Web-based machine learning models for real-time screening of thermoelectric materials properties

Michael W. Gaultois,[1, a)] Anton O. Oliynyk,[2] Arthur Mar,[2] Taylor D. Sparks,[3] Gregory J. Mulholland,[4] and Bryce Meredig[4, b)]

[1)] *Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, United Kingdom*

[2)] *Department of Chemistry, University of Alberta, Edmonton, Alberta, T6G 2G2, Canada*

[3)] *Department of Materials Science and Engineering, University of Utah, Salt Lake City, Utah, 84112, USA*

[4)] *Citrine Informatics, Redwood City, California, 94063, USA*

(Dated: 2 May 2016)

The experimental search for new thermoelectric materials remains largely confined to a limited set of successful chemical and structural families, such as chalcogenides, skutterudites, and Zintl phases.[1–3] In principle, computational tools such as density functional theory (DFT) offer the possibility of rationally guiding experimental synthesis efforts toward very different chemistries. However, in practice, predicting thermoelectric properties from first principles remains a challenging endeavor,[4] and experimental researchers generally do not directly use computation to drive their own synthesis efforts. To bridge this practical gap between experimental needs and computational tools, we report an open machine learning-based recommendation engine (http://thermoelectrics.citrination.com) for materials researchers that suggests promising new thermoelectric compositions based on pre-screening about 25,000 known materials, and also evaluates the feasibility of user-designed compounds. We show this engine can identify interesting chemistries very different from known thermoelectrics. Specifically, we describe the experimental characterization of one example set of compounds derived from our engine, $RE_{12}Co_5Bi$ (RE = Gd, Er), which exhibits surprising thermoelectric performance given its unprecedentedly high loading with metallic d and f block elements, and warrants further investigation as a new thermoelectric material platform. We show our engine predicts this family of materials to have low thermal and high electrical conductivities, but modest Seebeck coefficient, all of which are confirmed experimentally. We note that the engine also predicts materials that may simultaneously optimize all three properties entering into zT; we selected $RE_{12}Co_5Bi$ for this study due to its interesting chemical composition and known facile synthesis.

Keywords: Materials discovery, thermoelectric materials, rapid screening, data mining, machine learning

I. INTRODUCTION

For any materials problem, breaking out of "local optima" in composition space to discover entirely new chemistries remains a notoriously difficult challenge.[5] Many of the most notable materials classes under investigation today–from Na_xCoO_2 derived thermoelectrics[6] to iron arsenide superconductors[7]–were discovered fortuitously. As a result, experimental efforts often gravitate toward incrementally improving *known* chemistries (via doping, nanostructuring, etc.), as these efforts are more likely to bear fruit than high-risk searches through chemical whitespace for entirely new materials.

The consequence of research communities' focus on further exploitation of known chemistries rather than exploration of unknown chemistries is that much of composition space simply remains uncharacterized. We illustrate the remarkable chemical homogeneity of most thermoelectric materials investigated to date by plotting each material from the thermoelectric database of Gaultois *et al.*[8] on the periodic table based on the composition-weighted average of the positions of elements in the material (Fig. 1). The tight cluster of previously investigated chemistries is, as expected, dominated by chalcogenides and *p*-block elements such as Sn and Sb. In contrast, we also show the positions of $Gd_{12}Co_5Bi$ and $Er_{12}Co_5Bi$, materials derived from our recommendation engine, which we characterize as a new class of thermoelectrics in this work. These materials are almost pure intermetallics, in sharp contrast to thermoelectric compounds investigated to date (Fig. 2). The objective of our recommendation engine is to directly enable experimental researchers to rapidly identify new materials, such as $RE_{12}Co_5Bi$, that are very distinct from known compound classes, and worthy of further study.

A materials recommendation engine

Our recommendation engine is a machine learning-based approach[9,10] for efficiently driving synthetic efforts toward promising new chemistries. We have trained a machine learning model to make a confidence level prediction of whether the (1) Seebeck coefficient, (2) electrical resistivity, (3) thermal conductivity, and (4) band gap of input materials are within acceptable ranges for thermoelectric applications. We define these ranges

[a)] Electronic mail: mg757@cam.ac.uk

[b)] Electronic mail: bryce@citrine.io

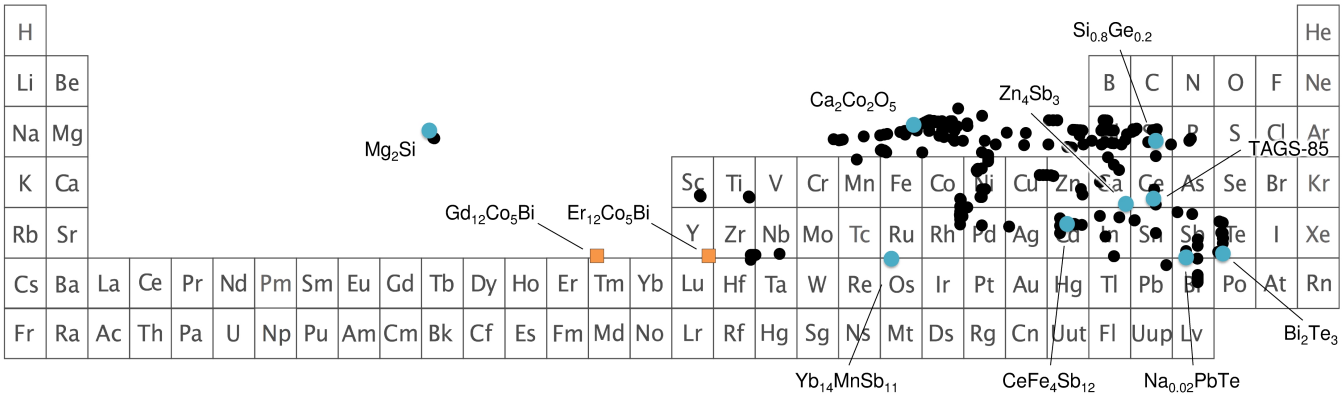

FIG. 1. Most known thermoelectric materials lie in a tight cluster in composition space (black and blue dots; blue dots have chemical formulae explicitly labelled). The recommendation engine presented here allows the identification of new thermoelectric materials families that are well outside the existing composition space of common systems in the Gaultois *et al.* database.[8] In particular, we report the characterization of $RE_{12}Co_5Bi$ (RE = Gd, Er; orange squares), which are chemically and structurally distinct from known thermoelectrics.

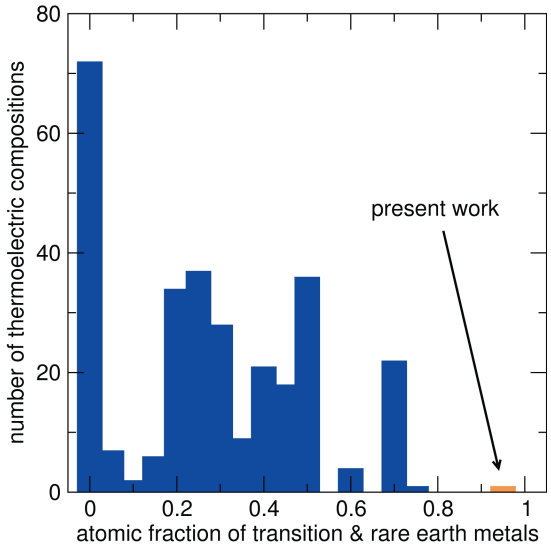

FIG. 2. The strongly intermetallic $RE_{12}Co_5Bi$ compounds we report here lie far outside the norm for metal loading among collected thermoelectric compositions in the Gaultois *et al.* database.[8] The recommendation of these materials was neither the result of simple interpolation between known compounds nor obvious from a strict chemical intuition standpoint.

as follows: (1) $|S| > 100\,\mu\mathrm{V\,K^{-1}}$; (2) $\rho < 10^{-2}\,\Omega\,\mathrm{cm}$; (3) $\kappa < 10\,\mathrm{W\,m^{-1}\,K^{-1}}$; and (4) $E_g > 0\,\mathrm{eV}$, all at room temperature.

For each range of thermoelectric property, the engine gives a confidence score between 0% and 100% that a given material's measured value for that property at room temperature will fall within the targeted range. We would classify any material for which the answer to all these questions is likely "yes" as a potentially promising thermoelectric that may warrant further study. The purpose of our recommendation engine is thus neither to make *quantitative* predictions of these thermoelectric properties, nor to definitively identify record-setting compounds–these remain open challenges for future work. Rather, the engine is intended to greatly augment the chemical intuition of experimental researchers working on materials discovery. In particular, we have found that our model's ability to screen vast numbers of possible compositions and short-list interesting candidates can inspire materials syntheses that would not have been obvious *a priori*.

Machine learning models such as those developed here differ considerably from atomistic simulation approaches such as density functional theory (DFT). DFT is already a well-established tool for accelerating materials discovery,[11,12] and high-throughput methods have already been applied successfully in the search for new thermoelectric materials.[13–17] Nevertheless, accurately predicting thermoelectric properties from first principles remains challenging.[4] Recent works, for example, use the BoltzTraP code[18] to estimate the Boltzmann transport properties of candidate materials based on DFT-predicted band structures.[19] The nascent field of materials informatics–algorithmically extracting new knowledge by mining large-scale materials databases–has emerged alongside these traditional physics-based simulations as a key means of predicting materials behavior.[20,21]

The present machine learning-based recommendation engine looks for empirical, chemically meaningful patterns in *experimentally reported* data on known thermoelectric compounds to make statistical predictions for the performance of new materials. Further, while efforts such as the Materials Project are making the results of DFT calculations more accessible to the experimental materials community than ever before,[5] most experimentalists still are not able to run DFT calculations continually to inform their laboratory work in real-time.

To make predictive computation more widely accessible, we make the results of the present work available as a web application (http://thermoelectrics.citrination.com) that any materials researcher can utilize to request real-time predictions and search for new thermoelectric candidates.

II. METHODS

Modelling and informatics

Here we describe the approach used to construct the recommendation engine. Our engine is an example of materials informatics,[22,23] or the application of empirical machine learning methods to the prediction of materials behaviour. Any machine learning approach for materials relies on three key ingredients: training data, descriptors, and choice of algorithm. Training data are the example sets from which the machine learning approach should extract meaningful chemical trends. Descriptors are the low-level characteristics of materials (*e.g.*, crystal structure, chemical formula, etc.) that might correlate with materials properties of interest. Specifically, descriptors are either numerical (*e.g.*, average atomic number Z) or categorical (*e.g.*, crystal structure = perovskite) variables that enable us to "vectorize" materials in such a way that they become amenable to machine learning techniques. Finally, learning algorithms interrogate descriptor-vectorized training data for relevant patterns.

In this work, the training set comprises a large body of both experimental thermoelectric characterization data,[8] experimental materials property data from the NIMS MatNavi database, and first principles-derived electronic structure data.[5,24] These data are publicly available via the Citrination platform (http://www.citrination.com), the Materials Project API (http://www.materialsproject.org/open), and NIMS (http://mits.nims.go.jp/index_en.html). These data consist of the Seebeck coefficients, thermal conductivities, electrical conductivities, and band gaps measured for thousands of materials as a function of temperature and a variety of other metadata conditions. Our model uses these input data to learn interesting chemical trends that could be exploited to design new materials. As large, high-quality training data sets are scarce in materials science relative to the biological sciences, where bioinformatics has become a standard tool, we urge the materials community to consider contributing to data infrastructures (Citrination, Materials Project, NIST's DSpace repository, EU's NoMaD, and others) that together will significantly expand open access to data for materials researchers.

Descriptors are the second key ingredient in materials informatics. The scientific literature around designing descriptors for materials has grown substantially in just the past several years.[25,26] Indeed, recent work has shown that the predictive power of machine learning models for materials is strongly dependent upon the selected descriptor set.[27] Our engine relies upon a tuned blend of descriptors designed in-house and drawn from a variety of sources.[4,9] By way of example, as materials scientists, we recognize that the periodic table contains a tremendous amount of information about how the elements behave and interact. We thus pre-bias our machine learning models with such knowledge (*e.g.*, the d block of the periodic table is metallic; Li and Na are chemically very similar but not identical; and the lanthanides behave similarly in ionic compounds). This step allows us to create predictive models with data sets that have thousands (rather than tens or hundreds of thousands) of examples.

The ability of materials informatics techniques to extract signal from materials data is strongly dependant on effective descriptor design and access to large quantities of training data. With respect to the latter point, machine learning algorithms are only able to identify patterns that are (at least sparsely) sampled by the training data. An important manifestation of this requirement in the context of the present work is modeling doping. Doping represents a minute change in materials composition (on an atomic percentage basis), but may result in orders of magnitude changes in properties. As most of the training data used in this work correspond to *undoped* bulk compounds, we expect the recommendation engine to perform best in identifying new such bulk systems which could be potentially be further optimized via doping. Given more training data, we could readily extend the current work to dilutely doped thermoelectric systems.

Finally, our recommendation engine is built using the so-called random forest algorithm.[28] This algorithm constructs a large number of decision trees, all trained on slightly different subsets of the training data. Random forest is an ensembling technique, which takes advantage of the fact that a collection of "weak" learners such as decision trees can, in concert, model extraordinarily complex nonlinear behaviour. An example rule that a single decision tree might learn is that if a material contains two elements with very different electronegativities (*e.g.*, Na and Cl), that material is likely to have a large band gap. Of course, the thermoelectric phenomena we seek to model here are substantially more subtle, and thus a large random forest of decision trees is useful in untangling the underlying physics. We refer the reader elsewhere[4,9,29] for more detailed discussions and tutorials on how to apply random forests to materials data.

Model validation

We visualize the accuracy of our recommendation engine's predictions in Fig. 3, which represents the results of leave-one-out cross-validation (LOOCV) on our training data (in the case of the band gap data, we performed LOOCV on a subset of the extremely large training set).

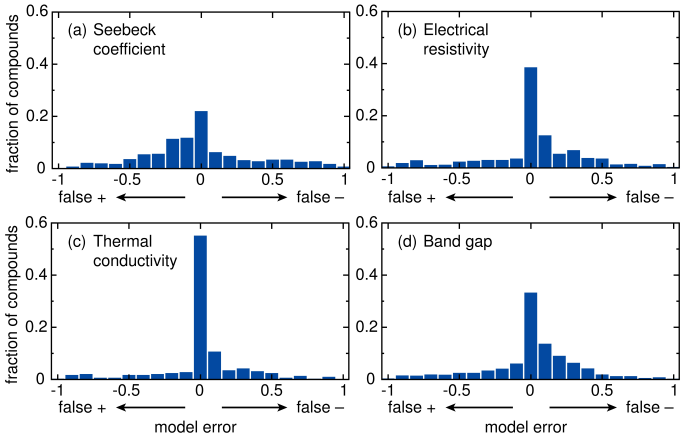

FIG. 3. Leave-one-out cross validation error histograms for the four key properties estimated by our recommendation engine: (a) Seebeck coefficient; (b) electrical resistivity; (c) thermal conductivity; and (d) band gap. For each material in our training set and each property, the recommendation engine gives a confidence score between 0 and 1 that the property value falls within the ideal windows we have defined for thermoelectric applications. Errors approaching $+1$ represent false negatives (our engine was extremely confident the material would be poor for that property, but the property is actually good); and an error of -1 is a false positive (our engine was extremely confident the material would be good for that property, but the property is actually poor). The peak around 0 for each property shows that the engine generally gives confidence values very close to unity for materials possessing properties in the desired ranges, or close to zero for materials whose property values fall outside the target range.

In the LOOCV procedure, if we have n total measurements of a particular property such as thermal conductivity, we train our machine learning model on $n-1$ of these values and predict the nth (left out) value. We perform one training step and prediction for each property value, and present the error distribution for all n values in Fig. 3. The error distribution then provides us with a sense of how we may expect the model to perform on new materials of which we have no prior knowledge.

Fig. 3 indicates that our engine generally makes very reliable assessments of thermoelectric materials properties. The modes of the error distributions are in each case close to 0. For each property, the engine's errors skew toward false negatives (resistivity, band gap, thermal conductivity) or false positives (Seebeck), which reflects the fact that the underlying training data do not contain equal fractions of positive and negative examples. Seebeck coefficients prove most difficult to assess (*i.e.*, the error distribution for that property has the largest standard deviation), likely because there are strikingly different mechanisms that underpin the values, for example, strongly correlated oxides as opposed to degenerate semiconductors. Owing to the difficulty in assessing the Seebeck coefficient, initial predictive models using only the electrical resistivity, thermal conductivity, and Seebeck coefficient produced too many candidates that were good metals with poor Seebeck coefficients. To remedy this shortcoming and provide more robust recommendations, the band gap was added as a secondary metric, where we determine the probability whether a given composition will have a non-zero bandgap.

Experimental details

$RE_{12}Co_5Bi$ (RE = Gd, Er) samples were made by arc-melting freshly filed Er or Gd pieces (99.9%, Hefa), Co powder (99.8%, Cerac), and Bi powder (99.999%, Alfa Aesar). Stoichiometric mixtures (0.5 g total mass) with 5% to 7% excess Bi were pressed into pellets and melted twice in arc-melting furnace under argon atmosphere (Edmund Bühler Compact Arc Melter MAM-1). The total mass loss after melting was $<1\%$. The samples were sealed in silica tubes and annealed at 1070 K for one week, then quenched in cold water. To produce enough material for physical property measurement, $\sim$70 samples of each compound were prepared, and pure samples were combined by melting into a single ingot of $\sim$5 g, which was sanded to yield the appropriate geometry (either a rectangular bar, or a cylinder). Density was measured using Archimedes' method; the final pellets had densities 100% of the single crystal values ($\rho_{Gd_{12}Co_5Bi} = 8.6\,g/cm^3$, $\rho_{Er_{12}Co_5Bi} = 9.9\,g/cm^3$).

Powder X-ray diffraction patterns were collected using an INEL CPS 120 diffractometer with Cu Kα_1 radiation at room temperature, and Rietveld refinement was used to confirm the structure and phase purity (see Supporting Information).[30] Backscatter electron microscopy and elemental analysis via energy dispersive X-ray spectroscopy (EDX) were performed with a JEOL JSM-6010LA InTouchScope scanning electron microscope. Backscatter micrographs reveal the samples are largely compositionally homogeneous (see Supporting Information).[30] Quantitative elemental analysis on several polished pieces found an atomic composition of Gd69(2)Co26(2)Bi5(2) which is in a good agreement with expected $RE_{12}Co_5Bi$ composition. $Er_{12}Co_5Bi$ samples were not appropriate for quantitative analysis because of overlapping Co Kα (6.924 keV) and Er Lα (6.947 keV) lines.

High-temperature thermoelectric properties (electrical resistivity and Seebeck coefficient) were measured with an ULVAC Technologies ZEM-3. Sample bars had approximate dimensions of 9 mm$\times$4 mm$\times$4 mm. Measurements were performed with a helium under-pressure, and data was collected from 300 K to 800 K through three heating and cooling cycles over 18 hours to ensure sample stability and reproducibility.

III. DISCUSSION

In this work, we are interested not only in developing a model that gives accurate predictions of materials properties, but also in making it immediately accessible and useful for experimental researchers. To that end, we have published our recommendation engine as a web app at http://thermoelectrics.citrination.com, where researchers may explore a pre-computed list of around 25 000 known compounds (representing a sizable subset of the Inorganic Crystal Structure Database, or ICSD), and also use our model to evaluate their own materials candidates in real-time. In this way, we hope that the app serves as a rapid triage tool for ideas for potential new thermoelectric materials.

This adds to a growing toolbox of computational tools designed to be a user-friendly aid to experimental workers, such as TEDesignLab, and the Materials Project.[5,31] Our pre-computed list may be arranged according to the probabilities associated with any one of the four properties we are modelling, and is sorted by default according a composite score that takes all four properties into account. Furthermore, the user may specify cutoff thresholds for any of the properties, and thereby greatly reduce the size of the list.

As we believe our extensive precomputed list contains some interesting and heretofore uncharacterized candidate thermoelectric materials, we now comment on a select set of high-ranking compounds. Several of these compounds are given in Table I.

$TaVO_5$ and $TaPO_5$ occur in an analogous crystal structure to the phosphate tungsten bronzes.[32,33] These materials can be expected to have good thermoelectric performance given the heavy atoms, the potential for low electrical resistivity provided by the repeating ReO_3-type structural network that is highly connected in three dimensions, and the intrinsic crystallographic shear provided by the crystal structure. Although the phosphate tungsten bronzes themselves are not highly rated, their metallic electrical transport properties are encouraging for structural analogues.[34] Moreover, $TaVO_5$ has a negative coefficient of thermal expansion and a structural transition at $600°$C.[35] This structural transition may lead to softening of phonon modes and anharmonic scattering, which may lead to low thermal conductivity.

Other interesting suggestions to come from the recommendation engine are Tl_9SbTe_6, Ba_2Pb, and $FeAs_2$. Although none of these compounds were included in the thermoelectric database, they all scored highly within the recommendation engine. This prediction provides experimental validation since good thermoelectric performance has recently been demonstrated for these materials through property measurements or high-level DFT calculations.[36–38]

The suggestion of $TaAlO_4$, $SrCrO_3$, $TaSbO_4$ and other oxides expected to be insulators can be understood because the recommendation engine uses as training data references where stoichiometric formulas were primarily reported rather than doping details.[39,40] Nevertheless, with doping through substitution or reduction, these compound may exhibit moderate electrical performance. Further, these materials all feature extended structures that are highly connected in three dimensions, an important feature for low electrical resistivity. Moreover, the large mass contrast on the cation sublattice in $TaAlO_4$ (edge shared TaO_6 and AlO_6 octahedra) could lead to low thermal conductivity, and previous reports have shown that $SrCrO_3$ is metallic when synthesized under pressure.[41]

Many of the high-ranking candidate materials are interesting because of their highly connected extended structures, even though the recommendation engine does not use features of crystal structure to make its suggestions. The chief disadvantage to training prediction algorithms using crystal structure is that structure then becomes a *required input* for making predictions, and yet structure is by definition not available for uncharacterized materials. However, the absence of crystal structure does cause our engine difficulty where changes in crystal structure with similar elemental compositions cause large changes in physical properties. For example, both $DyPO_4$ and $LaPO_4$ are predicted to have low thermal conductivity. However, $LaPO_4$ is monazite, a corner edge-shared structure, whereas $DyPO_4$ is xenotime,[42] an edge-shared structure leading to inherently higher thermal conductivity.[43]

New materials and their properties

Our final and most important task in this work is to demonstrate that our recommendation engine can indeed guide researchers toward interesting experimental discoveries. Among the set of high-scoring candidate materials, we selected $Er_{12}Co_5Bi$ and $Gd_{12}Co_5Bi$ to characterize as thermoelectric materials due to their facile synthesis through arc melting, and due to the fact they are chemically quite distinct from known thermoelectrics (Fig. 1). While the $RE_{12}Co_5Bi$ (RE = rare earth) family of compounds has only been sparsely studied in the literature, their crystal structure and initial low-temperature electrical and magnetic properties have been reported by Mar and coworkers.[44] The crystal structure of $RE_{12}Co_5Bi$ is shown in Figure 4.

Interestingly, the crystal structure of our candidate thermoelectric exhibits notable similarity to the structures of known thermoelectrics, in spite of the fact that crystal structure was not an input feature for our recommendation engine. $Ho_{12}Co_5Bi$ is the eponymous structure prototype (orthorhombic, space group *Immm*) adopted by a series of rare-earth intermetallics $RE_{12}Co_5Bi$ ($RE = \mathrm{Y, Gd, \ldots, Tm}$). In this structure, the $Ho_{12}Bi$ icosahedra play an analogous role to the LaP_{12} icosahedra in the filled skutterudite prototype $LaFe_4P_{12}$; rare-earth atoms "rattling" within their 12-fold coordinated cages is the idiosyncratic feature of filled skutteru-

TABLE I. Several promising new thermoelectric compounds selected from our pre-computed list. The P values refer to the engine's confidence level that a given material will exhibit a room-temperature value for a particular property (*e.g.*, S or ρ) within the target ranges specified above. The full compound list is available for exploration at http://thermoelectrics.citrination.com.

Material	P_S	P_ρ	P_κ	P_{gap}	Composite	Comments
$TaPO_5$ and $TaVO_5$	0.894	0.793	0.958	0.987	3.537	High polyhedral connectivity and structural superlattices
Tl_9SbTe_6	0.845	0.871	0.999	0.876	3.46	Recently reported to be a good thermoelectric ($zT \approx 1$ at 600 K)
$TaAlO_4$	0.893	0.703	1	0.977	3.477	High mass contrast, high polyhedral connectivity (edge- and corner-sharing TaO_6 octahedra)
$SrCrO_3$	0.772	0.767	0.996	0.95	3.308	High polyhedral connectivity (3-D corner-sharing CrO_6 octahedra), metallic when made under high pressure
$TaSbO_4$	0.892	0.919	1	0.997	3.559	High polyhedral connectivity: layered, edge-sharing MO_6 octahedra
TiCoSb	0.981	0.714	0.958	0.833	3.467	TiCoSb is not a new compound, but has been studied as a high-zT material. However it was not included in training data.

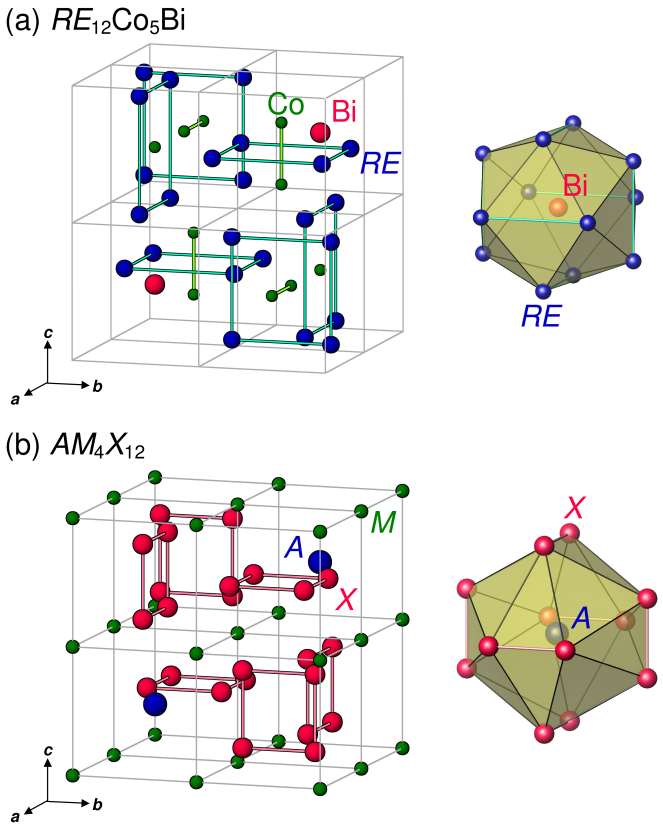

FIG. 4. (a) Crystal structure of $RE_{12}Co_5Bi$ (prototype $Ho_{12}Co_5Bi$), of which $Er_{12}Co_5Bi$ and $Gd_{12}Co_5Bi$ are exemplars. (b) Crystal structure of the filled skutterudites, which have the generic chemical formula AM_4X_{12}. These two structure types share an icosahedral motif consisting of $RE_{12}Bi$ and AX_{12} units, respectively.

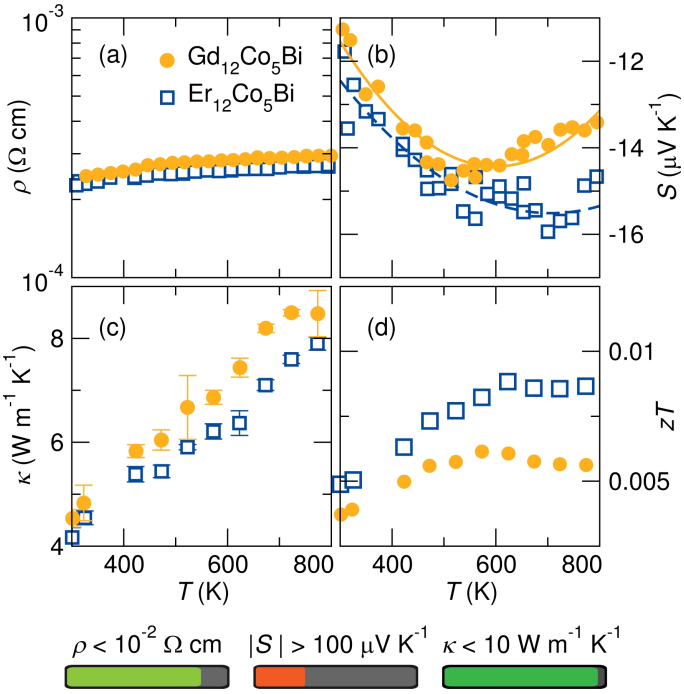

FIG. 5. Thermoelectric characterization of $RE_{12}Co_5Bi$ (RE = Gd, Er). (a) Electrical resistivity, (b) Seebeck coefficient, (c) thermal conductivity, and (d) thermoelectric figure of merit zT as a function of temperature. We also include the recommendation engine's confidence levels for the first three properties; the lowest-probability property, the Seebeck coefficient, is indeed found to be below the 100μV K^{-1} threshold.

dites that imparts low thermal conductivity so prized in thermoelectric materials. In fact, if the transition metal atoms, which occupy different sites in these structures, are disregarded, the $Ho_{12}Bi$ framework is an antitype to the LaP_{12} framework, with the roles of the rare-earth and group 15 elements reversed. We hypothesize its crystallographic similarity to skutterudite could be partly responsible for the thermoelectric behaviour of $RE_{12}Co_5Bi$ (RE = Gd, Er).

We give a full thermoelectric characterization of $Er_{12}Co_5Bi$ and $Gd_{12}Co_5Bi$ in Fig. 5. Based on these results, we report the discovery of a new thermoelectric class, which remains a completely unoptimized, pure bulk material and thus lends itself to further study. No-

tably, the material falls far outside the usual search space for thermoelectrics (Fig. 1 and Fig. 2), and was neither the result of simple interpolation between known compounds nor obvious from a strict chemical intuition standpoint. The electrical resistivity is commensurate with other high-performing materials such as chalcogenides, although the Seebeck coefficient is too low for the material to be competitive with the best-known thermoelectrics. Furthermore, the thermal conductivity is relatively high, but the filled cage structure lends itself to substitution that has successfully reduced thermal conductivity in the skutterudite systems.[3,45] In $RE_{12}Co_5Bi$ (RE = Gd, Er), the thermal conductivity from 300 K to 800 K ranges from $4\,W\,m^{-1}\,K^{-1}$ to $8\,W\,m^{-1}\,K^{-1}$, comparable to the half-Heuslers.[46,47] Note that these results are consistent with the engine's predictions (Fig. 5); the models give a high probability of achieving the thresholds for electrical conductivity (a) and thermal conductivity (c) (see confidence bar insets), while also suggesting a low probability of observing a large Seebeck coefficient (b). The electrical performance figure of merit κzT is around $0.03\,W\,m^{-1}\,K^{-1}$ at 400 K, which is actually higher than that of nearly 30% of the thermoelectrics in the Gaultois *et al.* thermoelectrics database;[8] of course, the database is a highly self-selected set of materials, consisting of literature-reported thermoelectrics, and would skew toward much higher κzT values than would a random subset of all crystalline materials. We note, of course, that the zT of several other thermoelectric materials can be significantly improved through carrier concentration tuning and microstructural engineering. For example, undoped polycrystalline Si has a 60-fold increase in performance after optimization, going from $zT < 0.01$ to 0.6 at 300 K.[48]

Another observation from Fig. 5 illustrates the scientific boon of studying entirely new classes of materials. Unexpectedly, $RE_{12}Co_5Bi$ (RE = Gd, Er) exhibits *increasing* thermal conductivity with temperature. (We note the recommendation engine successfully chose a material with a low thermal conductivity at room temperature, which would normally decrease with increasing temperature.) The increasing electrical resistivity with temperature indicates metallic electrical transport, so the electrical contribution to the total thermal conductivity should therefore decrease with increasing temperature. Additionally, the phonon contribution to thermal conductivity should also decrease with increasing temperature due to more phonon–phonon (Umklapp) scattering.[49] Thermal conductivity is calculated from the following relation: $\kappa = \alpha\,\rho\,C_p$, where α is thermal diffusivity, C_p is heat capacity, and ρ is density. Normally, thermal diffusivity has a negative temperature dependence whereas heat capacity and density both have positive temperature dependence. However, for this compound we observe a *positive* temperature dependence for the thermal diffusivity even after multiple measurements, the origin of which is not presently understood. Materials with increasing thermal conductivity with temperature are rare, though not unprecedented,[50,51] and further studies on this class of compounds to shed light on this anomaly could thus lead to new strategies for thermoelectric materials optimization.

IV. CONCLUSIONS

This initial experimental validation of our recommendation engine is encouraging. The present work represents the first time that machine learning has been used to suggest an experimentally viable new compound from true chemical white space, where no prior characterization had hinted at promising chemistries. The implication is that our approach–wherein a data-driven computational tool directly augments experimental capabilities and intuition–is a semi-rational way to discover new materials families that may have desirable properties. We suggest that such an paradigm could eventually replace trial-and-error and fortuity in the search for new materials across a wide variety of application areas.

ACKNOWLEDGMENTS

We thank Ram Seshadri for helpful discussions and insight. We thank the National Science Foundation for support of this research through NSF-DMR 1121053, as well as the Natural Sciences and Engineering Research Council of Canada (NSERC), and the DARPA SIMPLEX program N66001-15-C-4036. Additionally, this research made extensive use of shared experimental facilities of the Materials Research Laboratory: a NSF MRSEC, supported by NSF-DMR 1121053. MWG is thankful for support from NSERC through a Postgraduate Scholarship, support from the US Department of State through an International Fulbright Science & Technology Award, and support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska–Curie grant agreement No. 659764. BM and GJM are founders and significant shareholders in Citrine Informatics Inc.

[1] G. Nolas, J. Sharp, and H. J. Goldsmid, *Thermoelectrics: Basic principles and new materials developments* (Springer Science & Business Media, 2001) p. 292.

[2] G. J. Snyder and E. S. Toberer, "Complex thermoelectric materials," Nat. Mater. **7**, 105–114 (2008).

[3] G. S. Nolas, J. Poon, and M. Kanatzidis, "Recent developments in bulk thermoelectric materials," MRS Bull. **31**, 199–205 (2006).

[4] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, "Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling," Physical Review X **4**, 011019 (2014).

[5] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," APL Materials **1**, 011002 (2013).

[6] I. Terasaki, Y. Sasago, and K. Uchinokura, "Large thermoelectric power in $NaCo_2O_4$ single crystals," Phys. Rev. B **56**, R12685–R12687 (1997).

[7] Y. Kamihara, T. Watanabe, M. Hirano, and H. Hosono, "Iron-based layered superconductor La[O1 − xF$_x$]FeAs (x = 0.05–0.12) with T$_c$ = 26 K." J. Am. Chem. Soc. **130**, 3296–3297 (2008).

[8] M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio, and D. R. Clarke, "Data-driven review of thermoelectric materials: Performance and resource considerations," Chem. Mater. **25**, 2911–2920 (2013).

[9] B. Meredig, A. Agrawal, S. Kirklin, J. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," Physical Review B **89**, 094104 (2014).

[10] B. Meredig and C. Wolverton, "Dissolving the periodic table in cubic zirconia: Data mining to discover chemical trends," Chemistry of Materials **26**, 1985–1991 (2014).

[11] A. Jain, Y. Shin, and K. A. Persson, "Computational predictions of energy materials using density functional theory," Nat. Rev. Mater. **1**, 15004 (2016).

[12] G. Hautier, A. Jain, and S. P. Ong, "From the computer to the laboratory: materials discovery and design using first-principles calculations," J. Mater. Sci. **47**, 7317–7340 (2012).

[13] J. Carrete, N. Mingo, S. Wang, and S. Curtarolo, "Nanograined half-heusler semiconductors as advanced thermoelectrics: An ab initio high-throughput statistical study," Adv. Funct. Mater. **24**, 7427–7432 (2014).

[14] S. Bhattacharya and G. K. H. Madsen, "High-throughput exploration of alloying as design strategy for thermoelectrics," Phys. Rev. B **92** (2015), 10.1103/PhysRevB.92.085205.

[15] P. Gorai, P. Parilla, E. S. Toberer, and V. Stevanovic, "Computational exploration of the binary A$_1$B$_1$ chemical space for thermoelectric performance," Chem. Mater. **27**, 6213–6221 (2015).

[16] H. Zhu, G. Hautier, U. Aydemir, Z. M. Gibbs, G. Li, S. Bajaj, J.-H. Pöhls, D. Broberg, W. Chen, A. Jain, M. A. White, M. Asta, G. J. Snyder, K. Persson, and G. Ceder, "Computational and experimental investigation of TmAgTe$_2$ and XYZ$_2$ compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening," J. Mater. Chem. C **3**, 10554–10565 (2015).

[17] W. Chen, J.-H. Pöhls, G. Hautier, D. Broberg, S. Bajaj, U. Aydemir, Z. M. Gibbs, H. Zhu, M. Asta, G. J. Snyder, B. Meredig, M. A. White, K. Persson, and A. Y. Jain, "Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment," J. Mater. Chem. C (2016), 10.1039/c5tc04339e.

[18] G. K. H. Madsen and D. J. Singh, "Boltztrap. a code for calculating band-structure dependent quantities," Comput. Phys. Commun. **175**, 67–71 (2006).

[19] W. Khan, S. Borek, and J. Minar, "Correlation between the electronic structure, effective mass and thermoelectric properties of rare earth tellurides Ba$_2$MYTe$_5$ (M = Ga, In)," RSC Adv. **5**, 51461–51469 (2015).

[20] K. Rajan, "Materials informatics: The materials "gene" and big data," Annu. Rev. Mater. Res. **45**, 153–169 (2015).

[21] P. V. Balachandran, J. Theiler, J. M. Rondinelli, and T. Lookman, "Materials prediction via classification learning," Sci. Rep. **5**, 13285 (2015).

[22] Z.-K. Liu, L.-Q. Chen, and K. Rajan, "Linking length scales via materials informatics," Jom **58**, 42–50 (2006).

[23] K. Rajan, "Materials informatics," Materials Today **8**, 38–45 (2005).

[24] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis," Computational Materials Science **68**, 314–319 (2013).

[25] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," Nature materials **12**, 191–201 (2013).

[26] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," Physical review letters **114**, 105503 (2015).

[27] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," The Journal of Physical Chemistry Letters (2015).

[28] L. Breiman, "Random forests," Machine learning **45**, 5–32 (2001).

[29] B. Meredig, "Machine learning for the materials scientist," (2015).

[30] *See supplemental material at [URL to be inserted by AIP] for X-ray diffraction patterns and backscatter electron micrographs collected from the prepared materials.*

[31] P. Gorai, D. Gao, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, Q. Lv, V. Stevanovic, and E. S. Toberer, "Te design lab: A virtual laboratory for thermoelectric material design," Comput. Mater. Sci. **112**, 368–376 (2016).

[32] H. Chahboun, D. Groult, M. Hervieu, and B. Raveau, "β-NbPO$_5$ and β-TaPO$_5$: Bronzoïds, second members of the monophosphate tungsten bronze series (PO$_2$)$_4$(WO$_3$)$_{2m}$," J. Solid State Chem. **65**, 331–342 (1986).

[33] H. Chahboun, D. Groult, and B. Raveau, "TaVO$_5$, a novel derivative of the series of monophosphate tungsten bronzes (PO$_2$)$_4$(WO$_3$)$_{2m}$," Mater. Res. Bull. **23**, 805–812 (1988).

[34] M. Greenblatt, "Phosphate tungsten bronzes: A new family of quasi-low-dimensional metallic oxides," Int. J. Mod. Phys. B **07**, 3937–3971 (1993).

[35] X. Wang, Q. Huang, J. Deng, R. Yu, J. Chen, and X. Xing, "Phase transformation and negative thermal expansion in TaVO$_5$," Inorg. Chem. **50**, 2685–2690 (2011).

[36] Q. Guo, M. Chan, B. A. Kuropatwa, and H. Kleinke, "Enhanced thermoelectric properties of variants of Tl$_9$SbTe$_6$ and Tl$_9$BiTe$_6$," Chem. Mater. **25**, 4097–4104 (2013).

[37] D. Parker and D. J. Singh, "Alkaline earth lead and tin compounds Ae$_2$Pb, Ae$_2$Sn, Ae = Ca, Sr, Ba, as thermoelectric materials," Sci. Technol. Adv. Mater. **14**, 055003 (2013).

[38] P. Sun, N. Oeschler, S. Johnsen, B. B. Iversen, and F. Steglich, "Huge thermoelectric power factor: FeSb$_2$ versus FeAs$_2$ and RuSb$_2$," Appl. Phys. Express **2**, 091102 (2009).

[39] L. E. Depero and L. Sangaletti, "Cation sublattice and coordination polyhedra in ABO_4 type of structures," J. Solid State Chem. **129**, 82–91 (1997).

[40] K. M. Ok, N. Bhuvanesh, and P. Halasyamani, "SbSb$_x M_{1-x}$O$_4$ (M = NbV or TaV): Solid solution behavior and second-harmonic generating properties," J. Solid State Chem. **161**, 57–62 (2001).

[41] A. C. Komarek, T. Möller, M. Isobe, Y. Drees, H. Ulbrich, M. Azuma, M. T. Fernández-Díaz, A. Senyshyn, M. Hoelzel, G. André, Y. Ueda, M. Grüninger, and M. Braden, "Magnetic order, transport and infrared optical properties in the ACrO$_3$ system (A = Ca, Sr, and Pb)," Phys. Rev. B **84**, 125114 (2011).

[42] Y. X. Ni, J. M. Hughes, and A. N. Mariano, "Crystal chemistry of the monazite and xenotime structures," Am. Mineral. **80**, 21–26 (1995).

[43] M. R. Winter and D. R. Clarke, "Oxide materials with low thermal conductivity," J. Am. Ceram. Soc. **90**, 533–540 (2007).

[44] A. V. Tkachuk and A. Mar, "Structure and physical properties of ternary rare-earth cobalt bismuth intermetallics RE_{12}Co$_5$Bi (RE = Y, Gd, Tb, Dy, Ho, Er, Tm)," Inorg. Chem. **44**, 2272–2281 (2005).

[45] G. S. Nolas, G. A. Slack, D. T. Morelli, T. M. Tritt, and A. C. Ehrlich, "The effect of rare-earth filling on the lattice thermal conductivity of skutterudites," J. Appl. Phys. **79**, 4002–4008 (1996).

[46] T. Graf, C. Felser, and S. S. Parkin, "Simple rules for the understanding of Heusler compounds," Progress in Solid State Chemistry **39**, 1–50 (2011).

[47] J. E. Douglas, C. S. Birkel, M.-S. Miao, C. J. Torbet, G. D. Stucky, T. M. Pollock, and R. Seshadri, "Enhanced thermoelectric properties of bulk TiNiSn via formation of a TiNi$_2$Sn second phase," Applied Physics Letters **101**, 163514 (2012).

[48] A. I. Hochbaum, R. Chen, R. D. Delgado, W. Liang, E. C. Garnett, M. Najarian, A. Majumdar, and P. Yang, "Enhanced thermoelectric performance of rough silicon nanowires," Nature **451**, 163–167 (2008).

[49] G. Grimvall, *Thermophys. Prop. Mater.* (Elsevier, 1999) pp. 255–285.

[50] P. Jacobsson and B. Sundqvist, "Thermal conductivity and electrical resistivity of gadolinium as functions of pressure and temperature," Physical Review B **40**, 9541–9551 (1989).

[51] T. M. Tritt, *Thermal conductivity: Theory, properties, and applications*, Physics of Solids and Liquids (Springer, 2004).