{4}$$

C. 2^{nd}-order fingerprint, $\mathbf{f}^{(2)}$

Both $\mathbf{f}^{(0)}$ and $\mathbf{f}^{(1)}$ are local, representing the density of the atom types of a material. The equilibrium interatomic distance is somehow captured by the 2^{nd}-order fingerprint $\mathbf{f}^{(2)}$ where all the possible bonds are counted. $\mathbf{f}^{(2)}$ is a 13-dimensional vector whose components, $f^{(2)}_{\mathcal{A}i\text{-}\mathcal{B}j}$, represent the normalized number $n^{(2)}_{\mathcal{A}i\text{-}\mathcal{B}j}$ of the $\mathcal{A}i$-$\mathcal{B}j$ bonds in the structure. From $\mathbf{f}^{(2)}$, $\mathbf{f}^{(1)}$ can readily be determined by a recursion relation

$$f^{(1)}_{\mathcal{A}i} = \sum_{\mathcal{B}j} \frac{2^{\delta_{\mathcal{A}i,\mathcal{B}j}-1}}{i} f^{(2)}_{\mathcal{A}i\text{-}\mathcal{B}j} \tag{5}$$

where $\delta_{\mathcal{A}i,\mathcal{B}j}$ is used to remove the double counting when $\mathcal{A}i \equiv \mathcal{B}j$ [see Appendix B for the derivation of (5)]. Through this recursion relation, all the constraints that $\mathbf{f}^{(1)}$ obeys are applicable for $\mathbf{f}^{(2)}$. We note that $\mathbf{f}^{(2)}$ was discussed in several previous works, e.g., in Refs. 25, 46, and 47 under the name of "bond counting". This fingerprint can also be regarded as a generalization of "doubles", the fingerprint defined in Ref. 20 for the chain models of polymers.

D. 3^{rd}-order fingerprint, $\mathbf{f}^{(3)}$

In the 3^{rd}-order fingerprint $\mathbf{f}^{(3)}$, the number of two-bond catenation is represented, i.e., $\kappa \equiv \mathcal{A}i\text{-}\mathcal{B}j\text{-}\mathcal{C}k$. In particular, the definition (1) for $f^{(3)}_{\kappa\equiv\mathcal{A}i\text{-}\mathcal{B}j\text{-}\mathcal{C}k}$ involves $n_{\mathcal{A}i\text{-}\mathcal{B}j\text{-}\mathcal{C}k}$, which is the number of $\mathcal{A}i$-$\mathcal{B}j$-$\mathcal{C}k$ sequences, or equivalently, the catenation of two bonds $\mathcal{A}i$-$\mathcal{B}j$ and $\mathcal{B}j$-$\mathcal{C}k$. Considering compounds of C, O, and H, there are 125 possible distinct catenation of two bonds $\mathcal{A}i$-$\mathcal{B}j$ and $\mathcal{B}j$-$\mathcal{C}k$. From $\mathbf{f}^{(3)}$, $\mathbf{f}^{(2)}$ can be determined as (see Appendix B)

$$\begin{aligned} f^{(2)}_{\mathcal{A}i\text{-}\mathcal{B}j} &= \sum_{\mathcal{C}k}\left[\frac{2^{\delta_{\mathcal{A}i,\mathcal{C}k}-1}}{j-1} f^{(3)}_{\mathcal{A}i\text{-}\mathcal{B}j\text{-}\mathcal{C}k}\right] \\ &= \sum_{\mathcal{C}k}\left[\frac{2^{\delta_{\mathcal{B}j,\mathcal{C}k}-1}}{i-1} f^{(3)}_{\mathcal{B}j\text{-}\mathcal{A}i\text{-}\mathcal{C}k}\right]. \end{aligned} \tag{6}$$

Similar to $\mathbf{f}^{(2)}$, $\mathbf{f}^{(3)}$ can be viewed as a generalization of "triples", the fingerprint examined in Ref. 20.

IV. PROPERTY PREDICTION MODEL

A learning model is critical in order to map the fingerprints to properties. In this work, we chose Gaussian kernel ridge regression (KRR),[5,48,49] the technique which has successfully been used in material properties predictions[20,25,28–30] Within this model, the input fingerprints are transformed into higher-dimensional space whereby a linear relation between the transformed fingerprints and the associated properties can be established. This mapping involves the distances between fingerprints and can be regarded as a similarity-based prediction model, i.e., similar properties may be predicted for materials with similar fingerprints.

In the KRR model, the property $\mathcal{P}_\mu$ of a structure μ is predicted as an weighted sum of Gaussians

$$\mathcal{P}_\mu = \sum_\nu \alpha_\nu \exp\left[-\frac{1}{2}\left(\frac{d_{\mu\nu}}{\sigma}\right)^2\right], \tag{7}$$

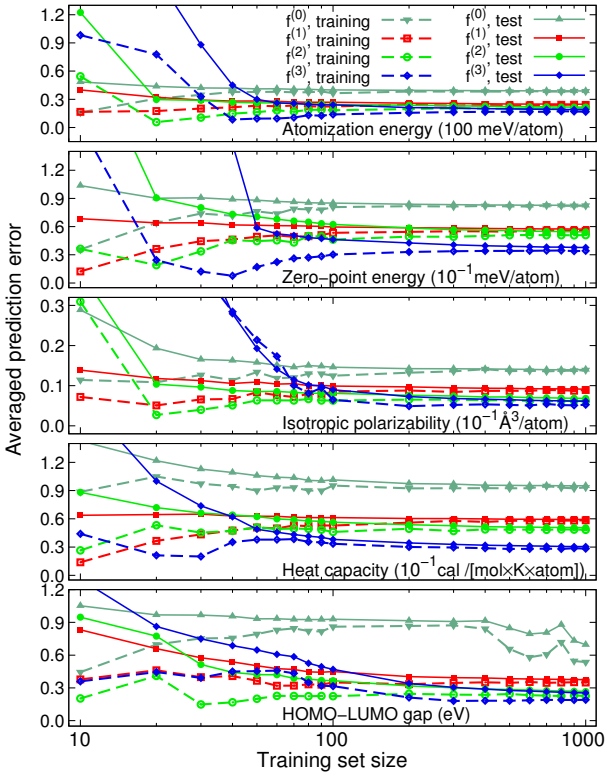

FIG. 2. (Color online) Learning curves corresponding to $\mathcal{E}_{\rm at}$, $\mathcal{E}_{\rm ZP}$, α, C_v, and $E_{\rm HL}$. For each model, $\mathbf{f}^{(0)}$, $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, and $\mathbf{f}^{(3)}$ are used to represent the molecules. Calculated data is given by symbols while curves are the guide for the eyes.

where ν runs over all the fingerprints in the training dataset. Here, $d_{\mu\nu}$ is the distance between fingerprints μ and ν, defined as the Euclidean metric $d_{\mu\nu} = \sqrt{\sum_\kappa \left(f_\kappa^\mu - f_\kappa^\nu\right)^2}$. The Gaussian width parameter σ and the regression coefficients α_ν are determined within the training phase whence a regularized objective function is minimized.[5,48,49] During this phase, σ and the regularization parameter are determined by k-fold cross validation on the training set ($k = 5$ in this work). Within this method, the training dataset is split into k bins, any of the bins is considered to be a new test dataset while the remaining $k-1$ bins form a new training datatest. This procedure is repeated for each of the k bins and for every value of σ and λ on a preselected logarithmic-scale grid. The optimal values of σ and λ, i.e., those leading to the minimum k-fold cross-validation (mean absolute) error, are used to compute α_ν of the entire dataset.

V. PROPERTY PREDICTION RESULTS

A. Molecules dataset

The four fingerprints considered, namely $\mathbf{f}^{(0)}$, $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, and $\mathbf{f}^{(3)}$, were used to represent the molecules dataset. To mimic the learning and prediction processes, the dataset was randomly partitioned into a training dataset and a test dataset. The KRR model was then trained on the training dataset using five-fold cross validation before predictions were made on the test dataset. We show in Fig. 2 the learning curves of $\mathcal{E}_{\rm at}$, $\mathcal{E}_{\rm ZP}$, α, C_v, and $E_{\rm HL}$, plotting the training and test errors against the number of molecules in the training dataset (data reported in this figure was averaged over 30 independent runs). In addition, predictions for the test dataset of 44,708 molecules after training the KRR model on a dataset of 1,000 molecules are shown in Fig. 3. As discussed in detail below, both Fig. 2 and Fig. 3 indicate that all of these properties can be very well predicted by using either $\mathbf{f}^{(2)}$ or $\mathbf{f}^{(3)}$, provided that the KRR model is trained on a training dataset of $\simeq 200$ or more data points.

The general tendency, as revealed by Fig. 2, is that higher-order fingerprints offer more accurate predictions. The $0^{\rm th}$-order fingerprint $\mathbf{f}^{(0)}$ can be used to roughly estimate energy-related quantities, i.e., $\mathcal{E}_{\rm at}$ and $\mathcal{E}_{\rm ZP}$ while it can not be used for others. For instance, $E_{\rm HL}$ can not be predicted with $\mathbf{f}^{(0)}$ because this fingerprint is totally local in nature, encoding no information at any finite range. Consequently, the finite conjugation length, known to signal the energy gap reduction in complex (conjugated) systems (see, for example Ref. 50), is not captured by $\mathbf{f}^{(0)}$. Fingerprints of higher orders, e.g., $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$ and $\mathbf{f}^{(3)}$, contain some information at increasing ranges, allowing for systematically better predicting $E_{\rm HL}$. These fingerprints also work sufficiently well in predicting $\mathcal{E}_{\rm at}$ and $\mathcal{E}_{\rm ZP}$. With $\mathbf{f}^{(1)}$, the averaged error in predicting $\mathcal{E}_{\rm at}$ is $\simeq 25$ meV/atom while this error is reduced to $\simeq 20$ meV/atom and $\simeq 18$ meV/atom if $\mathbf{f}^{(2)}$ and $\mathbf{f}^{(3)}$, respectively, are used. The very good power of $\mathbf{f}^{(2)}$ in predicting $\mathcal{E}_{\rm at}$ reproduces the similar conclusions drawn for the "bond counting" fingerprint by Ref. 47. This behavior is understandable because the dissociation energy of chemical bonds in organic molecules and crystals, which dominates the stability of these systems, are well-defined[46] in the same fashion with the bond length as previously discussed. Interestingly, this predictive power can significantly be improved if more advanced fingerprints, i.e., those can capture the small perturbations of interatomic distances like Coulomb matrix, are used.[29,30] Compared to $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$, $\mathbf{f}^{(3)}$ is significantly better in predicting C_v. The considerable improvement in the predictions of α when $\mathbf{f}^{(2)}$ is used instead of $\mathbf{f}^{(1)}$ may indicate the key contribution from polar bonds to the high-value regime of α.

B. Crystals dataset

We performed similar predictions for the dataset of 215 crystals containing C, O, and H. Using the KRR model coupled with $\mathbf{f}^{(0)}$, $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, and $\mathbf{f}^{(3)}$, five properties of these crystals, including the atomization energies $\mathcal{E}_{\rm at}$, the band gap $E_{\rm g}$, the electronic dielectric constant $\epsilon_{\rm elec}$, the

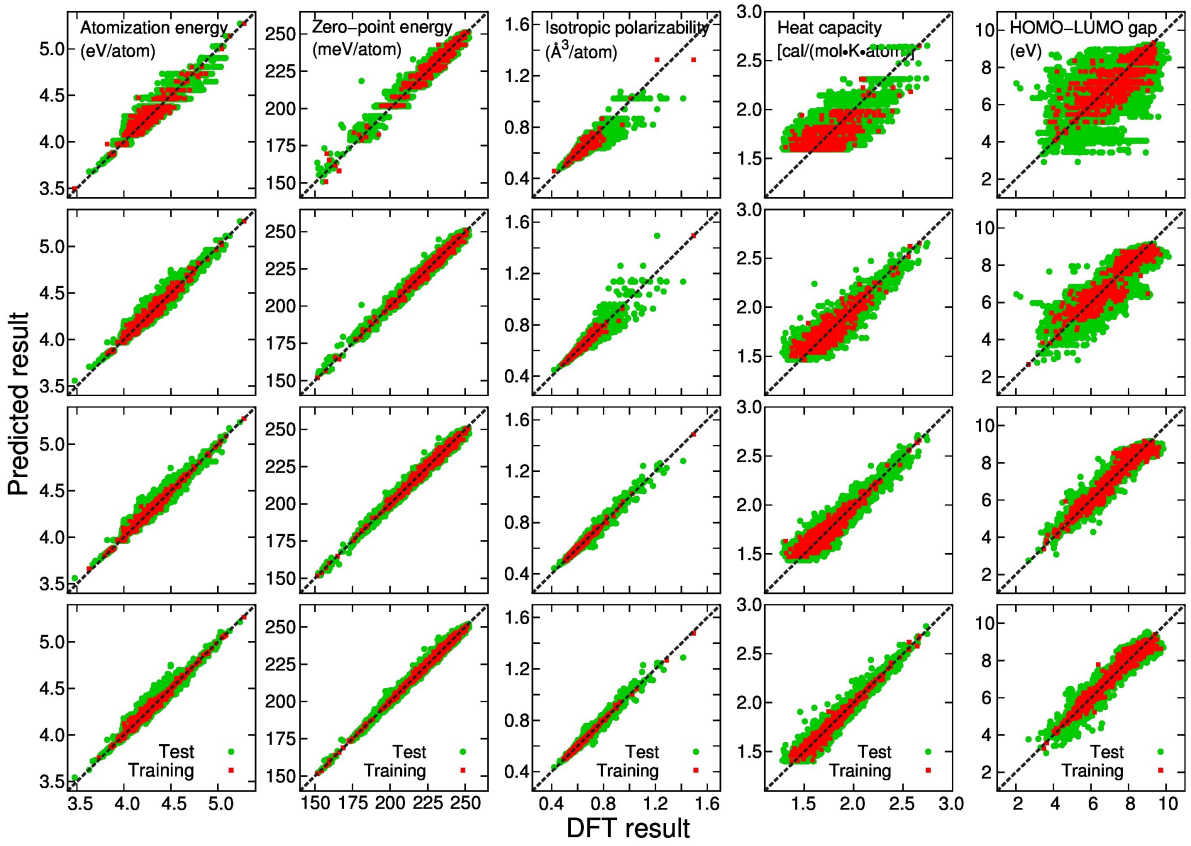

FIG. 3. (Color online) Predictions for $\mathcal{E}_{\rm at}$, $\mathcal{E}_{\rm ZP}$, α, C_v, and $E_{\rm HL}$ of the molecules dataset, using $\mathbf{f}^{(0)}$, $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, and $\mathbf{f}^{(3)}$ (from top row to bottom row). For each prediction, the training dataset consists of 1,000 points while the test dataset includes the remaining 44,708 data points.

ionic dielectric constant $\epsilon_{\rm ion}$, and the total dielectric constant $\epsilon_{\rm tot} = \epsilon_{\rm elec} + \epsilon_{\rm ion}$, were predicted. We show in Fig. 4 the learning curves, representing the errors of the predictions using these fingerprints, averaged over 100 independent runs. In Fig. 5, the predictions for the five properties are given, using the KRR model trained on a random training set of 150 data points.

Clearly, the tendency of the prediction performances on the crystals dataset is similar to those of the molecules dataset, i.e., high accuracies are obtained with fingerprints of higher orders, and properties which are governed by long-ranged information, e.g., band gap $E_{\rm g}$, can only be predicted with high-order fingerprints. For the atomization energy $\mathcal{E}_{\rm at}$, predictions with $\mathbf{f}^{(0)}$ and $\mathbf{f}^{(1)}$ leads to quite high averaged errors, which reduced to $\simeq 18$ meV/atom and $\simeq 15$ meV/atom when $\mathbf{f}^{(2)}$ and $\mathbf{f}^{(3)}$, respectively, were used. Overall, all the five examined properties can be predicted well when high-order fingerprints are used to represent the crystals. For instance, by employing $\mathbf{f}^{(3)}$, the averaged error in predicting $E_{\rm g}$ is $\simeq 0.45$ eV while the electronic dielectric constant $\epsilon_{\rm elec}$ and the ionic dielectric constant $\epsilon_{\rm ion}$ can be predicted with an averaged error of $0.1 - 0.2$.

VI. UTILITIES OF THE FINGERPRINTS

The demonstrated predictive power of the KRR model, which uses $\mathbf{f}^{(i)}$ to represent materials structures, inspires the idea of using this model to rationally optimize materials for a targeted property $\mathcal{P}_{\rm opt}$, the concept often referred to as "inverse design".[51–54] In fact, a large number of success stories along this direction have been reported in the past, using various approaches, e.g., iteratively optimizing the properties of a given compound or *on-the-fly* screening when searching for stable structures.[55–67] Here, our idea is that starting from a trained KRR model, fingerprints which correspond to the desired properties can be predicted. Then, molecular structures will be reconstructed from the predicted fingerprints. Finally, the targeted properties will be verified by DFT calculations at the same level with those used for the training dataset.

The greatest challenge of this procedure is to ensure that the predicted fingerprint is physically and chemically meaningful, i.e., at least one material structure can be reconstructed from it.[68,69] Therefore, one must mathematically define the subspace of the meaningful fingerprints, and then limit the search for desired fingerprints

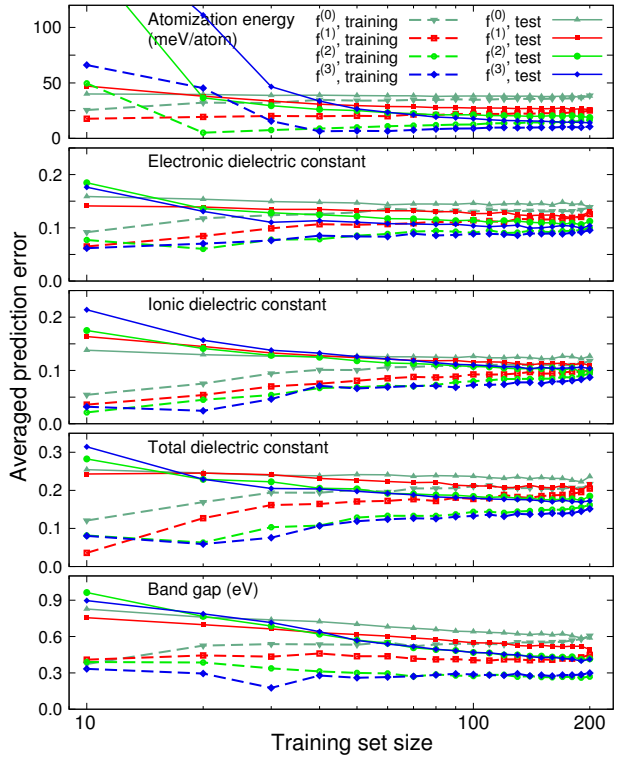

FIG. 4. (Color online) Learning curves corresponding to $\mathcal{E}_{\rm at}$, $\epsilon_{\rm elec}$, $\epsilon_{\rm ion}$, ϵ, and $E_{\rm g}$ determined by using $\mathbf{f}^{(0)}$, $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, and $\mathbf{f}^{(3)}$ for representing the crystals structures. Calculated data is shown by symbols while curves are the guide for the eyes.

within this subspace. We present two approaches which can be used for designing molecules (the work of designing crystals is not considered here).

A. Design via enumeration

The central idea of this approach is that the components of a given fingerprint can be enumerated in a given way so that it is meaningful. We used $\mathbf{f}^{(2)}$ for a demonstration because predictions using this fingerprint are good while its dimensionality is not too high like $\mathbf{f}^{(3)}$. We first implemented the applicable rules involving bonds and coordination numbers by defining five "backbone" blocks. They include C4, C=C (a pair of C3 atoms with a double bond), C≡C (a pair of C2 atoms with a triple bond), C=O (one C3 and one O1 atom linked by a double bond), and O2. By definition, all of the dangling bonds starting from these blocks are single, thus any of them can be connected to others without any constraint. Then, given a set of backbone blocks, all the possible arrangements can be scanned, keeping track of the connectivity to eliminate some dangling bonds, and saturating the remaining dangling bonds by either H1 or OH, referred to as "ending" blocks. From the obtained arrangements, $\mathbf{f}^{(2)}$ can be unambiguously determined and their properties were predicted. Those with targeted properties were singled out to rebuild molecular structures for validating calculations. We show in Fig. 6 two optimized molecules constructed from two of the predicted fingerprints, labeled by A and B, accompanied by the predicted and calculated $E_{\rm HL}$ and α. The results given in Fig. 6 indicate that the desired molecules are indeed obtained.

B. Design via inversion

Different from the enumeration approach, this procedure aims to directly determine the fingerprints, starting from desired properties. This goal can be achieved by optimizing an objective function, aiming towards the desired properties while applying the constraints that ensure the fingerprints considered are meaningful. Because the reconstruction step requires a simple enough fingerprint, $\mathbf{f}^{(1)}$ was selected for this approach. Among the constraints established for $\mathbf{f}^{(1)}$, (2) and (3) are explicitly imposed in the objective function defined below

$$\begin{aligned} G[\mathbf{f}^{(1)},\lambda_1,\lambda_2] &= (\mathcal{P}-\mathcal{P}_{\rm opt})^2+\lambda_1\left[\sum_{\mathcal{A}i} f^{(1)}_{\mathcal{A}i}-1\right]^2 \\ &+\lambda_2\left[f^{(1)}_{\rm H1}-2f^{(1)}_{\rm C4}-f^{(1)}_{\rm C3}+f^{(1)}_{\rm O1}\right]^2. \end{aligned} \quad (8)$$

Here, λ_1, and λ_2 are the Lagrange multipliers associated with the constraints while $\mathcal{P}$ is the property (or properties) of the trial fingerprint $\mathbf{f}^{(1)}$ predicted by the trained KRR model. In practice, we evaluated $\mathcal{P}$ by averaging many predictions, each of them was given by the KRR model trained on a randomly selected training dataset of 1,000 data points. All the terms in (8) are given in the quadratic form to smoothen G. Generally, the problem of minimizing $G[\mathbf{f}^{(1)},\lambda_1,\lambda_2]$ (performed with simulated annealing[70] in this work) returns many solutions $\mathbf{F}^{(1)}$. For each of them, $N_{\rm at}$ was determined by minimizing another objective function $D[\mathbf{F}]$ defined as

$$D[\mathbf{F}^{(1)}]=\sum_{\mathcal{A}i}\left[N_{\rm at}F^{(1)}_{\mathcal{A}i}-{\rm nint}\left(N_{\rm at}F^{(1)}_{\mathcal{A}i}\right)\right]^2, \quad (9)$$

where ${\rm nint}(x)$ returns the closest integer to x. Once $N_{\rm at}$ is determined, a post-screening step is performed to consider the possibility of $N_{\bigcirc}>0$ and to single out the fingerprints so that $N_{\rm at}F^{(1)}_{\rm C2}$ and $N_{\rm at}\left[F^{(1)}_{\rm C3}-F^{(1)}_{\rm O1}\right]$ are positive even numbers. Such fingerprints are meaningful, i.e., molecules can be built up from any of them.

We demonstrate this procedure by optimizing two properties simultaneously, i.e., the HOMO-LUMO gap $E_{\rm HL}$ and the isotropic polarizability α. Because α is a measure of the response, in terms of charge redistribution, of a molecule to the external electric field, this quantity is closely related to the electronic contribution of the dielectric constant $\epsilon_{\rm elec}$. We note that these properties seem to be competing, as shown in Fig. 7 where

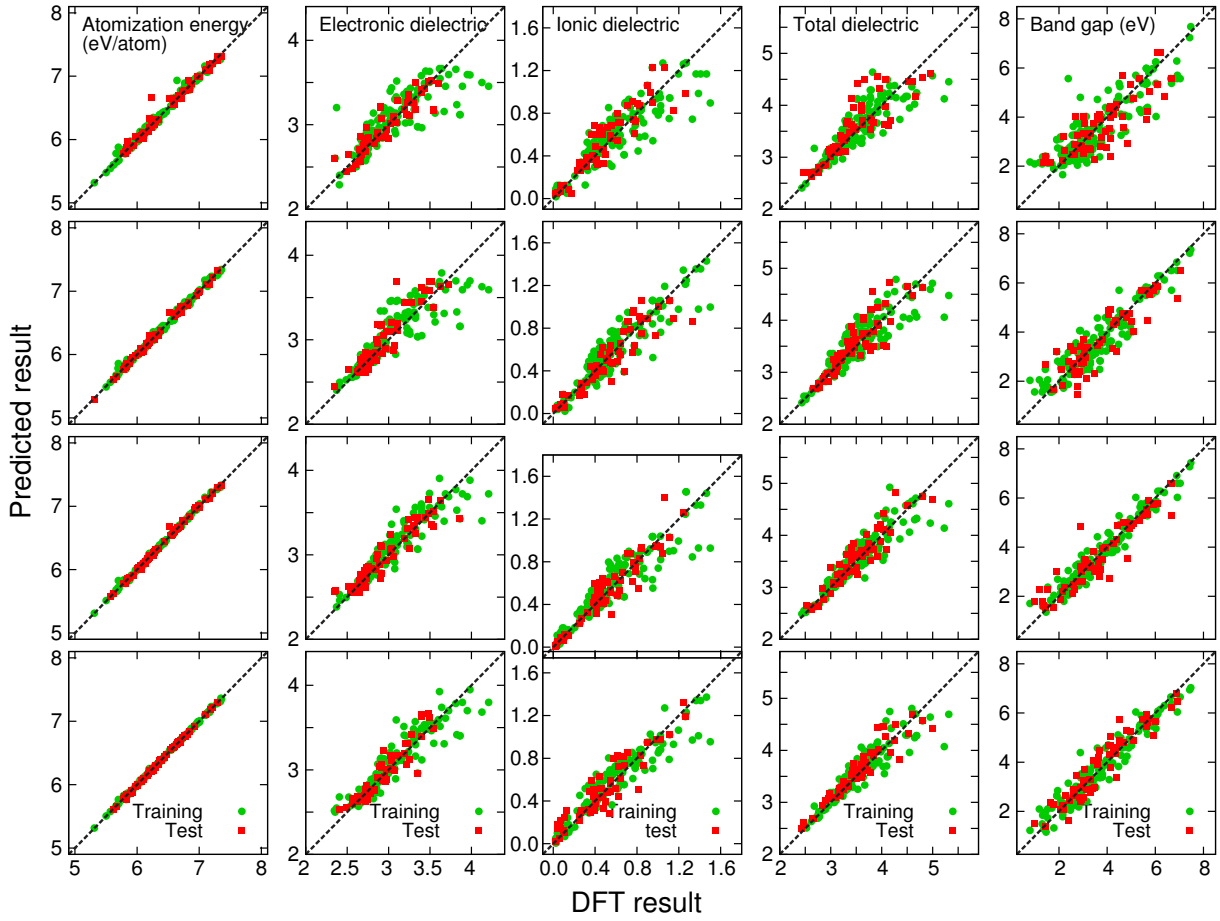

FIG. 5. (Color online) Predictions for $\mathcal{E}_{\rm at}$, $\epsilon_{\rm elec}$, $\epsilon_{\rm ion}$, ϵ, and $E_{\rm gap}$ of the crystals dataset, using $\mathbf{f}^{(0)}$, $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, and $\mathbf{f}^{(3)}$ (from top row to bottom row). For each prediction, the training set size is 150 and the remaining 70 points form the test set.

an asymptotic limit of the form $\alpha \sim 1/E_{\rm HL}$ can be seen (similar limit between two related properties of crystals, namely $\epsilon_{\rm elec}$ and $E_{\rm g}$ was documented earlier in Ref. 71). An examination of Fig. 3 reveals that the prediction of α using $\mathbf{f}^{(1)}$ is fairly good in the region of $\alpha < 0.8$ Å^3/atom. For this reason, we searched for new molecules, i.e., those that do not exist in the molecules dataset, of which $0.6 \leq \alpha \leq 0.7$ Å^3/atom while $E_{\rm HL} \geq 7$ eV and show the results in Fig. 7. While the calculated $E_{\rm HL}$ of the molecules dataset can reach the upper limit of $\simeq 10$ eV, all the predictions for $E_{\rm HL}$ by the KRR model are below 9 eV. The reason is given in Fig. 3 which clearly implies that when $\mathbf{f}^{(1)}$ is coupled with the KRR model, high values of $E_{\rm HL}$ ($8 \leq E_{\rm HL} \leq 10$ eV) are generally underestimated by roughly 1 eV. Three of the predicted fingerprints, labeled by C, D, and E, were selected for rebuilding new molecules. From either C or E, only one molecule can be constructed while many different molecules correspond to D. All of the molecules reconstructed from C, D, and E were optimized and then their α and $E_{\rm HL}$ were calculated with Gaussian 09,[72] using the 6-31G(2df,p) basis set and the B3LYP XC functional.[73,74] The results are summarized in Table I and in the inset of Fig. 7, demonstrating that the molecules with desired values of α and $E_{\rm HL}$ were actually obtained. Detailed information on all of the designed molecules can be found in the Supplemental Material.[42]

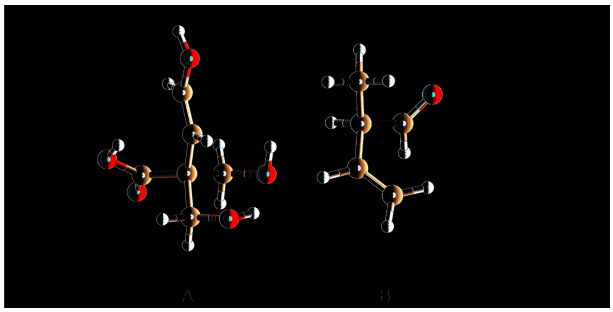

FIG. 6. (Color online) Optimized molecules, constructed from two predicted fingerprints A and B, shown with the predicted and calculated values of $E_{\rm HL}$ and α. Carbon, oxygen, and hydrogen atoms are given in dark brown, red, and pink.

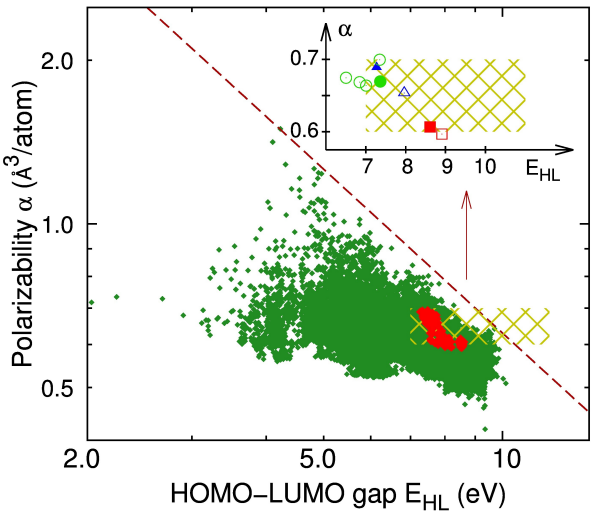

FIG. 7. (color online) $E_{\rm HL} - \alpha$ log-log plot of the molecules dataset, shown by forest-green symbols while the predicted fingerprints are shown by red diamonds within the regime of desired properties, i.e., $0.6 \leq \alpha \leq 0.7$Å^3/atom and $E_{\rm HL} \geq 7.0$ eV. In the inset, the predicted and calculated properties of the molecules reconstructed from three predicted fingerprints, i.e., C, D, and E, are shown by closed and open symbols: triangles for C, circles for D, and squares for E. The dashed line sketches the limit $\alpha \sim 1/E_{\rm HL}$ addressed in the text.

C. Remarks

It is worth noting that the key feature of $\mathbf{f}^{(i)}$ which is useable for the described enumeration and inversion design procedures is their discontinuity with respect to slight configurational perturbations. Because all the possible chemical bonds appearing in a molecule comprising C, O, and H are well-defined, it is very likely that the optimization step performed on the reconstructed molecules preserves the predicted fingerprint. Moreover, the efficiency of the designing approaches depends on several factors, including the prediction accuracy of the fingerprints used. Although predictions by using high-order fingerprints are systematically better, the complexity generated by their high dimensionality is significant. Comparing to the procedure described above, that utilizing $\mathbf{f}^{(2)}$ or $\mathbf{f}^{(3)}$ needs roughly 10 and 100 more constraints for ensuring the considered fingerprints are meaningful. If the dimensionality of $\mathbf{f}^{(2)}$ can considerably be reduced, it may then be used for the inversion approach.

VII. CONCLUSIONS

To summarize, we have systematically studied a family of motif-based topological fingerprints which can numerically represent major classes of molecules and crystals. By using a similarity based learning algorithm, these fingerprints can be mapped onto various properties of molecules and crystals, significantly accelerating their properties prediction. A major advantage of these fingerprints is clearly demonstrated via two procedures for designing molecules, one by enumeration and the other by inversion. These procedures rely on the accelerated properties prediction to identify the desired fingerprints, and then to reconstruct molecules that possess one or more targeted properties. We note that although only molecules and crystals comprising C, O, and H are considered in this contribution, our results can straightforwardly be generalized to those containing other light elements whose coordination preferences are well established, e.g., N and F.

TABLE I. Predicted and calculated values of α (in Å^3/atom) and $E_{\rm HL}$ (in eV) of the molecules designed from three predicted fingerprints C, D, and E. Data from this Table is also shown in the inset of Fig. 7.

Label	$N_{\rm at}$	Predicted		Calculated	
		α	$E_{\rm HL}$	α	$E_{\rm HL}$
C	11	0.689	7.273	0.654	7.964
D	18	0.670	7.363	$0.664 - 0.699$	$6.502 - 7.348$
E	14	0.607	8.612	0.597	8.909

ACKNOWLEDGMENTS

The authors thank Venkatesh Botu, Ghanshyam Pilania, and Vinit Sharma for useful discussions and O. Anatole von Lilienfeld for drawing our attention to some important relevant works. The present work was supported by a Multi-University Research Initiative (MURI) grant from the Office of Naval Research, under award number N00014100944. Computational work was made possible through the XSEDE computational resource allocation number TG-DMR080058N.[75]

Appendix A: Constraint of $\mathbf{f}^{(1)}$ derived from elementary chemical rules

Constraint (3) was derived with an assumption that the desired molecular structure is connected, i.e., any pair of atoms are connected by at least one sequence of the allowed chemical bonds. Let us take a molecule in which $n_{\mathcal{A}i}$ is the number of the blocks $\mathcal{A}i$. Starting from the applicable chemical rules, all the two-fold coordinated carbon atoms are grouped by pairs, forming $n_{\rm C2}/2$ units of C $\equiv$ C, each of which is a pair of carbon atoms linked by a triple bond. Next, $n_{\rm O1}$ one-fold coordinated oxygen atoms must bond with $n_{\rm O1}$ three-fold coordinated carbon atoms to form $n_{\rm O1}$ units of C = O. Then, the remaining $n_{\rm C3} - n_{\rm O1}$ three-fold coordinated carbon atoms are grouped together by pairs, forming $(n_{\rm C3} - n_{\rm O1})/2$ units of C = C. Therefore, the set of the blocks $\mathcal{A}i$ now contains $n_{\rm C2}/2 + n_{\rm O1} + (n_{\rm C3} - n_{\rm O1})/2 + n_{\rm C4} + n_{\rm O2}$ units of C $\equiv$ C,

CO, C = C, C4 and O2. Assuming that these units are isolated, the total number of dangling bonds starting from them is $2(n_{\mathrm{C2}}/2)+2n_{\mathrm{O1}}+4[(n_{\mathrm{C3}}-n_{\mathrm{O1}})/2]+4n_{\mathrm{C4}}+2n_{\mathrm{O2}}$, or simply

$$n_{\mathrm{C2}} + 2n_{\mathrm{C3}} + 4n_{\mathrm{C4}} + 2n_{\mathrm{O2}}. \tag{A1}$$

By joining $n_{\mathrm{C2}}/2+n_{\mathrm{O1}}+(n_{\mathrm{C3}}-n_{\mathrm{O1}})/2+n_{\mathrm{C4}}+n_{\mathrm{O2}}$ units together, the number of dangling bonds that will be annihilated to form inter-unit bonds is $2[n_{\mathrm{C2}}/2+n_{\mathrm{O1}}+(n_{\mathrm{C3}}-n_{\mathrm{O1}})/2 + n_{\mathrm{C4}} + n_{\mathrm{O2}} - 1] + 2n_{\bigcirc}$ where $n_{\bigcirc}$ is the number of loops of bonds, each of which costs extra 2 bonds. Therefore, the number of remaining dangling bonds is

$$n_{\mathrm{C3}} + 2n_{\mathrm{C4}} - n_{\mathrm{O1}} - 2n_{\bigcirc} + 2. \tag{A2}$$

All of these dangling bonds must be saturated by n_{H1} hydrogen atoms, thus

$$n_{\mathrm{H1}} = n_{\mathrm{C3}} + 2n_{\mathrm{C4}} - n_{\mathrm{O1}} - 2n_{\bigcirc} + 2. \tag{A3}$$

The constraint (3) can then be obtained when we divide Eq. (A3) by N_{at}. This constraint is applicable not only for molecules but also for crystals formed by repeatedly placing an isolated molecule in a periodic grid. If these molecules are not isolated, i.e., they form a network of d dimensions, $2d$ dangling bonds are used to form the network (assuming that the network are formed only by single bonds). Thus, Eq. A3 is given as

$$n_{\mathrm{H1}} = n_{\mathrm{C3}} + 2n_{\mathrm{C4}} - n_{\mathrm{O1}} - 2n_{\bigcirc} - 2d + 2. \tag{A4}$$

In the general case when not only single bonds involve the network formation, the parameter d used in Eq. A4 is not necessarily an integer.

Appendix B: Derivation of the recursion relations of $\mathbf{f}^{(2)}$ and $\mathbf{f}^{(3)}$

1. Recursion relations of $\mathbf{f}^{(2)}$

The number $n_{\mathcal{A}i}$ of blocks $\mathcal{A}i$ can be determined by counting all the bonds of $\mathcal{A}i$-$\mathcal{B}j$ type. By summing all the number of $\mathcal{A}i$-$\mathcal{B}j$ bonds, the $\mathcal{A}i$-$\mathcal{A}i$ bonds are counted twice. Therefore

$$n_{\mathcal{A}i} = \frac{1}{i}\left[\sum_{\mathcal{B}j} n_{\mathcal{A}i\text{-}\mathcal{B}j} - \frac{1}{2} n_{\mathcal{A}i\text{-}\mathcal{A}i}\right]. \tag{B1}$$

Then, the recursion relation of $\mathbf{f}^{(2)}$ can be obtained by dividing (B1) by the total number of atoms N_{at}.

2. Recursion relations of $\mathbf{f}^{(3)}$

Similar to the derivation of (B1), the fingerprint component $f^{(2)}_{\mathcal{A}i\text{-}\mathcal{B}j}$ can be determined by counting the number of $\mathcal{A}i$-$\mathcal{B}j$-$\mathcal{C}k$ sequences before dividing by $j-1$. In such a procedure, the $\mathcal{A}i$-$\mathcal{B}j$-$\mathcal{A}i$ sequences are counted twice. Thus, after removing the double counting, we obtain

$$n_{\mathcal{A}i\text{-}\mathcal{B}j} = \frac{1}{j-1}\left[\sum_{\mathcal{C}k} n_{\mathcal{A}i\text{-}\mathcal{B}j\text{-}\mathcal{C}k} - \frac{1}{2} n_{\mathcal{A}i\text{-}\mathcal{B}j\text{-}\mathcal{A}i}\right]. \tag{B2}$$

We note that one can also count the number of $\mathcal{B}j$-$\mathcal{A}i$-$\mathcal{C}k$ sequences before dividing the total number by $i-1$. Thus

$$n_{\mathcal{A}i\text{-}\mathcal{B}j} = \frac{1}{i-1}\left[\sum_{\mathcal{C}k} n_{\mathcal{B}j\text{-}\mathcal{A}i\text{-}\mathcal{C}k} - \frac{1}{2} n_{\mathcal{B}j\text{-}\mathcal{A}i\text{-}\mathcal{B}j}\right]. \tag{B3}$$

By dividing (B2) and (B3) by N_{at}, two equivalent recursion relations are obtained. Moreover, we note that (B2) and (B3) set up a constraint that $\mathbf{f}^{(3)}$ must also satisfy.

* rampi@ims.uconn.edu

[1] G. Hautier, A. Jain, and S. Ong, J. Mater. Sci. **47**, 7317 (2012).
[2] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nat. Matter. **12**, 191 (2013).
[3] V. Sharma, C. C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs, and R. Ramprasad, Nat. Commun. **5**, 4845 (2014).
[4] T. Mueller, A. G. Kusne, and R. Ramprasad (unpublished).
[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer-Verlag, New York, 2009).
[6] C. M. Breneman and M. Rhem, J. Comput. Chem. **18**, 182 (1997).
[7] N. Sukumar, M. Krein, Q. Luo, and C. Breneman, J. Mater. Sci. **47**, 7703, (2012).
[8] T. Le, V. C. Epa, F. R. Burden, and D. A. Winkler, Chem. Rev. **112**, 2889 (2012).
[9] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Phys. Rev. Lett. **91**, 135503 (2003).
[10] K. Rajan, Mater. Today **8**, 38 (2005).

[11] J. C. Schön, Z. Anorg. All. Chem. **640**, 2717 (2014).
[12] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Phys. Rev. B **89**, 094104 (2014).
[13] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, J. Chem. Theor. Comput. **9**, 3404 (2013).
[14] R. Guha and A. Bender, *Computational Approaches in Cheminformatics and Bioinformatics* (John Willey & Sons, New York, 2011).
[15] A. Varnek, in *Chemoinformatics and Computational Chemical Biology*, Vol. 672 of *Methods in Molecular Biology*, edited by J. Bajorath (Humana Press, New York, NY, 2011), pp. 213–243.
[16] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Materials **1**, 011002, (2013).
[17] G. Bergerhoff, I. Brown, F. Allen, G. Bergerhoff, and R. Sievers, *Crystallographic Databases* (International Union of Crystallography, Chester, 1987).
[18] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail, Nucleic Acids Res. **40**, D420 (2012).
[19] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Sci. Data **1**, 140022 (2014).
[20] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Sci. Rep. **3**, 2810 (2013).
[21] A. N. Andriotis, G. Mpourmpakis, S. Broderick, K. Rajan, S. Datta, M. Sunkara, and M. Menon, J. Chem. Phys. **140**, 094705 (2014).
[22] H. C. Dam, T. L. Pham, T. B. Ho, A. T. Nguyen, and V. C. Nguyen, J. Chem. Phys. **140**, 044101 (2014).
[23] R. D. Brown and Y. C. Martin, J. Chem. Inf. Comput. Sci. **36**, 572 (1996).
[24] R. D. Brown and Y. C. Martin, J. Chem. Inf. Comput. Sci. **37**, 1 (1997).
[25] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[26] K. Hansen, F. Biegler, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko (unpublished).
[27] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, Int. J. Quant. Chem. (2015), accepted.
[28] V. Botu and R. Ramprasad, Int. J. Quant. Chem. (2015), DOI: 10.1002/qua.24836.
[29] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, arXiv:1503.04987 (unpublished).
[30] R. Ramakrishnan and O. A. von Lilienfeld, Chimia **69**, 182 (2015).
[31] S. Goedecker, in *Modern Methods of Crystal Structure Prediction*, edited by A. R. Oganov (Wiley-VCH, Weinheim, Germany, 2011), Chap. 7, pp. 147–180.
[32] S. Goedecker, J. Chem. Phys. **120**, 9911 (2004).
[33] M. Amsler and S. Goedecker, J. Chem. Phys. **133**, 224104 (2010).
[34] C. W. Glass, A. R. Oganov, and N. Hansen, Comput. Phys. Commun. **175**, 713 (2006).
[35] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
[36] G. Kresse, Ph.D. thesis, Technische Universität Wien, 1993.
[37] G. Kresse and Furthmüller, J. Comput. Mater. Sci. **6**, 15 (1996).
[38] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[39] E. D. Murray, K. Lee, and D. C. Langreth, J. Chem. Theor. Comput. **5**, 2754 (2009).
[40] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).
[41] K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, Phys. Rev. B **82**, 081101(R) (2010).
[42] See Supplemental Material for more information reported in this paper.
[43] L. Pauling, J. Am. Chem. Soc. **54**, 3570 (1932).
[44] F. H. Allen, O. Kennard, D. G. Watson, L. Brammer, A. G. Orpen, and R. Taylor, J. Chem. Soc., Perkin Trans. II S1 (1987).
[45] F. H. Allen, D. G. Watson, L. Brammer, A. G. Orpen, and R. Taylor, in *International Tables for Crystallography, Mathematical, Physical and Chemical Tables*, 3rd ed., edited by E. Prince (Kluwer Academic Publishers, Norwell, MA, USA, 2004), Chap. 9.5.
[46] S. W. Benson, J. Chem. Educ. **42**, 502 (1965).
[47] J. Moussa, Phys. Rev. Lett. **109**, 059801 (2012).
[48] T. Hofmann, B. Schlkopf, and A. J. Smola, Ann. Stat. **36**, 1171 (2008).
[49] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, IEEE Trans. Neural Netw. **12**, 181 (2001).
[50] P. Stallinga, *Electrical characterization of organic electronic materials and devices* (John Wiley & Sons, West Sussex, UK, 2009).
[51] G. Ceder, Y. M. Chiang, D. R. Sadoway, M. K. Aydinol, Y. I. Jang, and B. Huang, Nature **392**, 694 (1998).
[52] F. Besenbacher, I. Chorkendorff, B. S. Clausen, B. Hammer, A. M. Molenbroek, J. K. Nrskov, and I. Stensgaard, Science **279**, 1913 (1998).
[53] A. Franceschetti and A. Zunger, Nature **402**, 60 (1999).
[54] T. Weymuth and M. Reiher, Int. J. Quant. Chem. **114**, 823 (2014).
[55] O. A. von Lilienfeld, R. D. Lins, and U. Rothlisberger, Phys. Rev. Lett. **95**, 153002 (2005).
[56] V. Marcon, O. A. von Lilienfeld, and D. Andrienko, J. Chem. Phys. **127**, 064305 (2007).
[57] O. A. von Lilienfeld, J. Chem. Phys. **131**, (2009).
[58] D. Sheppard, G. Henkelman, and O. A. von Lilienfeld, J. Chem. Phys. **133**, 084104 (2010).
[59] M. Wang, X. Hu, D. N. Beratan, and W. Yang, J. Am. Chem. Soc. **128**, 3228 (2006).
[60] S. Keinan, X. Hu, D. N. Beratan, and W. Yang, J. Phys. Chem. A **111**, 176 (2007).
[61] S. Keinan, W. D. Paquette, J. J. Skoko, D. N. Beratan, W. Yang, S. Shinde, P. A. Johnston, J. S. Lazo, and P. Wipf, Org. Biomol. Chem. **6**, 3256 (2008).
[62] B. C. Rinderspacher, J. Andzelm, A. Rawlett, J. Dougherty, D. N. Beratan, and W. Yang, J. Chem. Theor. Comput. **5**, 3321 (2009).
[63] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Phys. Rev. Lett. **91**, 135503 (2003).
[64] J. Greeley and M. Mavrikakis, Nat. Mater. **3**, 810 (2004).
[65] M. d'Avezac, J.-W. Luo, T. Chanier, and A. Zunger, Phys. Rev. Lett. **108**, 027401 (2012).
[66] H. J. Xiang, B. Huang, E. Kan, S.-H. Wei, and X. G. Gong, Phys. Rev. Lett. **110**, 118702 (2013).
[67] C. L. Phillips and G. A. Voth, Soft Matter **9**, 8552 (2013).
[68] O. A. von Lilienfeld, Int. J. Quant. Chem. **113**, 1676 (2013).
[69] O. A. von Lilienfeld and M. E. Tuckerman, J. Chem. Phys. **125**, 154104 (2006).

[70] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
[71] C. C. Wang, G. Pilania, S. A. Boggs, S. Kumar, C. Breneman, and R. Ramprasad, Polymer **55**, 979 (2014).
[72] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, T. Ishida, Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, *Gaussian 09, Revision A.02*, Gaussian, Inc., Wallingford CT, 2009.
[73] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[74] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, J. Phys. Chem. **98**, 11623 (1994).
[75] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, Comput. Sci. Eng. **16**, 62 (2014).