npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Empirical modeling of dopability in diamond-like semiconductors

Samuel A. Miller[1], Maxwell Dylla[1], Shashwat Anand[1], Kiarash Gordiz[2], G. Jeffrey Snyder[1] and Eric S. Toberer[2,3]

Carrier concentration optimization has been an enduring challenge when developing newly discovered semiconductors for applications (e.g., thermoelectrics, transparent conductors, photovoltaics). This barrier has been particularly pernicious in the realm of high-throughput property prediction, where the carrier concentration is often assumed to be a free parameter and the limits are not predicted due to the high computational cost. In this work, we explore the application of machine learning for high-throughput carrier concentration range prediction. Bounding the model within diamond-like semiconductors, the learning set was developed from experimental carrier concentration data on 127 compounds ranging from unary to quaternary. The data were analyzed using various statistical and machine learning methods. Accurate predictions of carrier concentration ranges in diamond-like semiconductors are made within approximately one order of magnitude on average across both *p*- and *n*-type dopability. The model fit to empirical data is analyzed to understand what drives trends in carrier concentration and compared with previous computational efforts. Finally, dopability predictions from this model are combined with high-throughput quality factor predictions to identify promising thermoelectric materials.

npj Computational Materials (2018)4:71 ; doi:10.1038/s41524-018-0123-6

INTRODUCTION

Control of charge-carrier concentration of semiconducting materials is vitally important in a variety of applications, including photovoltaics,[1,2] optoelectronics,[3,4] transistors,[5,6] and thermoelectrics (TE).[7,8] To maximize efficiency, many of these applications require tuning both the type (*p*- or *n*-type) as well as the concentration of carriers. For many well-studied systems, the methods of controlling the carrier concentration are well-established, both in choice of dopant species and synthetic technique.[9,10] However, the control of carrier concentration is not well-understood in novel material systems. Traditionally, experimentalists have relied on basic metrics to guide the choice of doping species, namely ionic charge counting and radius ratio "rules of thumb."[11] These rules may not directly translate to more complex chemistries and structures. Theorists have recently been able to better guide efforts using defect calculations as computational capabilities have improved.[12–16] Despite these improvements, issues remain in the widespread use of these calculations due to their computational costs and inaccuracy. Therefore, methods to address dopability are critical for advances in complex semiconductors.

One such example is the high-throughput prediction of material properties, which has become increasingly common in the TE community.[17–20] Various groups have developed their own models which can predict the optimal potential thermoelectric performance for a material based on its structure and simple density functional theory (DFT) calculations. One of these metrics, quality factor (β), is a descriptor for the potential of a material to exhibit high thermoelectric performance, and has been previously shown to track well with experimental TE performance.[21] Despite the reasonable accuracy of these models in predicting the potential of a compound's performance, they rely on a key assumption. In order to predict this quality factor, it must be assumed that the chemical potential (i.e., Fermi level) can be sufficiently tuned to the type and concentration of charge carrier that optimizes performance. In the absence of dopability guidance, experimental investigation of high β compounds is inefficient due to the large number of false positive compounds that cannot be doped.

In the discussion of dopability, we find it helpful to identify the distinct sources that limit dopability. The discussion here is in relation to *p*-type dopability, but the schematic and discussion for *n*-type would be a mirror image. In the first case, Fig. 1a, the red native donor defect represents a hole "killer" defect, one that prevents the Fermi level (E_F) from being driven beyond some energy range, as it spontaneously produces an electron which increases E_F. This donor defect pins the minimum thermodynamically achievable limit of the Fermi level ($E_{F,lim}$) at the location of the red tick, with the possible doping range shown by the red horizontal gradient bar. As the native donor energy ($E_{n,d}$) increases, $E_{F,lim}$ moves toward the valence band, allowing a larger possible doping range shown by the green bar. Eventually, the donor energy is great enough such that the native donor dopability window ($W_{n,d}$) becomes positive, allowing greater *p*-type carrier concentration. Beyond killer defects, a system may exhibit limited dopability due to the lack of chemical flexibility in the native structure or extrinsic dopants. The Fermi level of the material will be set near the intersection of the lowest energy acceptor and donor defects, regardless of whether they are native or extrinsic (Fig. 1b). In some cases, the lowest energy extrinsic acceptor dopant is too high to substantially lower the Fermi level, meaning there is no extrinsic dopant which increases the dopability window,[22] represented with the red extrinsic acceptor.

[1]Northwestern University, Evanston, IL 60208, USA; [2]Colorado School of Mines, Golden, CO 80401, USA and [3]National Renewable Energy Laboratory, Golden, CO 80401, USA
Correspondence: G Jeffrey Snyder (jeff.snyder@northwestern.edu) or Eric S. Toberer (etoberer@mines.edu)

Received: 17 July 2018 Accepted: 29 October 2018
Published online: 06 December 2018

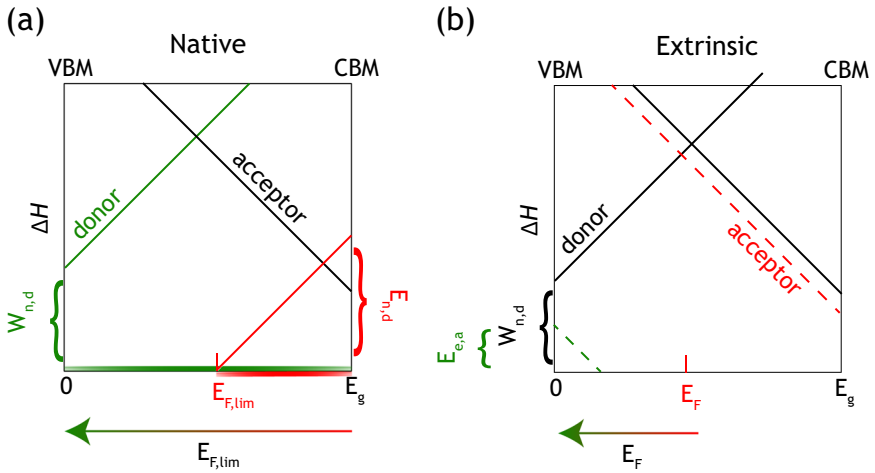

Fig. 1 **a** Defect diagram schematic showing native defects, including an acceptor defect (black line) and two possible variations of a native donor defect (red and green lines). The intersection at the valence band maximum (VBM) of the native donor defect gives the *p*-type dopability window ($W_{n,d}$). The achievable thermodynamic limit of the Fermi level ($E_{F,lim}$) is set by the charge (which determines slope) of the native donor defect and the conduction band minimum (CBM) defect energy (native donor energy or $E_{n,d}$). **b** Defect diagram schematic showing the effect of extrinsic dopants (dashed colored lines), given native acceptor and donor defects (solid black lines). The Fermi level will be near the intersection of the lowest energy donor and acceptor defects. The red extrinsic acceptor is a poor dopant as it does not significantly lower E_F. A good *p*-type dopant is one where the extrinsic acceptor energy ($E_{e,a}$) is less than or equal to $W_{n,d}$ ($W_{n,d} - E_{e,a} \geq 0$), allowing high *p*-type carrier concentration

This arises when there is significant phase competition for the dopant element and dopant solubility limits are reached. In more well-behaved systems, high dopant solubility is achieved due to a lack of phase competition and the minimal energetic penalty for dopant incorporation. This scenario yields a dopant where the energy of the extrinsic acceptor ($E_{e,a}$) is lower than the window of the native donor ($W_{n,d}$), drastically altering the Fermi level toward or into the valence band, leading to an associated high *p*-type carrier concentration (green extrinsic defect).

To date, there have been few efforts to model charge-carrier concentrations, and no comprehensive, analytical/physical model exists to estimate the dopability of materials. Conventional wisdom posits that large band gap materials are harder to dope, elemental properties such as size and electronegativity should be considered when choosing a substituting species, and that the structure and lattice energy have some effect.[11] Yet little is known about the relationship (sign, magnitude, functional form) between the physical properties and carrier concentration. DFT can predict intrinsic defects and external dopants, guiding experimentalists to regions of phase space and the necessary dopants needed to achieve the desired carrier concentration.[23] In the field of dopability, diamond-like semiconductors (DLS) have received the most computational attention. This includes an amphoteric-defect model,[15,24] phenomenological models for doping limits based on universal band alignment,[12,13,22,25] and detailed analysis of defects in individual DLS systems.[26–28] Despite this success, defect calculations can't be used for high-throughput screening due to their computational cost.

One possible solution is the development of semi-empirical models, as they have proven successful in combining experimental data with physics-based models.[18,21] In the absence of an analytic model or high-throughput defect calculations, statistical learning from experimental or computational data can serve as an alternative to create empirical models and rules of thumb to make predictions of dopability in new compounds. Machine learning has proven successful in understanding and predicting energy and entropy,[29] potentials and forces,[30–32] structure, physical, and elastic properties,[33–38] bandgap,[34,39,40] and defects,[41] as well as enabling high-throughput screening and discovery,[42–46] and guiding experimental synthesis.[47,48]

In order to properly model and interpret dopability, the construction of an empirical dataset for cross-validation is of vital importance. While other physical properties have been tabulated in databases, there are few resources where carrier concentration in semiconductors has been collected. Minimization of the number of uncontrolled variables and maximization of the size of the dataset is helpful in improving accuracy, statistical significance, and applicability to the largest possible group of materials.[40,49] Again, DLS stands out from this perspective as there are a large number of compounds and they are technologically relevant,[50–53] including recent discovery of high β quaternary compounds.[54] DLS compounds have the same tetrahedral local bonding environment and span an impressive fraction of the periodic table.[55,56]

The goal of the following is to establish a broader understanding of the drivers underpinning dopability and develop a method that predicts the possible carrier concentration range in DLS compounds. By performing a careful and extensive literature search with DLS as the model system, the experimentally realized carrier concentration range of 127 compounds have been obtained. Input features for modeling have been generated using structural information, periodic table properties of constituent elements, and widely-available, inexpensive DFT calculation results. We show, using cross-validation, that accurate predictions are possible for this dataset across the entire family, with the model capturing experimental trends in subsets of compounds as well. The features determined to be important in the linear regression are explained and matched with intuition and previous computational results. Finally, the dopability prediction engine is applied to additional DLS compounds which have not been experimentally studied to assess their ultimate potential as TE materials.

RESULTS

Experimental and prediction comparison

The dopability dataset scraped from literature reports of carrier concentration is presented in Fig. 2. The width of each bar represents the dopability range for each of the 127 compounds found in the comprehensive literature search. While the theoretical limits on dopability are determined by defects and chemistry

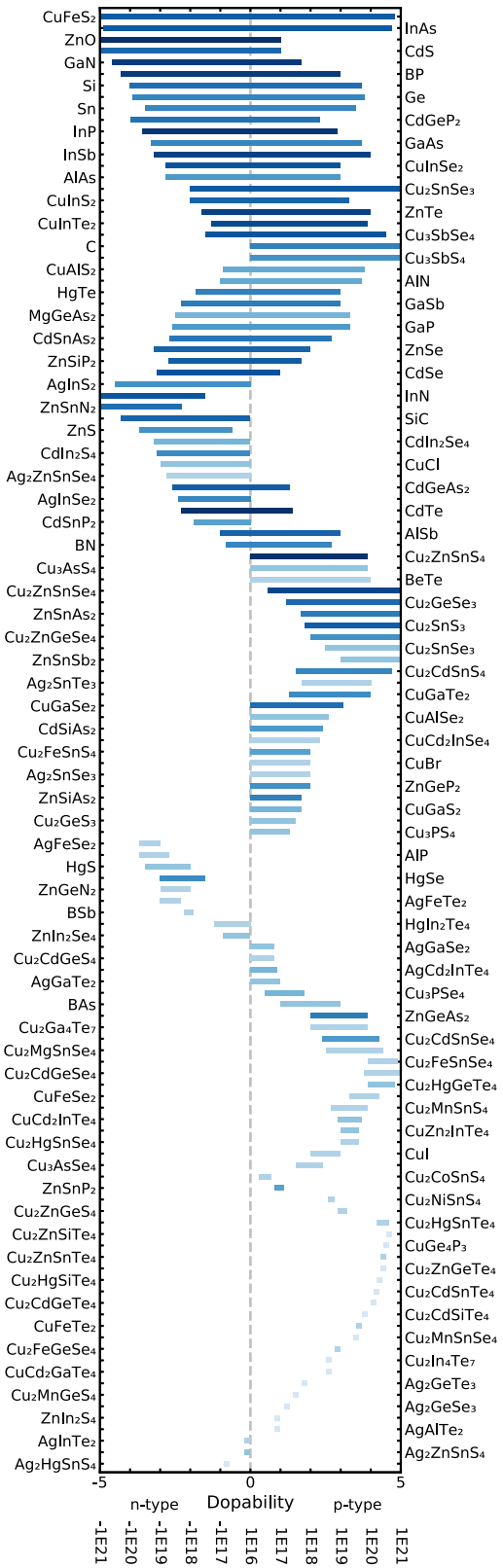

Fig. 2 Experimental dopability range for diamond-like semiconductors collected from literature data. Left end of bar represents highest *n*-type carrier concentration while right side shows highest *p*-type achieved. Top to bottom order chosen to minimize both the difference in dopability range and the left/right displacement of the bar. Compounds with more experimental measurements are darker blue

(see Fig. 1), there are also practical limits due to historical research which we call "persistence". The more times a compound has been reported, the more likely it is that someone has pushed the dopability limit. This is highlighted by the color shading in Fig. 2. The persistence is quite varied, from one report of carrier concentration in a compound to dozens, with the average or median value being approximately five per compound. For compounds which have not been measured with high persistence, for only a single application, or that have not been made with the explicit goal of exploring the the full dopability range, it is likely that only a single distribution (*n*-type, intrinsic, or *p*-type) has been investigated. Whether intentionally or not, compounds which have been studied extensively are more likely to have been sampled in each distribution, thus pushing the dopability limits (Fig. S6). While one may intuit that the low-persistence compounds could be ignored, we provide two virtual experiments (Fig. S7) which demonstrate that we should not ignore these compounds in modeling the dataset and instead use this persistence value in weighing the data for fitting.

Using this set of dopability data, linear regression, random forest, and neural network models were compared based on cross-validated prediction accuracy, with the results shown in Fig. S3. It was found that a linear regression provided similar or better prediction accuracy and superior interpretability, thus it was chosen for further refinement. Feature downselection for the linear model was performed using LASSO (least absolute shrinkage and selection operator), shown in Fig. S4. Due to the effect of persistence in this dataset, sample weighting was also applied to increase prediction accuracy, followed by further feature downselection. Linear regression and other models predict the mean value, while we are interested in the extreme limits of dopability, therefore confidence and prediction intervals were calculated to determine reasonable estimates of these limits.

The resulting predictions using leave-one-out cross-validation (LOOCV) are shown in Fig. 3. This figure contains only the subset of DLS compounds for which both experimental dopability data could be found as well as properties from DFT databases (OQMD and MP) had been calculated, and defect structures were removed. The experimental range is shown with a blue bar representing the maximum extent of dopability in both directions observed in a given compound. A seperate model is used to predict the maximum dopability on each side, with the predicted dopability range being all values between these maxima and given by the red bar. Since this model predicts the mean value for dopability, where we are interested in the maximum extent, a 50% prediction interval is given by the grey error bars. The prediction interval is calculated individually for each type (*n*/*p*) and compound, but in this dataset the bars are largely of the same length, indicating the compounds are fairly evenly distributed across the feature space. As the left and right sides of this dopability range are predicted separately, there is a difference between the accuracy of these individual models. The MAE of the CB is 1.16, while for the VB it is 1.22, giving an average MAE of 1.19 (about one order of magnitude in carrier concentration on average), demonstrating the model is roughly equally predictive of both *n*- and *p*-type carrier concentration. The predictive quality of this model has thus been established using LOOCV demonstrating the ability to predict dopability using this data collection and statistical fitting method.

Careful observation of the experimental dopability dataset reveals some trends, and these trends are captured by the model. A few of these trends are highlighted here and are discussed in more detail in Fig. S5. The first trend that is captured is in the Group IV compounds, where Si, Ge, and Sn all have quite large carrier concentration ranges across *n*- and *p*-type, whereas C can only be made *p*-type and SiC only *n*-type. Second, in binary materials, there is a clear trend in II–VI compounds where dopability shifts from left to right (*n*- towards *p*-type) as you

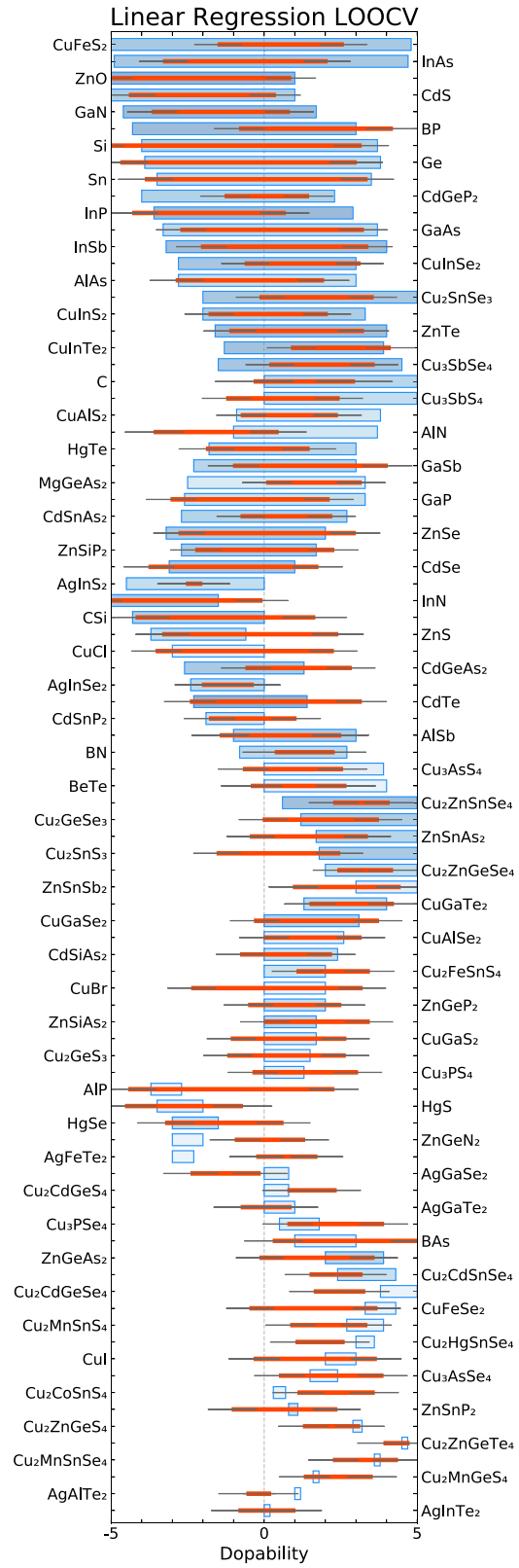

Fig. 3 For each compound, the experimental range is shown in blue and the prediction in red (shade of blue denotes experimental persistence). Grey error bar style lines represent a 50% prediction interval for both the *n*- and *p*-type models. Top to bottom order is the same as in Fig. 1

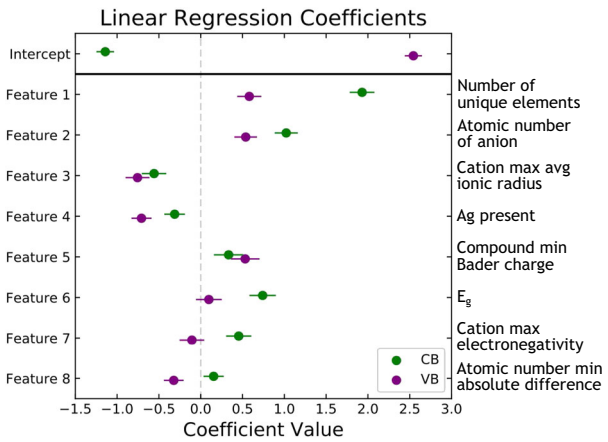

Fig. 4 Linear regression intercepts and coefficients that were most important in determining carrier concentration ranges. Separate models were fit for *n*-type (CB) shown in green and *p*-type (VB) in purple. Error bars represent a 50% confidence interval for the coefficients. Features are in descending order based on the sum of the absolute value of the coefficients

move down the anion group which is not observed in experimental III–V data. Third, compounds of the II–IV–V_2 family show a similar trend in the experimental data for those with Zn but not those with Cd. Finally, in I–III–VI_2 compounds, Cu-containing compounds are much more *p*-type whereas Ag ones are intrinsic or *n*-type. The predictions for all of these sub-families of DLS match qualitatively with the observation of experimental trends and are discussed in more detail in the Supplemental Information.

Model interpretation

Rather than serving only as a prediction engine, the hope is that some physical insight can be gained through interpretation of the resulting model. As this is a linear regression, it has an intercept (baseline *n*/*p*-type dopability) and a number of features that modify the intercept based on the coefficient value associated with each of them. Therefore, Fig. 4 can be interpreted by looking at the sign and magnitude of the coefficient value of each feature in relation to that of the intercept. Features whose coefficients are opposite in sign to their intercept contribute to lowering the carrier concentration range for either the *n*/*p*-type side. The eight features of Fig. 4 constitute the set which provide the best fit (lowest MAE) before overfitting begins by the addition of more features. Each of the features and their associated coefficient values will be discussed individually but viewed together, there are three drivers of dopability in this model: substitutional defects, other chemistry related features (including lattice and electronic energy), and practical limits due to historical persistence.

The first set of coefficients to discuss are those that are partly related to the persistence limitation and more broadly the historical bias of prior experimental work: the intercept, Feature 1, and Feature 4. The sign of the intercept is as we should expect, negative for *n*-type and positive for *p*-type. A large number of *p*-type Cu-containing compounds form the learning set (over 45%), thereby making the intercept for the *p*-type prediction quite large. Conversely, the low individual persistence of these Cu-containing compounds yields a narrow dopability window and thus a smaller *n*-type intercept.

Feature 1 represents the number of unique elements in a compound; both coefficients are positive and the *n*-type value is significantly larger than the *p*-type value. In other words, the range shrinks and becomes more *p*-type with increasing number of unique elements. Unary and binary compounds have been well-studied and thus their experimental dopability range is quite

large, while ternary and quaternary compounds have less reports in the literature and smaller experimental dopability ranges. The impact of a number of unique elements is not a priori obvious, as the number of sites increases, there are more possible sites to dope but also more possible substitutional defects that could pin the Fermi level. Again, the Cu-containing compounds induce this shift due to their native *p*-type behavior and limited persistence.

Feature 4 is whether silver is present in the compound, driving the dopability more *n*-type with a smaller range. Since many of the Ag-containing compounds have low *p*-type carrier concentration (more intrinsic) or are *n*-type, the presence of silver as one of the elements in the composition drives both the CB and VB more negative (Feature 4). As the model is heavily biased by Cu-containing compounds, the presence of Ag thus requires a correction to overcome this difference. Once again, persistence is relevant as the Ag-compounds likewise have limited persistence and a small range.

The next set of features all relate to the cations and their similarity to the anion: Feature 3 is the maximum cation average ionic radius, Feature 7 is the maximum electronegativity of the cations, and Feature 8 is the minimum absolute pairwise difference in atomic number between the anion and each cation. The effect of Features 7 and 8 are quite similar; when there is cation that is more like the anion in electronegativity or atomic number, the carrier concentration is pushed toward intrinsic and the compound has a lower dopability range. This can be seen as the mean VB coefficient for both of these features is negative and the mean CB coefficient is positive, both opposite of the VB/CB intercept. Since the VB coefficient for Feature 3 is negative (opposite VB intercept) and more negative than the CB coefficient, the dopability range decreases and shifts toward *n*-type when there is a cation with a similar ionic radius as the anion. Our interpretation of these trends is that these three features, each relating to having at least one cation that is similar to an anion, whether in size, electronegativity, or atomic radius, reduce the dopability range. Such reduced range may be due to the emergence of low energy "killer" compensating defects.

The other important features are related to chemistry, but not closely linked to the concept of substitutional defects. Feature 2 is the atomic number of the anion; therefore the trend is that as the anionic species is found further to the right and bottom of the periodic table, the compound is more *p*-type and has a smaller dopability range. Feature 5 is the minimum Bader for the compound, and in this case, the minimum Bader charge is always that of the anion. The Bader charge is an approximation of the total electronic charge of an atom and depends on both the chemistry and structure of the compound. The coefficients for the linear model imply that as the anionic element becomes more positive (i.e., less negatively charged), the anion less strongly holds the electrons and the compound becomes more *p*-type and the dopability range increases. This matches our intuition that compounds that are more covalent and less ionic are more dopable. Similarly, Feature 6 is the band gap from OQMD, with larger band gaps leading to a slightly more *p*-type material but one with a much smaller dopability range, again matching our intuition.

Phase competition has been found to affect dopability in materials.[57] The number of elements plays a role in phase competition as there are likely to be more possible compounds in the phase diagram as the number of elements increases; as such, this prior work can be related to Feature 1. Another factor that enters into dopability is the energy of the lattice,[13,15] captured in this model by the Bader charge in Feature 5. It has long been assumed that compounds with larger band gaps are harder to dope, however more recent work has found that the band offset and the position of the band extrema relative to the Fermi stabilization or pinning energy is what controls doping limits.[12,15,22,58] These computational efforts find universal band alignment relate dopability in DLS compounds to the relative position of the VBM and CBM, creating "Pauling-esque" rules. However, these are not linked directly to the elements that are present and the trends in associated properties of those elements and compounds. Therefore, the remaining features are not directly comparable with previous computational work.

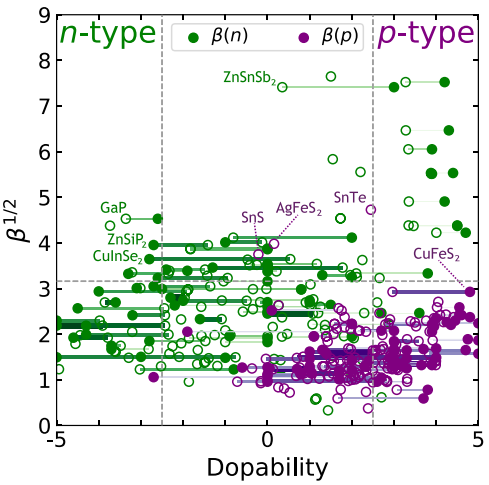

Fig. 5 Predicted quality factor (β) vs. dopability for compounds in the diamond-like structure. Filled circles are experimental dopability, open circles are predicted. Horizontal colored lines connect experimental and predicted values of dopability, with color and thickness of line showing persistence. Dashed gray lines represent approximate benchmarks for good TE performance (carrier concentration $> |3 \times 10^{18}|$ cm^{-3} and $\beta > 10$), meaning promising materials are in the regions labeled *n*-type and *p*-type

High-throughput dopability predictions

Combining dopability predictions with quality factor predictions allows the identification of potential new TE materials in the DLS family. While the optimized version of the dopability model includes features and weighting that do not lend well to high-throughput predictions, a simplified version can be used. A model for dopability was constructed utilizing only a feature set created using the formula and chemistry-based inputs, and LOOCV has a prediction accuracy reasonably similar to the optimized model (MAE/MSE of 1.34/3.05 compared with 1.19/2.23). This model was then used in tandem with β[20] to generate predictions for DLS compounds, with results shown in Fig. 5. Out of a total of 188 DLS compounds considered in this study, the dopabiilty learning set was based on 127 compounds. This yields 61 compounds that had no experimental dopability information; both β and dopability could then be predicted for these compounds. An additional 67 compounds were part of the dopability learning set and had β predictions but had a persistence of less than five papers. In this latter case, the dopability prediction may highlight opportunities for extending the known dopant range in promising β compounds.

The first notable observation is that both the model predictions and experimental values (where they exist) indicate there are far more DLS compounds that are dopable to an appropriate *p*-type carrier concentration range to be useful for thermoelectric applications. Furthermore, there are far more DLS compounds with β predicted to be 10 or greater as *n*-type than *p*-type. The average hole mobility is approximately two orders of magnitude lower than the electron mobility, and the valence band mass is about one order of magnitude higher than the conduction band on average. This leads to many materials with high-potential *n*-type performance and fewer with high *p*-type β. Unfortunately, it appears that far fewer compounds can be realized *n*-type to a

reasonably high-carrier concentrations. Many of these are quaternary Cu-containing compounds which have been made *p*-type only, and while the quality factor indicates they have the potential to be very good thermoelectric materials, the dopability predictions show it is very unlikely they can be made *n*-type to realize that potential. However, the high quantity of low-persistence quaternary Cu compounds may bias this conclusion; further studies of extrinsic doping in these materials is needed to establish the attainable carrier concentration bounds.

Figure 5 also demonstrates there are a number of binary and ternary compounds which could be interesting thermelectric materials where the dopability limits have not been fully explored experimentally. For example, β calculations indicate GaP could be a good thermoelectric, with both experimental and predicted dopability in the approriate range. $ZnSnSb_2$ is predicted to have a much larger carrier concentration window than that found to date experimentally, and it could be a good thermoelectric if pushed from *p*-type through intrinsic to *n*-type. This also indicates that $CuInSe_2$, which has had recent attention as a *p*-type thermoelectric, should be looked at instead as an *n*-type TE. Likewise *p*-type $CuFeS_2$ and $AgFeS_2$ could be promising TE materials. SnTe[59–61] and SnS[62,63] have both received significant interest as TE materials in their typical structures (rocksalt and orthorhombic, respectively) and our predictions indicate they could exhibit high performance if stabilized in the diamond-like structure.

DISCUSSION

In this work, we show that dopability ranges can be predicted to approximately one order of magnitude against experimental results through linear modeling. The accuracy of a simple linear model is found to be similar to more complex machine learning techniques and allows greater interpretability regarding the features and properties that control dopability in diamond-like semiconductors. A number of features indicate that substitutional defects play a major role in driving carrier concentration ranges toward intrinsic ranges. In addition, features such as band gap and Bader charge match either "rules of thumb" or previous computational findings. Compounds that possess both promising thermoelectric quality factor (β) and complementary dopability are identified. Our results serve as a caution against pursuing a subset of high β compounds that have unfavorable dopability ranges. For compounds or subgroups of DLS where persistence is historically low, this work serves to inspire further experimental interrogation of carrier concentration limits when the predicted dopability range is far greater than the experimentally realized limits. These results also indicate where detailed defect calculation efforts are most impactful. We note that the predictive dopability model is not inherently limited to DLS compounds; an expanded learning set could incorporate structural descriptors so that the model could be applied to a more diverse set of compounds, including all thermoelectric materials.

METHODS

An extensive literature search was undertaken to compile a dataset that could be used for statistical modeling. Although a few sources of tabulated carrier concentration exist, these often do not note sample quality or processing conditions, even though these are very important in determining carrier concentration in semiconductors. For this reason, careful consideration was given to experimental conditions from which measurements could be relied upon and all measurements for this dataset were scraped from original sources where possible. An attempt was made to use bulk samples produced at or near equilibrium conditions, measured using the Hall effect technique near room temperature and pressure. Nanomaterials, thin films, and other non-equilibrium processing methods were avoided where possible. The results of this literature search are given in Table S1.

With DLS as the model system, the scale for dopability must be defined such that the data are distributed fairly normal so that it can be modeled. The primary goal is to describe the full extent of the achievable dopability range in a compound, from the maximum *n*-type carrier concentration to the maximum *p*-type concentration. In some compounds, carrier concentration can be experimentally varied across many orders of magnitude for both hole and electron majority carriers, while for others the range is narrow or limited to a single type. Although carrier concentration spans many orders of magnitude for both positive and negative values, we are primarily interested in the range from intrinsic to degenerately doped. Thus, our scale ranges from -1×10^{21} (degenerate *n*-type) to 1×10^{21} (degenerate *p*-type), with intrinsic considered any concentration less than $|1\times10^{16}|$ (all units given as cm^{-3}). This allows us to define a linear scale from −5 to 0 to 5, corresponding to -1×10^{21} to $\pm1\times10^{16}$ to 1×10^{21} where every integer value represents an order of magnitude in carrier concentration. Any concentration greater than 1×10^{21} is assigned a value of five and anything less than 1×10^{16} is considered to be 0. Using this scale for the carrier concentration data scraped from the literature results in a distribution that is relatively normal, as seen if Fig. S1.

To model this dataset, a list of features was first compiled. This includes chemistry-based features from periodic table properties of the consitutuent elements. Added to this feature set were inexpensive calculations from the Open Quantum Materials Database (OQMD)[64,65] and Materials Project (MP).[66–68] Lastly, structure and other miscellaneous features were added (Fig. S2). Modeling was performed using a number of Python packages, most notably Scikit-learn and StatsModels. Linear regression and leave-one-out cross-validation (LOOCV) were chosen due to a combination of prediciton accuracy and interpretability after comparing the results with other machine learning methods. A detailed discussion of data and feature set preparation, model comparison and analysis, and refinement can be found in the Supplemental Information. The primary metric for scoring accuracy used here is the mean absolute error (MAE) which is the average distance between the experimental and predicted value, though mean squared error (MSE) is also evaluated, where smaller values indicate less difference between experiments and predictions.

DATA AVAILABILITY

The dataset generated and analyzed during this study is made available in the Supplemental Information.

ACKNOWLEDGEMENTS

Special thanks to Vinay I. Hegde for assistance with collecting data from the Open Quantum Materials Database. We acknowledge support from National Science Foundation DMR program, grant no. 1333335, 1334713, 1729487, and 1729594.

AUTHOR CONTRIBUTIONS

S.A.M. compiled dataset, wrote code, analyzed data, and wrote the paper. M.D. and K. G. assisted in writing code and analysis. S.A. carried out and gathered first principles calculations. G.J.S. and E.S.T. contributed to writing of the paper. All authors discussed reults and commented on the paper.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-018-0123-6).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Chan, M., Lai, S., Fung, M., Lee, C. & Lee, S. Doping-induced efficiency enhancement in organic photovoltaic devices. *Appl. Phys. Lett.* **90**, 023504 (2007).
2. Oh, J., Yuan, H.-C. & Branz, H. M. An 18.2%-efficient black-silicon solar cell achieved through control of carrier recombination in nanostructures. *Nat. Nanotechnol.* **7**, 743 (2012).

3. Avouris, P. & Chen, J. Nanotube electronics and optoelectronics. *Mater. Today* **9**, 46–54 (2006).
4. Khokhlev, O. V., Bulashevich, K. A. & Karpov, S. Y. Polarization doping for III-nitride optoelectronics. *Phys. Status Solidi (A)* **210**, 1369–1376 (2013).
5. Pankratov, E. & Bulaeva, E. Doping of materials during manufacture p-n-junctions and bipolar transistors. Analytical approaches to model technological approaches and ways of optimization of distributions of dopants. *Rev. Theor. Sc.* **1**, 58–82 (2013).
6. Derycke, V., Martel, R., Appenzeller, J. & Avouris, P. Controlling doping and carrier injection in carbon nanotube transistors. *Appl. Phys. Lett.* **80**, 2773–2775 (2002).
7. Kim, G.-H., Shao, L., Zhang, K. & Pipe, K. P. Engineered doping of organic semiconductors for enhanced thermoelectric efficiency. *Nat. Mater.* **12**, 719 (2013).
8. Snyder, G. J. & Toberer, E. S. Complex thermoelectric materials. *Nat. Mater.* **7**, 105 (2008).
9. Toberer, E. S., May, A. F. & Snyder, G. J. Zintl chemistry for designing high efficiency thermoelectric materials. *Chem. Mater.* **22**, 624–634 (2009).
10. Pei, Y. et al. Stabilizing the optimal carrier concentration for high thermoelectric efficiency. *Adv. Mater.* **23**, 5674–5678 (2011).
11. Rocket, A. *Defects in Semiconductors*. (Springer US, Boston, MA, 2008; 289–356.
12. Zhang, S., Wei, S.-H. & Zunger, A. A phenomenological model for systematization and prediction of doping limits in II–VI and I-III-VI2 compounds. *J. Appl. Phys.* **83**, 3192–3196 (1998).
13. Zunger, A. Practical doping principles. *Appl. Phys. Lett.* **83**, 57–59 (2003).
14. Paudel, T. R., Zakutayev, A., Lany, S., d'Avezac, M. & Zunger, A. Doping rules and doping prototypes in A2BO4 spinel oxides. *Adv. Funct. Mater.* **21**, 4493–4501 (2011).
15. Walukiewicz, W. Intrinsic limitations to the doping of wide-gap semiconductors. *Phys. B: Condens. Matter* **302**, 123–134 (2001).
16. Faschinger, W., Ferreira, S. & Sitter, H. Doping limitations in wide gap II-VI compounds by Fermi level pinning. *J. Cryst. Growth* **151**, 267–272 (1995).
17. Gaultois, M. W. et al. A recommendation engine for suggesting unexpected thermoelectric chemistries. *arXiv preprint arXiv:1502.07635* (2015).
18. Miller, S. A. et al. Capturing anharmonicity in a lattice thermal conductivity model for high-throughput predictions. *Chem. Mater.* **29**, 2494–2501 (2017).
19. Toher, C. et al. High-throughput computational screening of thermal conductivity, Debye temperature, and Grüneisen parameter using a quasiharmonic Debye model. *Phys. Rev. B* **90**, 174107 (2014).
20. Gorai, P. et al. TE design lab: a virtual laboratory for thermoelectric material design. *Comput. Mater. Sci.* **112**, 368–376 (2016).
21. Yan, J. et al. Material descriptors for predicting thermoelectric performance. *Energy Environ. Sci.* **8**, 983–994 (2015).
22. Zhang, S., Wei, S.-H. & Zunger, A. Overcoming doping bottlenecks in semiconductors and wide-gap materials. *Phys. B: Condens. Matter* **273**, 976–980 (1999).
23. Ohno, S. et al. Phase boundary mapping to obtain n-type mg 3 sb 2-based thermoelectrics. *Joule* 2(1), 141-154, (2017) https://doi.org/10.1016/j.joule.2017.11.005.
24. Walukiewicz, W. Mechanism of schottky barrier formation: the role of amphoteric native defects. *J. Vac. Sci. & Technol. B: Microelectron. Process. Phenom.* **5**, 1062–1067 (1987).
25. Van de Walle, C. G. & Neugebauer, J. Universal alignment of hydrogen levels in semiconductors, insulators and solutions. *Nature* **423**, 626 (2003).
26. Van de Walle, C. G. & Neugebauer, J. First-principles calculations for defects and impurities: Applications to III-nitrides. *J. Appl. Phys.* **95**, 3851–3879 (2004).
27. Van de Walle, C. G. Hydrogen as a cause of doping in zinc oxide. *Phys. Rev. Lett.* **85**, 1012 (2000).
28. Lany, S., Osorio-Guillén, J. & Zunger, A. Origins of the doping asymmetry in oxides: Hole doping in NiO versus electron doping in ZnO. *Phys. Rev. B* **75**, 241203 (2007).
29. Legrain, F., Carrete, J., van Roekeghem, A., Curtarolo, S. & Mingo, N. How chemical composition alone can predict vibrational free energies and entropies of solids. *Chem. Mater.* **29**, 6220–6227 (2017).
30. Chen, C. et al. Accurate force field for molybdenum by machine learning large materials data. *Phys. Rev. Mater.* **1**, 043603 (2017).
31. Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
32. Takahashi, A., Seko, A. & Tanaka, I. Conceptual and practical bases for the high accuracy of machine learning interatomic potentials: application to elemental titanium. *Phys. Rev. Mater.* **1**, 063801 (2017).
33. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 144110 (2017).
34. Zhang, Y. & Ling, C. A strategy to apply machine learning to small datasets in materials science. *npj Comput. Mater.* **4**, 25 (2018).
35. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
36. De Jong, M. et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Sci. Rep.* **6**, 34256 (2016).
37. Mannodi-Kanakkithodi, A., Huan, T. D. & Ramprasad, R. Mining materials design rules from data: the example of polymer dielectrics. *Chem. Mater.* **29**, 9001–9010 (2017).
38. Ma, X., Li, Z., Achenie, L. E. & Xin, H. Machine-learning-augmented chemisorption model for CO_2 electroreduction catalyst screening. *J. Phys. Chem. Lett.* **6**, 3528–3533 (2015).
39. Zeng, Y., Chua, S. J. & Wu, P. On the prediction of ternary semiconductor properties by artificial intelligence methods. *Chem. Mater.* **14**, 2989–2998 (2002).
40. Dey, P. et al. Informatics-aided bandgap engineering for solar materials. *Comput. Mater. Sci.* **83**, 185–195 (2014).
41. Medasani, B. et al. Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning. *npj Comput. Mater.* **2**, 1 (2016).
42. Li, Z., Wang, S., Chin, W. S., Achenie, L. E. & Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* **5**, 24131–24138 (2017).
43. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73 (2016).
44. Sendek, A. D. et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306–320 (2017).
45. Oliynyk, A. O. et al. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
46. Lu, W., Xiao, R., Yang, J., Li, H. & Zhang, W. Data mining-aided materials discovery and optimization. *J. Mater.* **3**, 191–201 (2017).
47. Kim, E. et al. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).
48. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* **3**, 53 (2017).
49. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
50. Guo, Q. et al. Fabrication of 7.2% efficient CZTSSe solar cells using CZTS nanocrystals. *J. Am. Chem. Soc.* **132**, 17384–17386 (2010).
51. Panthani, M. G. et al. Synthesis of CuInS2, CuInSe2, and Cu (in x Ga1−x) Se2 (CIGS) nanocrystal "inks" for printable photovoltaics. *J. Am. Chem. Soc.* **130**, 16770–16777 (2008).
52. Fan, J. et al. Investigation of thermoelectric properties of Cu2GaxSn1−xSe3 diamond-like compounds by hot pressing and spark plasma sintering. *Acta Mater.* **61**, 4297–4304 (2013).
53. Liu, R. et al. Ternary compound CuInTe 2: a promising thermoelectric material with diamond-like structure. *Chem. Commun.* **48**, 3818–3820 (2012).
54. Ortiz, B. R. Ultralow thermal conductivity in diamond-like semiconductors: selective scattering of phonons from antisite defects. *Chem. Mater* **30**(10), 3395–3409 (2018).
55. Shay, J. L. & Wernick, J. H. *Ternary Chalcopyrite Semiconductors: Growth, Electronic Properties, and Applications: International Series of Monographs in The Science of The Solid State*, vol. 7 (Elsevier, 2017) https://www.elsevier.com/books/ternary-chalcopyrite-semiconductors-growth-electronic-properties-andapplications/shay/978-0-08-017883-7
56. Parthé, E. *Crystal Chemistry of Tetrahedral Structures* (CRC Press, 1964) http://www.worldcat.org/title/crystal-chemistry-of-tetrahedral-structures/oclc/1572050.
57. Walukiewicz, W. Application of the amphoteric native defect model to diffusion and activation of shallow impurities in III–V semiconductors. *Mater. Res. Soc. Symp. Proc.* **300**(421), 300–421 (1993).
58. Longini, R. & Greene, R. Ionization interaction between impurities in semiconductors and insulators. *Phys. Rev.* **102**, 992 (1956).
59. Li, W. et al. Advances in environment-friendly SnTe thermoelectrics. *ACS Energy Lett.* **2**, 2349–2355 (2017).
60. Zhou, M. et al. Optimization of thermoelectric efficiency in SnTe: the case for the light band. *Phys. Chem. Chem. Phys.* **16**, 20741–20748 (2014).
61. Moshwan, R., Yang, L., Zou, J. & Chen, Z.-G. Eco-friendly SnTe thermoelectric materials: Progress and future challenges. *Adv. Funct. Mater.* **27**, 1703278 (2017).
62. Wei, T.-R. et al. Thermoelectric SnS and SnS-SnSe solid solutions prepared by mechanical alloying and spark plasma sintering: Anisotropic thermoelectric properties. *Sci. Rep.* **7**, 43262 (2017).
63. Guo, R., Wang, X., Kuang, Y. & Huang, B. First-principles study of anisotropic thermoelectric transport properties of IV–VI semiconductor compounds SnSe and SnS. *Phys. Rev. B* **92**, 115202 (2015).

64. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the the Open Quantum Materials Database (OQMD). *Jom* **65**, 1501–1509 (2013).
65. Kirklin, S. et al. The open quantum materialsdatabase (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
66. Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
67. de Jong, M. et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**:150009 https://doi.org/10.1038/sdata.2015.9 (2015).
68. de Jong, M., Chen, W., Geerlings, H., Asta, M. & Persson, K. A. A database to enable discovery and design of piezoelectric materials. *Sci. Data* **2**:150053 https://doi.org/10.1038/sdata.2015.53 (2015).