Predicting thermoelectric transport properties from composition with attention-based deep learning

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
23 December 2022

REVISED
1 March 2023

ACCEPTED FOR PUBLICATION
15 March 2023

PUBLISHED
4 April 2023

PAPER

Predicting thermoelectric transport properties from composition with attention-based deep learning

Luis M Antunes[1,*], **Keith T Butler**[1,2] **and Ricardo Grau-Crespo**[1,*]

[1] Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6DX, United Kingdom
[2] School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom
* Authors to whom any correspondence should be addressed.

E-mail: l.m.antunes@pgr.reading.ac.uk and r.grau-crespo@reading.ac.uk

Keywords: thermoelectric, Seebeck, electrical conductivity, deep learning, machine learning

Supplementary material for this article is available online

Abstract

Thermoelectric materials can be used to construct devices which recycle waste heat into electricity. However, the best known thermoelectrics are based on rare, expensive or even toxic elements, which limits their widespread adoption. To enable deployment on global scales, new classes of effective thermoelectrics are thus required. *Ab initio* models of transport properties can help in the design of new thermoelectrics, but they are still too computationally expensive to be solely relied upon for high-throughput screening in the vast chemical space of all possible candidates. Here, we use models constructed with modern machine learning techniques to scan very large areas of inorganic materials space for novel thermoelectrics, using composition as an input. We employ an attention-based deep learning model, trained on data derived from *ab initio* calculations, to predict a material's Seebeck coefficient, electrical conductivity, and power factor over a range of temperatures and *n*- or *p*-type doping levels, with surprisingly good performance given the simplicity of the input, and with significantly lower computational cost. The results of applying the model to a space of known and hypothetical binary and ternary selenides reveal several materials that may represent promising thermoelectrics. Our study establishes a protocol for composition-based prediction of thermoelectric behaviour that can be easily enhanced as more accurate theoretical or experimental databases become available.

1. Introduction

Approximately 65%–70% of the energy used in industrial and transportation processes is wasted as heat [1]. Traditional means of converting waste heat into electricity involve the use of devices such as Rankine steam engines, but these methods tend to involve machines comprised of multiple moving parts, which require maintenance and upkeep, and are difficult to scale. Thermoelectric generators, which are solid-state devices without moving parts, provide an alternative and convenient solution to waste heat recovery [2]. A thermoelectric generator is typically built from two semiconducting materials, one with *n*-type conductivity, and the other with *p*-type conductivity. The materials are assembled with electrical and thermal connections between a heat source, at temperature T_{hot}, and a heat sink, at temperature T_{cold}. The efficiency of a thermoelectric generator depends strongly on the temperature difference, T_{hot}–T_{cold}, as well as on the physical characteristics of the materials used, which are usually summarized in the *figure of merit*:

$$zT = \frac{S^2 \sigma T}{\kappa}. \tag{1}$$

Here, S is the Seebeck coefficient, σ is the electrical conductivity, T is the absolute temperature, and κ is the thermal conductivity, which contains two main contributions: the lattice thermal conductivity κ_{latt} due to

crystal vibrations, and the electronic thermal conductivity κ_{elec} due to heat-carrying diffusion of electrons in the solid. The term $S^2\sigma$ is commonly referred to as the *power factor*. The higher the dimensionless figure of merit zT, the more efficient the thermoelectric material. Consequently, a good thermoelectric material must exhibit a large (absolute) Seebeck coefficient, good electrical conductivity, but low thermal conductivity.

Finding good thermoelectric materials with the right combination of properties is a difficult task, because of the interdependence of the properties that appear in the figure of merit. Other factors, like abundance and toxicity, further complicate the search for good candidate materials. While thermoelectricity has been a known phenomenon since the early 1800s [3, 4], relatively few materials have been discovered that are effective enough for practical applications. Well-studied thermoelectric materials, such as Bi_2Te_3 and PbTe, are suitable for various applications, but are often too expensive or too toxic for widespread adoption [5]. If thermoelectric generators are to be deployed on a scale large enough to have a positive environmental impact, new materials are needed [6]. The search for novel thermoelectrics is an active field of research [7–9]. A range of promising thermoelectric materials have been discovered experimentally, either serendipitously, or as a result of chemical intuition. In the low temperature range (near room temperature), where thermoelectric materials are typically used for cooling applications or low-grade heat recovery, top performances are achieved with Bi_2Te_3-based alloys (e.g. $zT = 1.2$ and power factor of 45 μW cm^{-1} K^{-2} for $(Bi_{1-x}Sb_x)_2Te_3$ at room temperature [10]). Materials based on PbTe exhibit some of the best performances in the temperature range between 500 K and 900 K (e.g. zT of 2.5 at around 800 K in p-doped $Pb_{1-x}Sr_xTe$, with a maximal power factor above 30 μW cm^{-1} K^{-2}) [11]. At very high temperatures, such as those used in radioisotope thermoelectric generators ($\sim$1000 K or above), Si–Ge alloys exhibit some of the highest figures of merit (e.g. peak zT of about 1.3 at 1173 K in an n-type nanostructured SiGe bulk alloy, corresponding a maximal power factor of $\sim$30 μW cm^{-1} K^{-2}) [12]. Other families of compounds that are attracting considerable attention as promising thermoelectric materials include the metal chalcogenides (e.g. SnSe, Cu_2Se) [13–15], skutterudites (e.g. $CoAs_3$, $CoSb_3$) [16], Zintl compounds (e.g. $YbZn_2Sb_2$) [17], clathrates (e.g. $Sr_8Ga_{16}Ge_{30}$) [18], Heusler and half-Heusler compounds (e.g. TiNiSn, ZrNiSn) [19–21], and metal oxides (e.g. $NaCo_2O_4$, $Ca_3Co_4O_9$) [22, 23]. Hole-doped polycrystalline SnSe is the record-holder in terms of thermoelectric figure of merit, and is reported to exhibit a zT of 3.1 at 783 K [24]. In principle, there are no theoretical or thermodynamic limits for the possible values of zT [25], so there is hope that materials with even higher values of zT can be found.

In addition to trial-and-error exploration, and the rational design of materials, computational techniques based on the combination of density functional theory (DFT) and high-throughput screening (HTS) are becoming increasingly prevalent in the search for new thermoelectrics [26–28]. The first report of such an approach was made in 2006 by Madsen, who screened a dataset of 1630 Sb-containing compounds derived from existing crystal structure databases, and based on the results of *ab initio* calculations, identified LiZnSb as an interesting thermoelectric material [29]. Since then, a number of studies involving the use of HTS in the search for new thermoelectric candidates have followed [30–38]. The increasing availability of distributed computing infrastructure, along with the development of workflow management software [39–46], has enabled the growing adoption of this approach.

While DFT-based HTS is becoming more prevalent, there remains a large gap between the size of chemical space that is accessible with this approach, and the size of the space of all possible inorganic materials. To bridge that gap, and to further accelerate computational predictions of thermoelectric behaviour, techniques involving the use of machine learning (ML) have been gaining popularity in the search for new thermoelectric materials [47–51]. Data for these ML approaches can come from either theoretical calculations, or from physical experiments. HTS studies have been producing *ab initio* results for thousands of materials, and these results can be assembled into datasets that are usable with ML algorithms. Since experimental data is scarcer, the outputs of *ab initio* calculations are often the source of data for ML approaches. Using ML to learn models that predict the output of *ab initio* calculations is sensible, since invoking an ML model is much faster (and less computationally expensive) than carrying out an *ab initio* calculation. ML models of various thermoelectric properties, such as the Seebeck coefficient [52–55], electrical conductivity [56, 57], power factor [58–61], lattice thermal conductivity [62–74], and even zT [75–80], have been developed.

Deep learning is a particular ML approach that has been very successful in recent years, and has seen adoption in many diverse areas of science [81, 82]. It is characterized by the combination of large datasets with various neural network architectures, together with advantages such as automatic feature extraction. In materials chemistry, deep learning approaches have been adopted for prediction of materials properties [83]. General purpose deep learning architectures for materials properties prediction, such as ElemNet [84], IRNet [85, 86], CGCNN [87], MEGNet [88], Roost [89], and CrabNet [90] have become powerful tools in the materials informatics toolbox.

Here, we utilize attention-based deep learning, together with existing datasets derived from high-throughput DFT calculations [91], to predict the thermoelectric transport properties of a material. The input to the model is a representation of a material's composition, and optionally the material's band gap. The output is a collection of predictions for a range of temperatures, for various doping levels, and for *n* and *p* doping types. This structure-free approach allows us to scan regions of materials space of hypothetical but plausible compounds, whose structures are not known. Our multi-output approach creates a thermoelectric behaviour profile for a material at a number of different conditions, which offers advantages over narrower models that only make predictions for specific conditions.

2. Methods

2.1. Datasets

Our models are trained on the dataset published in 2017 by Ricci *et al* [92] (henceforth the *Ricci database*). This is a freely available electronic transport database containing the computationally derived electronic transport properties for 47 737 inorganic compounds with stoichiometric compositions. The properties listed include the Seebeck coefficient, the electrical conductivity, and the electronic thermal conductivity, obtained using DFT in the generalized gradient approximation (GGA), and the Boltzmann Transport equation through the BoltzTraP computer software [93], under the constant relaxation time approximation (CRTA). They also associate the computed band gap with each entry, amongst several other properties. For each compound, the aforementioned properties were determined at various temperatures (100 K–1300 K in 100 K increments), for *p*- and *n*-doping types, and at five doping levels (ranging from 10^{16} to 10^{20} cm^{-3}). Moreover, each property is a tensor quantity reported as a 3×3 matrix. The database is altogether quite large, with 18 617 430 data points if one considers only the values of the diagonal elements S_{xx}, S_{yy}, and S_{zz} (i.e. 47 737 compounds × 13 temperatures × 2 doping types × 5 doping levels × 3 diagonal elements). Another important consideration is that there are duplicate compounds in the database in terms of composition (corresponding to possible polymorphs). While there are 47 737 unique compounds in the database when structure is considered, there are only 34 628 unique compositions. In this study, we form a dataset of compositions from the Ricci database and their associated thermoelectric transport properties. For cases where there are multiple entries with the same composition, we obtain the DFT-derived energy per atom of each polymorph, and use the transport properties and band gap of the entry corresponding to the polymorph with the lowest energy per atom.

Additionally, we form a dataset consisting solely of compositions and their associated electronic band gaps derived from DFT, by combining data from the Materials Project [94] and the Ricci database. We obtained 126 335 structures and their associated electronic band gaps from the Materials Project, which corresponded to 89 444 unique compositions, which are used to train the band gap predictor. Where there were multiple structures for a composition, again we used the band gap of the polymorph with the lowest computed energy per atom.

The Ricci database has some important limitations. As discussed in [92] and elsewhere (see [95] for a recent perspective), the use of the GGA and CRTA in the prediction of electronic transport can lead to large discrepancies with respect to experiment. In particular, GGA band structures generally exhibit too narrow gaps and too large bandwidths, which tends to exaggerate the electronic conductivity. The CRTA, especially when unaccompanied by physically-sound prediction of relaxation times, misses important differences in scattering mechanisms across compounds. Furthermore, the calculations in [92] did not consider spin–orbit coupling (SOC), which often has an important effect on the electron transport properties of materials [96]. Inevitably, any ML model based on this dataset will carry over these limitations of the underlying data, hindering the quality of the predictions with respect to experimental values. However, our approach establishes a protocol capable of efficiently mapping composition to thermoelectric behaviour, which can be easily refined once more accurate databases become available. This is important because, in addition to the improvement of existing *ab initio* databases, there are ongoing efforts to create large databases of thermoelectric properties from experiment [97], so we anticipate our model will keep evolving following the expansion of such datasets.

2.2. ML models

We build ML models that predict the Seebeck coefficient, the electrical conductivity, and the power factor using data from the Ricci database. Our multi-output regression models [98, 99] produce predictions of transport properties at 13 temperatures, 5 doping levels, for 2 doping types, given a material's composition and (optionally) band gap. The task is to predict the mean of the diagonal elements of the Seebeck tensor,

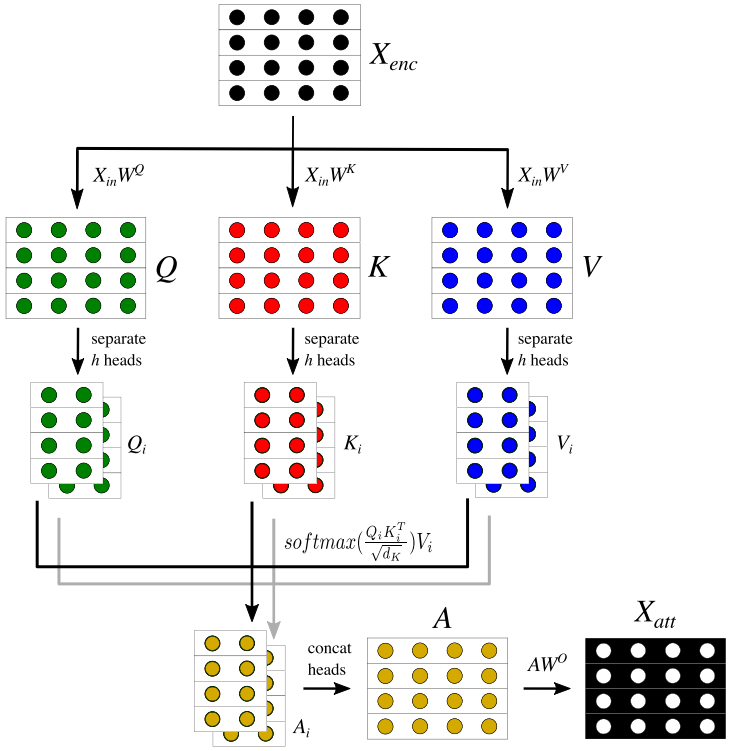

Figure 1. Depiction of the multi-head self-attention operation. The input, $X_{\text{enc}} \in \mathbb{R}^{n \times d_{\text{model}}}$, consists of the fractionally encoded representations for n constituent elements of the composition, each with d_{model} components. Linear transformations are applied to the input to produce the query, $Q \in \mathbb{R}^{n \times d_{\text{model}}}$, the key, $K \in \mathbb{R}^{n \times d_{\text{model}}}$, and the value, $V \in \mathbb{R}^{n \times d_{\text{model}}}$, using the learned parameters W^Q, W^K, and W^V, respectively. The query, key and value are each subsequently separated into h heads (indexed here by i). The corresponding queries, Q_i, keys, K_i, and values, V_i, are combined to produce the attention products, A_i, by multiplying the *softmax* of a scaled dot-product of the queries and keys with the values. After the attention products are concatenated to produce A, a linear transformation of A using the learned parameters W^O produces the output of multi-head self-attention, $X_{\text{att}} \in \mathbb{R}^{n \times d_{\text{model}}}$.

$(S_{xx} + S_{yy} + S_{zz})/3$, henceforth referred to as the Seebeck coefficient, S, and the mean of the diagonal elements of the electrical conductivity tensor, $(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})/3$, henceforth referred to as the electrical conductivity, σ. The values for electrical conductivity in the Ricci database are reported per unit of relaxation time. Hence, in this report, electrical conductivity, σ, will more precisely refer to electrical conductivity per unit relaxation time, σ/τ. The target power factor is also predicted, and is defined here as the mean of the directional power factors, $(S_{xx}^2\sigma_{xx} + S_{yy}^2\sigma_{yy} + S_{zz}^2\sigma_{zz})/3$. It will be denoted by *PF*, and is also given per unit of relaxation time.

More formally, the task is to learn a function $f: \mathcal{X} \to \mathcal{Y}$, given a training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \mid 1 \leqslant i \leqslant k\}$, with $\mathbf{x}_i \in \mathcal{X}$, $\mathbf{y}_i \in \mathcal{Y}$, and k labelled examples. Here, the $\mathbf{x}_i$ represent a multi-dimensional input describing the features of an exemplar, and $\mathbf{y}_i$ represent a multi-dimensional target associated with $\mathbf{x}_i$. A training procedure is used to find f, and involves the minimization of a loss, $L: \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}$, that specifies the degree of disagreement between the true values $\mathcal{Y}$, and $\hat{\mathcal{Y}}$, the output of f given members of $\mathcal{X}$.

Here, we use two different forms of f: a Random Forest (RF) [100], and an attention-based deep neural network based on the CrabNet architecture, which is the state-of-the-art tool for property prediction from materials composition, as demonstrated in the work by Wang *et al* [90]. The CrabNet architecture incorporates a multi-head self-attention mechanism, originally introduced in the Transformer deep learning model [101], which provides the added advantage of enhanced interpretability. Traditionally, a Transformer transforms an input sequence to an output sequence using an encoder followed by a decoder. However, CrabNet consists strictly of an encoder, followed by a number of Residual blocks [102]. Moreover, instead of a sequence of words, CrabNet operates on a bag of atoms, and consequently, instead of using a positional encoding of the input, it encodes the relative amounts of atoms present.

The input to the model thus consists of a material's composition. Formally, the input, $X_{\text{in}} \in \mathbb{R}^{n \times d_{\text{in}}}$, consists of d_{in}-dimensional representations for the n constituent elements of the composition. The first step involves the encoding of the relative amounts of atoms into X_{in}, referred to as *fractional encoding* (see [90] for more details), resulting in $X_{\text{enc}} \in \mathbb{R}^{n \times d_{\text{model}}}$, where d_{model} is given as a hyperparameter. This is followed by the sequential application of a number of Transformer blocks. Each Transformer block begins by performing a multi-head self-attention operation. (figure 1) The self-attention operation allows the model to learn to

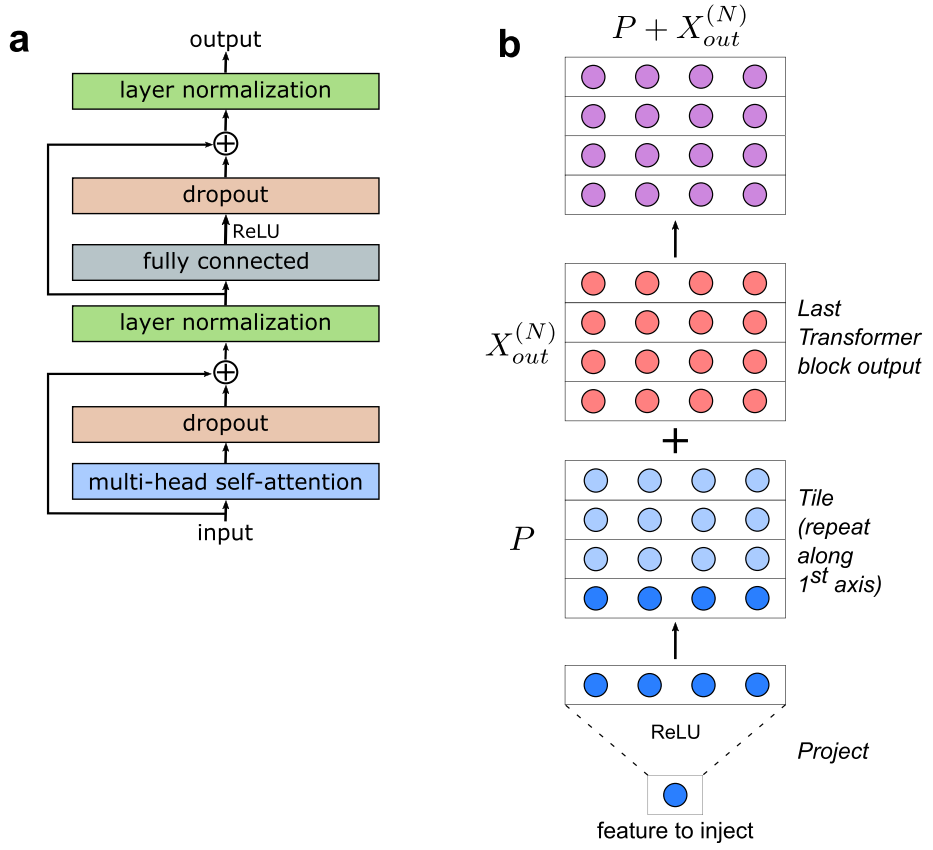

Figure 2. (a) Components of the Transformer block. (b) A depiction of how additional features, such as band gap, are incorporated into the CrabNet architecture. The addition operation refers to element-wise addition. The projection of the feature involves trainable parameters.

attend to the relationships between the atoms of the composition, in the context of the task. The 'attention weights' are encoded into a $n \times n$ matrix, associated with each of h attention heads, by applying the *softmax* operation to a scaled dot-product of a query, $Q_i \in \mathbb{R}^{n \times d_K}$, and a transposed key, $K_i^T \in \mathbb{R}^{d_K \times n}$, where $d_K = d_{\text{model}}/h$ specifies the key (and query) dimension for an attention head.

The Transformer block follows the multi-head self-attention operation with layer normalization [103], dropout [104], and feed-forward *ReLU* operations (figure 2(a)). The output of a Transformer block, $X_{\text{out}} \in \mathbb{R}^{n \times d_{\text{model}}}$, thus consists of the same dimensions as the input, which allows multiple Transformer blocks to be connected serially.

Since it may also be desirable to provide additional information beyond composition to the model, we augment the CrabNet architecture so that additional features may be provided. There are a number of ways this could be accomplished, but we choose to borrow an approach from computer vision [105], and perform a projection on v input features, $\mathbf{u} \in \mathbb{R}^v$, followed by a tiling operation, so that the resulting projected features, $P \in \mathbb{R}^{n \times d_{\text{model}}}$, have the same dimensions as the output of a Transformer block. Finally, we perform element-wise addition, $P + X_{\text{out}}^{(N)}$, where N is the number of Transformer blocks, and $X_{\text{out}}^{(N)}$ denotes the output of the last Transformer block (figure 2(b)). While any number of extra features may be supplied to the model this way, in this work, we (optionally) supply a single feature, the band gap E_g, associated with the material.

Finally, the output $P + X_{\text{out}}^{(N)}$ is given to three separate output heads. Each output head consists of a series of Residual blocks, followed by a fully connected linear layer that produces the final predictions for each of S, σ, and PF. This multi-head architecture has advantages in terms of convenience, efficiency, and also usually provides better overall performance on the task when compared to using a separate (single-head) model for each property predicted. (See supplementary table 1 for a comparison of the performance of architectures with different output head numbers.) For clarity, and to differentiate it from the original CrabNet architecture, we refer to this model as Compositionally-restricted attention-based ThermoElectrically-oriented Network (CraTENet); its architecture is illustrated in figure 3.

The CraTENet model thus expects a dataset consisting of compositions, $X_i \in \mathbb{R}^{n \times d_{\text{in}}}$, and associated thermoelectric transport properties, $\mathbf{y}_i^S, \mathbf{y}_i^\sigma, \mathbf{y}_i^{PF} \in \mathbb{R}^m$, where $\mathbf{y}_i^S$, $\mathbf{y}_i^\sigma$, and $\mathbf{y}_i^{PF}$, represent the S, σ, and PF transport values, respectively, at all temperatures, doping levels and doping types, each an m-dimensional

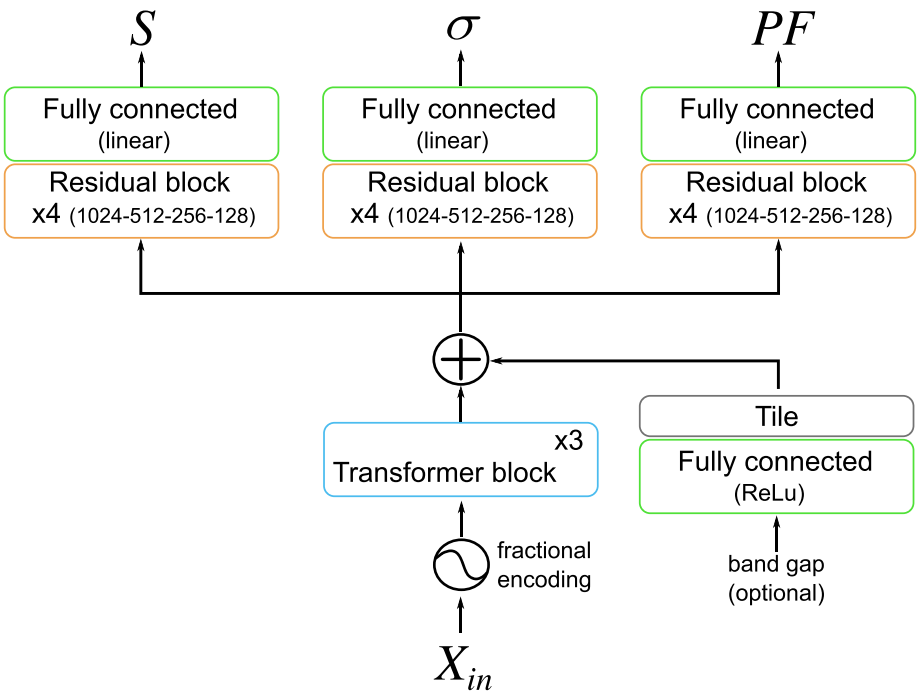

Figure 3. The multi-head attention-based architecture, CraTENet, used in this study. Each of the three output heads are multi-valued, containing the prediction of the Seebeck (S), electrical conductivity (σ), and power factor (PF), at different temperatures, doping levels, and doping types.

vector. Optionally, a band gap, $E_{g_i} \in \mathbb{R}$, may be associated with X_i. The dataset is thus $\{((X_i, E_{g_i}), (\mathbf{y}_i^S, \mathbf{y}_i^\sigma, \mathbf{y}_i^{PF})) \mid 1 \leqslant i \leqslant k\}$, where k is the number of examples.

As in the CrabNet and Roost models, the CraTENet model learns the heteroscedastic aleatoric uncertainty (i.e. how the variance of the predicted variable depends on the independent variables), explicitly through the loss function [106, 107]. Here, the calculated variance is a measure of the uncertainty associated with the incompleteness of the descriptor used (which is why the calculated variance decreases considerably when the band gap information is added to the descriptor). This variance is different from the epistemic variance related to the quality of the model parameterization. Whereas the CrabNet and Roost models use a Robust L1 loss to estimate the uncertainty, we find that a Robust L2 loss, which places an L2 distance on the residuals, results in superior performance for this task (see supplementary note 2 and supplementary table 4). The loss, L_p, for a particular thermoelectric transport property p, is given by:

$$L_p = \frac{1}{2k} \sum_{i=1}^{k} \sum_{j=1}^{m} (\hat{y}_{ij}^p - y_{ij}^p)^2 \exp(-\ln \hat{s}_{ij}^p) + \ln \hat{s}_{ij}^p \tag{2}$$

where k is the number of examples in the dataset, and m is the number of components of the output vector $\mathbf{y}_i^p$. The prediction of the ith example is $\hat{\mathbf{y}}_i^p$, and $\hat{y}_{ij}^p$ the jth component of the ith prediction (also considered the predictive mean in this context). The corresponding target value is y_{ij}^p. Finally, the predictive aleatoric variance for the jth component of the ith prediction is given by $\hat{s}_{ij}^p$. The form of this loss arises from the assumption that the uncertainty in the observations follows a Gaussian distribution. Also, the term $\exp(-\ln \hat{s}_{ij}^p)$ is used in place of the term $1/\hat{s}_{ij}^p$ for numerical stability reasons, such as to avoid a potential division by zero. Since the model utilizes a separate output head for each of the three thermoelectric transport properties being learned, the overall loss, L, to be minimized is given by:

$$L = \alpha L_S + \beta L_\sigma + \gamma L_{PF} \tag{3}$$

where α, β, and γ are constants which weight the importance of each of the terms in the loss L. In this work, $\alpha = \beta = \gamma = 1$.

Finally, we also train a band gap predictor from composition, using the original CrabNet model and the expanded band gap dataset described previously. The fact that the band gap predictor can be trained with a much larger dataset than the one used for training the CraTENet model justifies our attempt to use the band

gap as an additional input to CraTENet for the prediction of transport coefficients. As shown in the section 3, if the band gap predictor is sufficiently accurate, the inclusion of the predicted band gap in the CraTENet input can lead to overall performance enhancement, even if composition remains the only global input of the model.

2.3. ML model training and evaluation

For all CraTENet and CrabNet models, the input, X_{in}, consisted of $n = 8$ elements, and was zero-padded if the composition consisted of less than eight elements. Each element in the input was described with a SkipAtom distributed representation [108] with dimensions $d_{\text{in}} = 200$. (We performed experiments, as described in supplementary note 1 and supplementary table 3, to determine the performance of different descriptors). The default architectural hyperparameters of the original CrabNet model were used without further tuning. Specifically, both models consisted of $h = 4$ attention heads in each of three sequential Transformer blocks; the hyperparameter d_{model} was set to 512. The output (or output head) consisted of four sequential Residual blocks, with 1024, 512, 256, and 128 nodes respectively. For all neural network training procedures, a mini-batch size of 128 and a learning rate of 10^{-4} was used, which were derived from a hyperparameter grid search. The Adam optimizer, with an epsilon parameter of 10^{-8}, was used. [109] All neural network models were implemented using the TensorFlow [110] and Keras [111] software libraries.

The input for the RF models was a descriptor described by Meredig *et al* [112], as implemented in the Matminer software library [113]. It is a local descriptor of composition, containing properties such as atomic fractions, electronegativities, and radii. In some experiments, we concatenate an unscaled band gap feature to the descriptor. The RF model hyperparameters were determined using a grid search. The number of estimators was set to 200, the maximum depth was set to 110, the maximum number of features was set to 36, and bootstrapping was used. We used the implementation provided in the Scikit-learn software library [114].

Because the electrical conductivity values in the Ricci database are given per unit of relaxation time τ, which is an exceedingly small number (of the order of 10^{-15} s), the target values for σ and PF are numerically quite large. The values also vary by orders of magnitude, reflecting the distribution across metallic, semiconducting and insulating conductivity ranges. For these reasons, the models learn $\log_{10}\sigma$ and $\log_{10} PF$ instead. All output targets are standardized by removing the mean and scaling to unit variance. The band gap, when it is provided to the CraTENet model, is given in eV units and unscaled.

Neural network model training was carried out in one of two contexts: a 90–10 holdout experiment, or a ten-fold cross-validation experiment. For 90–10 holdout experiments, we split the dataset $\mathcal{D}$ into a set $\mathcal{A}$ consisting of 90% of the data, and a set $\mathcal{B}$ consisting of 10% of the data. For the neural network models, set $\mathcal{A}$ was further split into a training set $\mathcal{T}$ consisting of 90% of $\mathcal{A}$, and a validation set $\mathcal{V}$ consisting of 10% of $\mathcal{A}$. Early stopping was used (with a patience of 50) to determine the optimal number of epochs to train, using $\mathcal{V}$ as the validation set. Then, the model was re-trained on all of $\mathcal{A}$ for the number of epochs determined to be optimal, again starting from random parameters. Test set $\mathcal{B}$ was then used to evaluate performance of the re-trained model (see [115] for more information on this approach). The RF models were trained on $\mathcal{A}$, and evaluated on $\mathcal{B}$. The same random seed was used throughout when creating the splits, to ensure identical splits for all experiments.

For the ten-fold cross-validation experiments, we followed the same procedure as for the 90-10 holdout experiments, except that we create ten mutually exclusive splits, each consisting of 10% of $\mathcal{D}$ for testing and 90% of $\mathcal{D}$ for training, using the same random seed for all experiments, and repeating the hold-out procedure for each of the ten splits. The performance on $\mathcal{B}$ was averaged across the 10 splits to yield the final performance of the model.

The objective of all neural network training experiments was to minimize either the Robust L1 or Robust L2 loss. The objective of RF training was to minimize the mean squared error (MSE) criterion. The mean absolute error (MAE) and coefficient of determination (R^2) metrics were used to assess model performance. To produce the final neural network models to be used for inference on composition space outside the datasets used for training and evaluation, we train the models on all available data $\mathcal{D}$ for a number of epochs determined from the corresponding ten-fold cross-validation experiment, by averaging the number of epochs required for each fold. The final RF models to be used for inference were simply trained on all available data $\mathcal{D}$.

2.4. DFT calculations

We performed a small number of DFT calculations in systems not found in the Ricci database, for testing purposes. All calculations were carried out using the Vienna *Ab initio* Simulation Package (VASP) [116, 117], and the calculation settings were chosen to follow the work of Ricci *et al* [92] as closely as possible. The

Perdew–Burke–Ernzerhof [118] exchange-correlation functional, which is based on the GGA, was used in conjunction with the projector augmented-wave method [119, 120] to describe the interaction between core and valence electrons. All structures were fully relaxed until the force on each atom is below 0.02 eV $Å^{-1}$. Spin polarization was on, and magnetic moments on the ions were initialized in a high-spin ferromagnetic configuration, and then allowed to relax to the spin groundstate. A self-consistent static calculation was performed using 90 k-points $Å^3$ (in terms of reciprocal lattice volume) for systems with band gaps $\geqslant$0.5 eV, and 450 k-points $Å^3$ for systems with band gaps $<$0.5 eV. Subsequently, a non-self-consistent calculation was performed to evaluate the band structures on a uniform k-point grid, with 1000 k-points $Å^3$ for systems with band gaps $\geqslant$0.5 eV, and 1500 k-points $Å^3$ for systems with band gaps $<$0.5 eV. SOC was not considered.

The Seebeck coefficient, S, and the electrical conductivity, σ, were computed using the BoltzTraP2 software package [121]. Interpolation was first performed by sampling five irreducible k-points for each k-point from the VASP output. The band structure was then integrated to obtain sets of Onsager coefficients. The temperature range 100 K–1300 K was explored, in increments of 100 K, at 5 different doping levels (10^{16}–10^{20} cm^{-3}), for both n and p doping types. We verified that our *ab initio* procedure emulates the one that was used to create the Ricci database by comparing our results to those of the Ricci database for a number of compounds (see supplementary figure 3).

3. Results and discussion

3.1. Thermoelectric property prediction

Both the CraTENet model and a RF model were trained on the 34 628 entries of the Ricci database. To establish the generalization error of the models, ten-fold cross-validation was performed. Since multi-target regression of thermoelectric transport properties on composition is essentially a new task, unreported in the literature, there are no existing benchmarks to compare with. We created simple baseline models, such as linear regression with a Meredig feature vector, or simply taking the median of the target values, but these models performed considerably worse than the ML models presented here. To simplify presentation, we leave out the baseline results.

The results of ten-fold cross-validation are presented in table 1. For the remainder of this article, 'CraTENet' will refer to either the version of the model which does not accept a band gap input or to the CraTENet model in general, depending on the context, whereas 'CraTENet+gap' will specifically refer to the version of the model which requires a band gap input. As is evident from the results in table 1, the models which utilize the band gap clearly outperform those which do not. The band gap is thus an important predictor of thermoelectric transport properties. In both the case where band gap is or is not provided, the CraTENet model outperforms the RF model in terms of MAE. The RF performs better in terms of R^2, but generally only when band gap is absent. Moreover, the models appear to perform slightly better when predicting the $\log\sigma$ than the Seebeck. Prediction of the $\log PF$ appears to be the most problematic, with the R^2 for this property being noticeably lower than for the other two properties. The best thermoelectric materials have Seebeck coefficients in the order of several hundreds of $\mu V\,K^{-1}$, so the resulting MAE is still reasonably small by comparison.

The results in table 1 represent predictions made for all temperatures, doping levels and doping types. However, it is useful to understand how the models perform for different cross-sections of the data. For example, the ten-fold cross-validation results as a function of doping type are presented in table 2. To obtain the values in table 2, only the predictions for a given doping type were considered when computing the metrics, across all doping levels and temperatures. The CraTENet model appears to perform better on the p-type predictions, though it depends on which metric one considers. In figure 4, ten-fold cross-validation results are presented as a function of temperature and doping level. It is interesting (and useful to know) that the PF predictions are worse, in terms of R^2 values, at lower temperatures and higher doping levels. The MAE, on the other hand, does not show significant variations with the conditions of temperature and doping, remaining constant at around 0.40 for $\log PF$. The ability of the model to find the most promising compounds for further study depends on the magnitude of the error relative to the width of the distribution of values. If the absolute error is roughly constant, the ability of the model to discriminate between compounds can be expected to be worse for a dataset that is more narrowly distributed. In this sense, the R^2 is a better metric because it is related to the ratio between the MSE and the variance. Supplementary figure 14 shows that at high doping levels the distribution of values is narrower, and therefore the R^2 (as well as our ability to select the best compounds) decreases. The effect of temperature is a bit less pronounced, but because increasing temperature also tends to widen the distribution, the R^2 is slightly better at high temperatures. The variations in the distribution of PF at different conditions are related to the balance

Table 1. Ten-fold cross-validation results for each of the transport properties for the CraTENet and Random Forest (RF) models, both with and without a provided band gap, in terms of MAE and R^2. Each value represents the mean result across ten folds, across all temperatures, doping levels and doping types. Bold values represent the best result for a class of models (*i.e.* with or without band gap) for a particular property.

	S		$\log \sigma$		$\log PF$	
	MAE (μV K^{-1})	R^2	MAE	R^2	MAE	R^2
CraTENet	**114**	0.780	**0.576**	0.768	**0.452**	0.616
RF	141	**0.798**	0.696	**0.780**	0.476	**0.632**
CraTENet+gap	**49**	**0.962**	**0.260**	**0.968**	**0.380**	0.731
RF+gap	54	0.961	0.301	0.964	0.398	**0.737**

Table 2. Ten-fold cross-validation performance of the CraTENet model as a function of doping type. Each value represents the mean result for each doping type across all ten folds, across all temperatures and doping levels. Bold values represent the best result between *p*- and *n*-doping types for a class of models (*i.e.* with or without band gap) for a particular property.

		S		$\log \sigma$		$\log PF$	
	Doping	MAE (μV K^{-1})	R^2	MAE	R^2	MAE	R^2
CraTENet	*p*-type	119	**0.636**	0.589	**0.775**	0.465	**0.631**
CraTENet	*n*-type	**109**	0.627	**0.562**	0.758	**0.439**	0.594
CraTENet+gap	*p*-type	**49**	**0.945**	0.260	**0.972**	0.388	**0.747**
CraTENet+gap	*n*-type	50	0.925	0.260	0.962	**0.371**	0.709

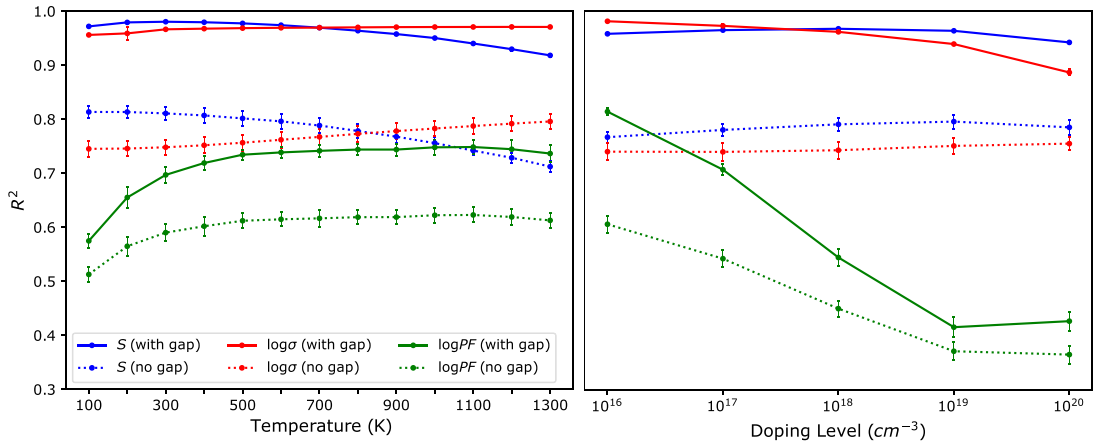

Figure 4. Ten-fold cross-validation performance (R^2) of the CraTENet model as a function of temperature and doping level. On the left, each point represents the mean performance for each temperature across all ten folds, across all doping levels and doping types. On the right, each point represents the mean performance for each doping level across all ten folds, across all temperatures and doping types. The dotted series represent the model's results without a provided band gap.

between the Seebeck coefficient and the conductivity in metallic *vs.* gapped materials in the calculations of [92]; further details can be found in the supplementary material.

To understand how the predictions compare to the 'true' values (i.e. the target DFT values), and how the prediction errors are distributed, it is useful to plot the true versus the predicted values, and also the distribution of absolute errors, as in figure 5. The plots show that most predictions lie close to the true values. Moreover, the distribution of absolute errors indicates that the majority of errors are less than the overall MAE values.

As the CraTENet model performs best when access to a band gap is available, it is important to understand how the performance of the model depends on the quality of the band gap provided, since, in many contexts, an experimental or *ab initio* band gap may not be available. In screening scenarios, the band gap could originate from a predictive model. Thus, to understand how the CraTENet model depends on the quality of the band gap, we performed sensitivity experiments, by incrementally degrading high quality band gaps (i.e. derived from *an initio* methods) by adding Gaussian noise, and then supplying these 'lower-quality' band gaps to the model. The results are presented in figure 6. In the figure, the horizontal axis along the top of the plot represents the resulting MAE (in eV) after a certain percentage of noise has been added to the

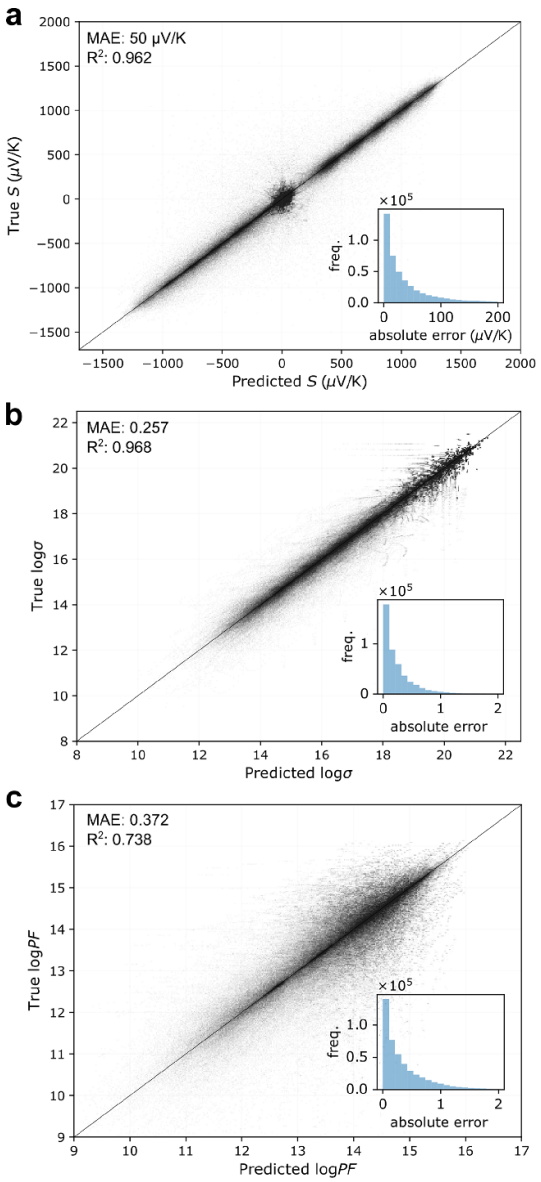

Figure 5. True values vs. predicted values of the test set of a 90–10 holdout experiment using the CraTENet+gap model, for each of the transport properties, across all temperatures, doping levels, and doping types. Each plot contains 450 190 points, as there are 3463 compositions in the test set, each with 130 (13 temperatures × 5 doping levels × 2 doping types) associated values. The inset plots depict the distribution of absolute errors.

band gaps. For example, when 10% noise has been added to the *ab initio* band gaps, the MAE when comparing these corrupted gaps to the true gaps is 0.065 eV. Figure 6 shows, as might be expected, that when more noise is added to the band gaps, the performance of the model falls. However, some thermoelectric transport properties are more robust (or more sensitive) to changes in the band gap quality. For example, in the case of the prediction of the Seebeck, even with band gaps exhibiting an MAE of 0.30 eV, the model is still able to achieve an R^2 of 0.85, in comparison to an R^2 of below 0.80 when no band gap is provided. However, in the case of $\log\sigma$, the model is much more sensitive. Current state-of-the-art band gap predictors that operate on composition alone typically achieve an MAE of 0.30–0.45 eV [122]. However, band gap predictor performance is expected to improve over time, and this will further increase the utility of the CraTENet model in screening scenarios with predicted band gaps.

3.2. Band gap prediction

A dataset consisting of compositions and their corresponding DFT-derived band gaps was formed by taking all of the unique compositions in the Materials Project, and consisted of 89 444 entries. A CrabNet model was trained on this dataset, using the minimization of the Robust L1 loss as the objective. To establish the

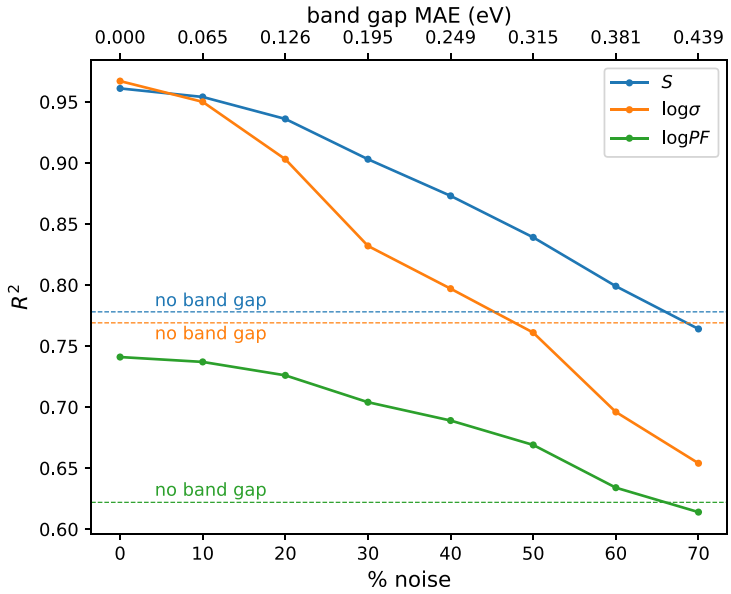

Figure 6. Performance of the CraTENet+gap model (in terms of R^2) as a function of band gap quality. A 90–10 holdout experiment was performed, and the actual gaps in the test set were corrupted by adding increasing amounts of Gaussian noise, before the performance of the model was assessed. The dotted lines represent the performance of the CraTENet model without a provided band gap.

generalization error of the model, ten-fold cross-validation was performed (as described in the Methods). Across the ten folds, the model achieved a mean R^2 of 0.71, and a mean MAE of 0.38 eV. A final model was trained on all 89 444 entries for 101 epochs, which was determined to be the ideal number of epochs required (i.e. the mean number of epochs required across the ten folds). This band gap predictor was subsequently used to provide band gaps when scanning composition space where structure and band gaps were unknown.

3.3. Searching composition space for new thermoelectrics

3.3.1. Materials project compounds not in the Ricci database

Of the 126 335 structures we obtained from the Materials Project, we derived 89 444 unique compositions. Since the compounds in the Ricci database originate from the Materials Project, we obtained 54 816 unique compositions when removing the compositions found in the Ricci database. This collection of 54 816 compositions forms a sizeable and convenient search space, since GGA band gaps have already been computed for these compounds, and their structures are known. Moreover, we verified that the distributions of compositions in this dataset and the Ricci database are similar (see supplementary figure 2). Thus, we apply our CraTENet+gap model to this space, in an attempt to surface novel compounds which may represent promising thermoelectrics. We verify the quality of our predictions by performing *ab initio* calculations for a small subset of these compounds.

Making predictions for tens of thousands of compounds with the CraTENet model is computationally inexpensive in comparison with *ab initio* calculations, since inference is fast, aided by the use of GPUs and the inherent parallelism in neural networks. After performing inference on this space, we selected 23 materials from this collection that spanned a range of different thermoelectric properties and band gaps. For example, the predicted Seebeck values ranged from -1200 to 1200 μV K^{-1}. When comparing the values produced using the CraTENet+gap model and those obtained through *ab initio* methods, we found that the R^2 was between 0.87 and 0.88, and the MAE was between 72 and 79 μV K^{-1} (figure 7 and supplementary figure 15). Although the agreement is generally good, there are some outliers (notably related to compositions SbTeIr and $LiNbN_2$). The performance of the model at specific compositions is difficult to rationalise, as it reflects both how well similar compositions are represented in the training set and the error related to the incompleteness of composition as a descriptor.

Moreover, we extracted the top 1000 compounds by predicted *PF*, for each of *p* and *n* doping types (the lists are provided in the dataset accompanying this article). We selected three *p*-type selenides for performing *ab initio* calculations: GaCuTeSe, InCuTeSe, and CeSbSe. These compounds do not appear to have been studied as thermoelectrics before, but they seem promising as they include elements like Cu, In, Sb, and Te that are present in well-known thermoelectrics. After carrying out *ab initio* calculations, we found generally

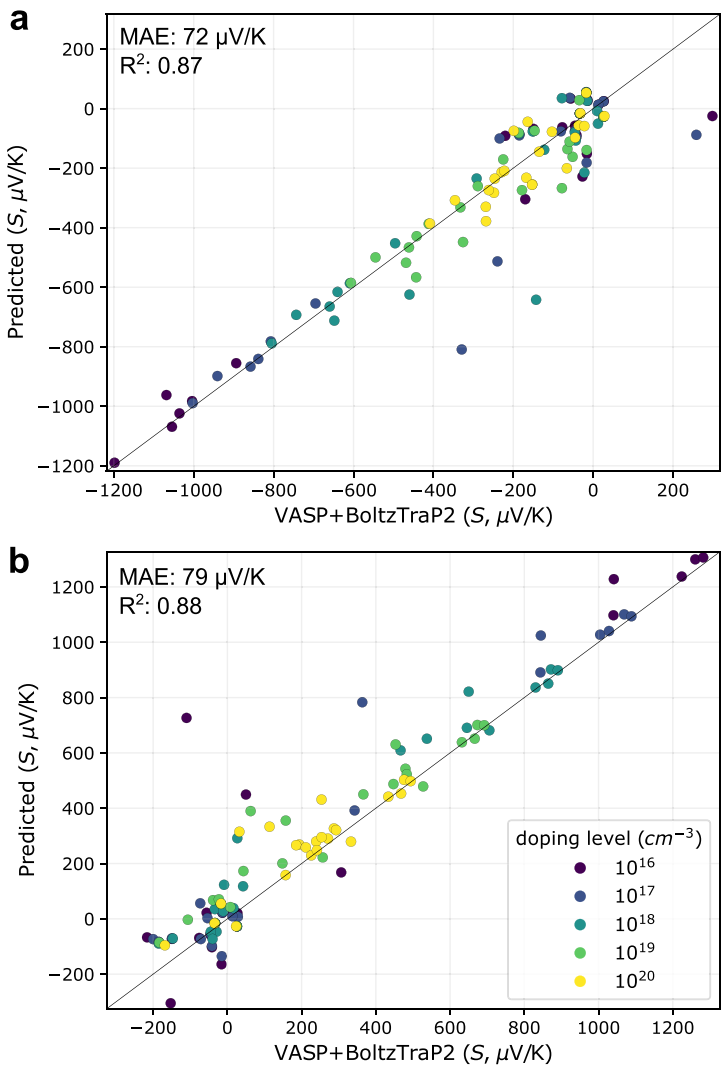

Figure 7. Seebeck coefficients at 700 K predicted with CraTENet+gap *vs.* those computed using the *ab initio* approach, for 23 Materials Projects compounds not found in the Ricci database, with (a) *n*-type doping, and (b) *p*-type doping. Each point represents a particular compound at a particular doping level (e.g. SbTeIr at 10^{20} cm^{-3}).

good agreement with the CraTENet predictions (figure 8; see supplementary figures 4–9 for more comprehensive plots of the predictions).

3.3.2. Hypothetical selenides

Since the CraTENet model requires only composition as input, it is conceivable that arbitrarily large hypothetical composition spaces could be generated and then processed by the model. SMACT is a software library that facilitates the generation of composition spaces, while adhering to chemical bonding rules, resulting in compositions which are chemically sensible [123]. Selenium-based materials are very promising thermoelectrics, because they exhibit similar properties as record-holding thermoelectric tellurides, but with the advantage that Se is much more Earth-abundant and cheaper than Te. We then chose to focus on creating a composition space of ternary selenides. Using SMACT, we generated 269 846 ternary selenide compositions, containing elements with an atomic number less than 84 (to avoid the heavy radioactive elements). The CraTENet and CraTENet+gap models were then used to make predictions of the thermoelectric transport properties of these compositions. As the CraTENet+gap model requires a band gap, we use our composition-only CrabNet band gap predictor as the source of the band gaps for this space. Since there is uncertainty in the band gap prediction, we make a separate prediction of thermoelectric transport properties using the predicted gap, the predicted gap plus the standard deviation, and the predicted gap minus the standard deviation. We find that this technique is useful for understanding the sensitivity of the predictions to the band gap value for a particular composition.

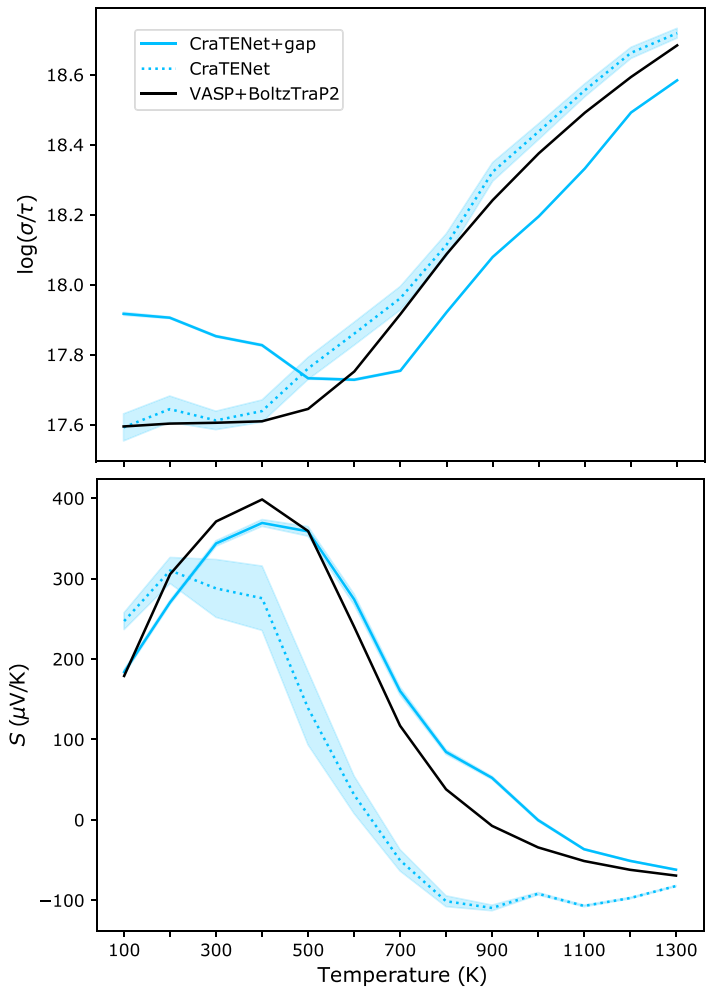

Figure 8. Predictions of the Seebeck and $\log\sigma$ for GaCuTeSe using the CraTENet models and the *ab initio* procedure, for *p*-type doping, at a level of 10^{19} cm^{-3}. The band gap value used, 0.387 eV, was obtained from the Materials Project. The shaded regions represent the $\pm$ standard deviation (i.e. the square root of the predicted variance).

Having made predictions on these SMACT-generated selenides, we then rank the compositions by *PF* (as described in the previous section). We make the top 1000 compositions publicly accessible in the code and dataset repository accompanying this article. There are several interesting selenides in that list, involving elements like bismuth (e.g. $LiBiSe_2$) or thallium (e.g. $NaTlSe_2$) which are often present in known thermoelectric materials. To the best of our knowledge, these compounds have not been studied as thermoelectrics in the literature. To validate the model's predictions, we carried out *ab initio* calculations on these two compounds, given that their structures are reported in the OQMD database [43]. A comparison of the predictions and the *ab initio* values for each is provided in figures 9(a) and (b). (See also supplementary figures 10–13 for more comprehensive plots of the predictions).

In the absence of DFT-calculated band gaps as input, the performance of the CraTENet model for these compounds is not as impressive in predicting the DFT-calculated values of the transport coefficients. The model using the predicted band gaps as an input seems to perform generally better than the model with no gap, but the deviations are still considerable, especially at high temperatures. All models, for example, overestimate the electrical conductivity of $LiBiSe_2$ by at least half an order of magnitude. Still, the DFT calculations confirm, within their own limitations, that these compounds have attractive values of the electronic transport coefficients; they deserve further investigation, either using more accurate theoretical predictions with methods beyond the GGA and the CRTA, or experimentally. Clearly, the main use of the methods presented here cannot be the quantitative prediction of the transport properties of individual compounds, but rather the identification of interesting candidates in unexplored regions of the compositional space.

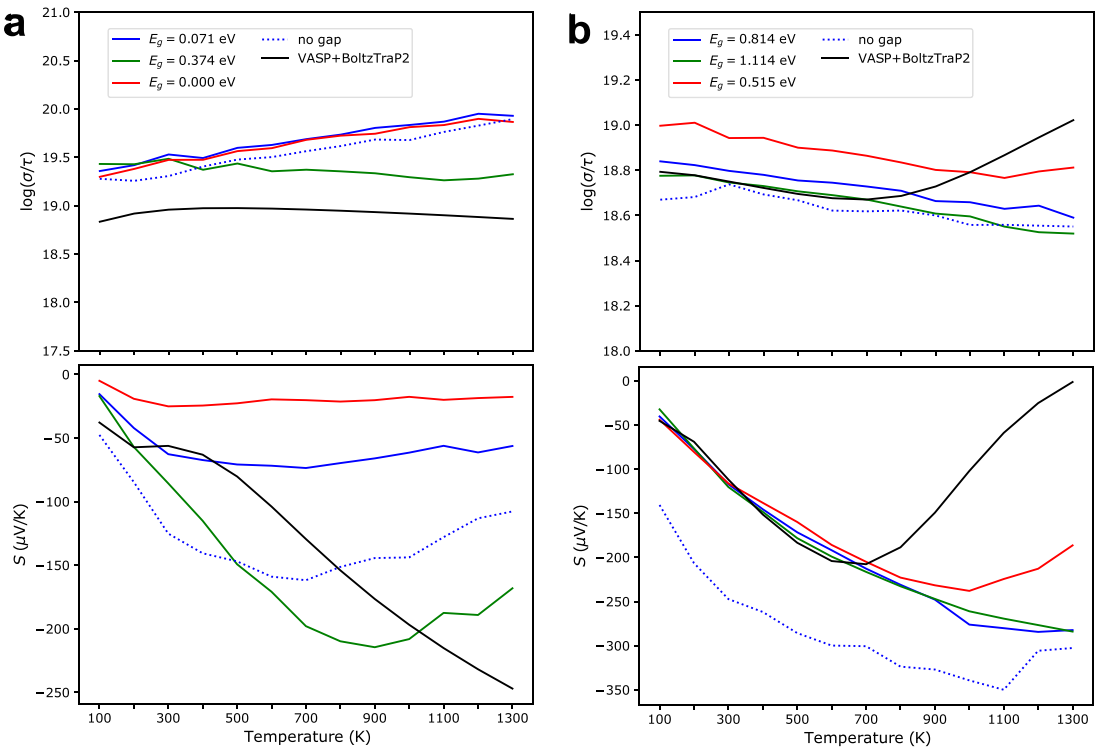

Figure 9. Predictions of the Seebeck and $\log\sigma$ for (a) $LiBiSe_2$, and (b) $NaTlSe_2$ using the CraTENet models and the *ab initio* procedure, for *n*-type doping, at a level of 10^{20} cm^{-3}. Predicted band gap values were used: blue represents the initial prediction, green represents the prediction plus the predicted standard deviation, and red represents the prediction minus the predicted standard deviation (i.e. square root of the predicted variance).

4. Conclusions

Approaches based on HTS combined with ML seem promising for suggesting novel candidate materials, since very large areas of chemical space can be examined quickly and efficiently. Here, we have shown that such an approach can be used to identify promising candidate thermoelectric materials based on the screening of potential compositions only, optionally supplemented with band gaps.

Several aspects of the approach described here contribute to its utility. First, the use of multi-output regression is helpful, and well-suited to the problem, since thermoelectric transport properties are dependent on factors such as temperature, doping level, and doping type. Conversely, an approach that requires parameters such as the temperature, doping level and doping type as input is problematic, since it increases the dimensionality of the input space, and also leads to inputs that resemble each other closely, as a result of the combinatorial nature of such a dataset [124].

Second, we believe that regression is a more useful choice for this learning task when compared to classification, in the context of searching for new materials. Several existing studies have involved the training of classification models of thermoelectric properties [53, 125]. These classification approaches involve predicting whether a thermoelectric property is in a desired range, or above (or below) a specified threshold. We argue that regression models, such as ours, provide a level of increased utility via their finer-grained predictions, which is critical when sifting through many thousands of potential candidates. A binary classifier simply provides no convenient means of differentiating the candidates labelled as promising. Although there is room for improvement in the quality of the predictions made by our regression models, we find that at the current performance level, the approach is effective at surfacing promising candidates.

Third, the use of an attention-based model, in combination with the Robust L2 loss, both leads to superior performance and provides unique advantages. The learned attention weights provide an opportunity to interpret the predictions made for a composition [126], and this could be a useful aspect of using the CraTENet model when analysing individual materials (rather than in bulk, as we have focused on here). Additionally, the Robust L2 loss is especially useful in that it allows the model to learn to quantify the uncertainty arising from mapping the composition (and optionally band gap) to thermoelectric properties. This provides users with a quantitative measure of the certainty of a prediction.

Future work will involve follow-up investigations of the candidate materials proposed here, using more rigorous *ab initio* methods. Should the candidates continue to appear promising, attempts may be made to synthesize the materials and measure their thermoelectric properties in the laboratory. In terms of the model itself, future work may involve augmenting the objective so that it takes into account the shape of the underlying manifold on which the multiple target values exist [127]. It is important to note that optimal thermoelectric transport properties are not the only criteria that establishes a material as a practical thermoelectric; other properties, such as dopability and stability, need to be considered. Thus, the computational discovery of novel thermoelectrics will be aided by the development of a suite of predictive models.

It is clear that the approach we describe depends heavily on the quality of the data it is trained on. The Ricci database was derived using theoretical constraints such as the CRTA for solving the Boltzmann transport equation, and the GGA for the exchange correlation functionals, which have important limitations. However, the approach we describe here can continue to be used with future databases of computed thermoelectric properties that will be obtained with more accurate theoretical methods, with improved data quality.

Finally, to demonstrate the predictions made by the CraTENet model, we have deployed an internet-accessible web browser-based application, located at https://thermopower.materialis.ai, that allows a user to submit a material's composition and (optionally) its band gap, and returns thermoelectric transport property predictions for the material, as made by the CraTENet model.

Data availability statement

The data that support the findings of this study are openly available. The Ricci database of thermoelectric transport coefficients is publicly available online at: https://datadryad.org/stash/dataset/doi:10.5061/dryad.gn001. The Materials Project data that was used to train the band gap predictor and form a composition search space are publicly available online at: https://materialsproject.org/. The pre-trained SkipAtom embeddings that were used as input to the neural network models are located at: https://github.com/lantunes/skipatom. The OQMD data that was used to provide structures for the SMACT-generated selenides are publicly available online at: https://oqmd.org/.

The data that support the findings of this study are openly available at the following URL: https://github.com/lantunes/CraTENet.

Acknowledgments

This work was partially supported by computational resource donations from Amazon Web Services through the AWS Activate program, obtained with assistance from the Communitech Hub. We are also grateful to the UK Materials and Molecular Modelling Hub for computational resources in the Young facility, which is partially funded by EPSRC (EP/P020194/1 and EP/T022213/1).

Code availability

The code with the CraTENet implementation, and for pre-processing the data and reproducing the experiments, is open source, released under the MIT License. The code repository is accessible online, at: https://github.com/lantunes/CraTENet.

Author contributions

R G-C conceived the project. L M A, K T B and R G-C designed the experiments. L M A conceived and implemented the model, and performed the experiments. L M A drafted the manuscript. R G-C and K T B supervised and guided the project. All authors reviewed, edited and approved the manuscript.

Conflict of interest

The authors declare no competing interests.

ORCID iDs

Luis M Antunes https://orcid.org/0000-0002-4867-5635
Keith T Butler https://orcid.org/0000-0001-5432-5597
Ricardo Grau-Crespo https://orcid.org/0000-0001-8845-1719

References

[1] Kajikawa T 2006 Thermoelectric power generation system recovering industrial waste heat *Thermoelectrics Handbook: Macro to Nano* ed D M Rowe (Boca Raton, FL: CRC Press) pp 50–51
[2] Snyder G J 2008 Small thermoelectric generators *Interface* **17** 54
[3] Seebeck T J 1826 Magnetische polarisation der metalle und erze durch temperatur-differenz *Ann. Phys., Lpz.* **82** 253–86
[4] Roget P M 1832 *Treatises on Electricity, Galvanism, Magnetism and Electro-Magnetism* (London: Baldwin and Cradock)
[5] Caballero-Calero O, Ares J R and Martín-González M 2021 Environmentally friendly thermoelectric materials: high performance from inorganic components with low toxicity and abundance in the earth *Adv. Sustain. Syst.* **5** 2100095
[6] Freer R and Powell A V 2020 Realising the potential of thermoelectric technology: a roadmap *J. Mater. Chem.* **8** 441–63
[7] Sootsman J R, Chung D Y and Kanatzidis M G 2009 New and old concepts in thermoelectric materials *Angew. Chem., Int. Ed.* **48** 8616–39
[8] Gayner C and Kar K K 2016 Recent advances in thermoelectric materials *Prog. Mater. Sci.* **83** 330–82
[9] Beretta D *et al* 2019 Thermoelectrics: from history, a window to the future *Mater. Sci. Eng.* R **138** 100501
[10] Poudel B *et al* 2008 High-thermoelectric performance of nanostructured bismuth antimony telluride bulk alloys *Science* **320** 634–8
[11] Tan G, Shi F, Hao S, Zhao L-D, Chi H, Zhang X, Uher C, Wolverton C, Dravid V P and Kanatzidis M G 2016 Non-equilibrium processing leads to record high thermoelectric figure of merit in PbTe–SrTe *Nat. Commun.* **7** 12167
[12] Wang X *et al* 2008 Enhanced thermoelectric figure of merit in nanostructured n-type silicon germanium bulk alloy *Appl. Phys. Lett.* **93** 193121
[13] Zhao L-D, Chang C, Tan G and Kanatzidis M G 2016 SnSe: a remarkable new thermoelectric material *Energy Environ. Sci.* **9** 3044–60
[14] Zhou M, Snyder G J, Li L and Zhao L-D 2016 Lead-free tin chalcogenide thermoelectric materials *Inorg. Chem. Front.* **3** 1449–63
[15] Liu H, Shi X, Xu F, Zhang L, Zhang W, Chen L, Li Q, Uher C, Day T and Snyder G J 2012 Copper ion liquid-like thermoelectrics *Nat. Mater.* **11** 422–5
[16] Caillat T, Fleurial J-P and Borshchevsky A 1996 Bridgman-solution crystal growth and characterization of the skutterudite compounds $CoSb_3$ and $RhSb_3$ *J. Cryst. Growth* **166** 722–6
[17] Gascoin F, Ottensmann S, Stark D, Haïle S M and Snyder G J 2005 Zintl phases as thermoelectric materials: tuned transport properties of the compounds $Ca_xYb_{1-x}Zn_2Sb_2$ *Adv. Funct. Mater.* **15** 1860–4
[18] Nolas G, Cohn J, Slack G and Schujman S 1998 Semiconducting Ge clathrates: promising candidates for thermoelectric applications *Appl. Phys. Lett.* **73** 178–80
[19] Aliev F, Brandt N B, Moshchalkov V V, Kozyrkov V V, Skolozdra R V and Belogorokhov A I 1989 Gap at the Fermi level in the intermetallic vacancy system RBiSn (R=Ti,Zr,Hf) *Z. Phys.* B **75** 167–71
[20] Aliev F, Kozyrkov V, Moshchalkov V, Scolozdra R and Durczewski K 1990 Narrow band in the intermetallic compounds MNiSn (M=Ti,Zr,Hf) *Z. Phys.* B **80** 353–7
[21] Hohl H, Ramirez A P, Kaefer W, Fess K, Thurner C, Kloc C and Bucher E 1997 A new class of materials with promising thermoelectric properties: MNiSn (M=Ti,Zr,Hf) *MRS Online Proc. Libr.* **478** 109–14
[22] Terasaki I, Sasago Y and Uchinokura K 1997 Large thermoelectric power in $NaCo_2O_4$ single crystals *Phys. Rev.* B **56** R12685(R)
[23] Tian R *et al* 2014 Enhancement of high temperature thermoelectric performance in Bi, Fe co-doped layered oxide-based material $Ca_3Co_4O_{9+\delta}$ *J. Alloys Compd.* **615** 311–5
[24] Zhou C *et al* 2021 Polycrystalline SnSe with a thermoelectric figure of merit greater than the single crystal *Nat. Mater.* **20** 1378–84
[25] Tritt T M and Subramanian M 2006 Thermoelectric materials, phenomena and applications: a bird's eye view *MRS Bull.* **31** 188–98
[26] Sparks T D, Gaultois M W, Oliynyk A, Brgoch J and Meredig B 2016 Data mining our way to the next generation of thermoelectrics *Scr. Mater.* **111** 10–15
[27] Gorai P, Stevanović V and Toberer E S 2017 Computationally guided discovery of thermoelectric materials *Nat. Rev. Mater.* **2** 1–16
[28] Recatala-Gomez J, Suwardi A, Nandhakumar I, Abutaha A and Hippalgaonkar K 2020 Toward accelerated thermoelectric materials and process discovery *ACS Appl. Energy Mater.* **3** 2240–57
[29] Madsen G K H 2006 Automated search for new thermoelectric materials: the case of LiZnSb *J. Am. Chem. Soc.* **128** 12140–6
[30] Wang S, Wang Z, Setyawan W, Mingo N and Curtarolo S 2011 Assessing the thermoelectric properties of sintered compounds via high-throughput *ab-initio* calculations *Phys. Rev.* X **1** 021012
[31] Carrete J, Mingo N, Wang S and Curtarolo S 2014 Nanograined half-Heusler semiconductors as advanced thermoelectrics: an *Ab Initio* high-throughput statistical study *Adv. Funct. Mater.* **24** 7427–32
[32] Toher C, Plata J J, Levy O, de Jong M, Asta M, Nardelli M B and Curtarolo S 2014 High-throughput computational screening of thermal conductivity, Debye temperature and Grüneisen parameter using a quasiharmonic Debye model *Phys. Rev.* B **90** 174107
[33] Gorai P, Parilla P, Toberer E S and Stevanovic V 2015 Computational exploration of the binary A_1B_1 chemical space for thermoelectric performance *Chem. Mater.* **27** 6213–21
[34] Zhu H *et al* 2015 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.* **3** 10554–65
[35] Xi L *et al* 2018 Discovery of high-performance thermoelectric chalcogenides through reliable high-throughput material screening *J. Am. Chem. Soc.* **140** 10785–93
[36] Gorai P, Ganose A, Faghaninia A, Jain A and Stevanović V 2020 Computational discovery of promising new n-type dopable ABX Zintl thermoelectric materials *Mater. Horiz.* **7** 1809–18
[37] Chen X, Zhang X, Gao J, Li Q, Shao Z, Lin H and Pan M 2021 Computational search for better thermoelectric performance in nickel-based half-Heusler compounds *ACS Omega* **6** 18269–80

[38] Pöhls J-H et al 2021 Experimental validation of high thermoelectric performance in $RECuZnP_2$ predicted by high-throughput DFT calculations *Mater. Horiz.* **8** 209–15
[39] Pizzi G, Cepellotti A, Sabatini R, Marzari N and Kozinsky B 2016 AiiDA: automated interactive infrastructure and database for computational science *Comput. Mater. Sci.* **111** 218–30
[40] Mathew K et al 2017 Atomate: a high-level interface to generate, execute and analyze computational materials science workflows *Comput. Mater. Sci.* **139** 140–52
[41] Jain A et al 2015 FireWorks: a dynamic workflow system designed for high-throughput applications *Concurr. Comput. Pract. Exp.* **27** 5037–59
[42] Curtarolo S et al 2012 AFLOW: an automatic framework for high-throughput materials discovery *Comput. Mater. Sci.* **58** 218–26
[43] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD) *JOM* **65** 1501–9
[44] Zapata F, Ridder L, Hidding J, Jacob C R, Infante I and Visscher L 2019 QMflows: a tool kit for interoperable parallel workflows in quantum chemistry *J. Chem. Inf. Model.* **59** 3191–7
[45] Adorf C S, Dodd P M, Ramasubramani V and Glotzer S C 2018 Simple data and workflow management with the signac framework *Comput. Mater. Sci.* **146** 220–9
[46] Mayeshiba T, Wu H, Angsten T, Kaczmarowski A, Song Z, Jenness G, Xie W and Morgan D 2017 The MAterials Simulation Toolkit (MAST) for atomistic modeling of defects and diffusion *Comput. Mater. Sci.* **126** 90–102
[47] Wang T, Zhang C, Snoussi H and Zhang G 2020 Machine learning approaches for thermoelectric materials research *Adv. Funct. Mater.* **30** 1906041
[48] Juneja R and Singh A K 2021 Accelerated discovery of thermoelectric materials using machine learning *Artificial Intelligence for Materials Science* (Berlin: Springer) pp 133–52
[49] Han G, Sun Y, Feng Y, Lin G and Lu N 2021 Machine learning regression guided thermoelectric materials discovery–a review *ES Mater. Manuf.* **14** 20–35
[50] Qian X and Yang R 2021 Machine learning for predicting thermal transport properties of solids *Mater. Sci. Eng.* R **146** 100642
[51] Antunes L M et al 2022 Machine learning approaches for accelerating the discovery of thermoelectric materials *Advancing Materials Innovation with Machine Learning* (*ACS Symp. Series* vol 1416) ed Y An (Washington, DC: American Chemical Society) ch 1, pp 1–32
[52] Furmanchuk A, Saal J E, Doak J W, Olson G B, Choudhary A and Agrawal A 2018 Prediction of Seebeck coefficient for compounds without restriction to fixed stoichiometry: a machine learning approach *J. Comput. Chem.* **39** 191–202
[53] Gaultois M W, Oliynyk A O, Mar A, Sparks T D, Mulholland G J and Meredig B 2016 Perspective: web-based machine learning models for real-time screening of thermoelectric materials properties *APL Mater.* **4** 053213
[54] Pimachev A K and Neogi S 2021 First-principles prediction of electronic transport in fabricated semiconductor heterostructures via physics-aware machine learning *npj Comput. Mater.* **7** 1–12
[55] Yuan H, Han S H, Hu R, Jiao W Y, Li M K, Liu H J and Fang Y 2022 Machine learning for accelerated prediction of the Seebeck coefficient at arbitrary carrier concentration *Mater. Today Phys.* **25** 100706
[56] Mukherjee M, Satsangi S and Singh A K 2020 A statistical approach for the rapid prediction of electron relaxation time using elemental representatives *Chem. Mater.* **32** 6507–14
[57] Yoshihama H and Kaneko H 2021 Design of thermoelectric materials with high electrical conductivity, high Seebeck coefficient and low thermal conductivity *Anal. Sci. Adv.* **2** 289–94
[58] Choudhary K, Garrity K F and Tavazza F 2020 Data-driven discovery of 3D and 2D thermoelectric materials *J. Phys.: Condens. Matter* **32** 475501
[59] Sheng Y, Wu Y, Yang J, Lu W, Villars P and Zhang W 2020 Active learning for the power factor prediction in diamond-like thermoelectric materials *npj Comput. Mater.* **6** 1–7
[60] Yang Z et al 2021 Accurate and explainable machine learning for the power factors of diamond-like thermoelectric materials *J. Materiomics* **8** 633–9
[61] Laugier L et al 2018 Predicting thermoelectric properties from crystal graphs and material descriptors—first application for functional materials (arXiv:1811.06219)
[62] Carrete J, Li W, Mingo N, Wang S and Curtarolo S 2014 Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling *Phys. Rev.* X **4** 011019
[63] Seko A, Togo A, Hayashi H, Tsuda K, Chaput L and Tanaka I 2015 Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization *Phys. Rev. Lett.* **115** 205901
[64] Zhang Y and Ling C 2018 A strategy to apply machine learning to small datasets in materials science *npj Comput. Mater.* **4** 1–8
[65] Chen L, Tran H, Batra R, Kim C and Ramprasad R 2019 Machine learning models for the lattice thermal conductivity prediction of inorganic materials *Comput. Mater. Sci.* **170** 109155
[66] Juneja R, Yumnam G, Satsangi S and Singh A K 2019 Coupling the high-throughput property map to machine learning for predicting lattice thermal conductivity *Chem. Mater.* **31** 5145–51
[67] Tewari A, Dixit S, Sahni N and Bordas Séphane P A 2020 Machine learning approaches to identify and design low thermal conductivity oxides for thermoelectric applications *Data-Centric Eng.* **1** E8
[68] Liu J, Han S, Cao G, Zhou Z, Sheng C and Liu H 2020 A high-throughput descriptor for prediction of lattice thermal conductivity of half-Heusler compounds *J. Phys. D: Appl. Phys.* **53** 315301
[69] Li R et al 2020 A deep neural network interatomic potential for studying thermal conductivity of β-Ga_2O_3 *Appl. Phys. Lett.* **117** 152102
[70] Loftis C, Yuan K, Zhao Y, Hu M and Hu J 2020 Lattice thermal conductivity prediction using symbolic regression and machine learning *J. Phys. Chem.* A **125** 435–50
[71] Miyazaki H, Tamura T, Mikami M, Watanabe K, Ide N, Ozkendir O M and Nishino Y 2021 Machine learning based prediction of lattice thermal conductivity for half-Heusler compounds using atomic information *Sci. Rep.* **11** 1–8
[72] Tranås R, Løvvik O M, Tomic O and Berland K 2022 Lattice thermal conductivity of half-Heuslers with density functional theory and machine learning: enhancing predictivity by active sampling with principal component analysis *Comput. Mater. Sci.* **202** 110938
[73] Jaafreh R, Kang Y S and Hamad K 2021 Lattice thermal conductivity: an accelerated discovery guided by machine learning *ACS Appl. Mater. Interfaces* **13** 57204–13

[74] Choi J M, Lee K, Kim S, Moon M, Jeong W and Han S 2022 Accelerated computation of lattice thermal conductivity using neural network interatomic potentials *Comput. Mater. Sci.* **211** 111472
[75] Tabib M V *et al* 2018 Discovering thermoelectric materials using machine learning: insights and challenges *Int. Conf. on Artificial Neural Networks* (Springer) pp 392–401
[76] Wang Z-L, Yokoyama Y, Onda T, Adachi Y and Chen Z-C 2019 Improved thermoelectric properties of hot-extruded Bi–Te–Se bulk materials with Cu doping and property predictions via machine learning *Adv. Electron. Mater.* **5** 1900079
[77] Na G S, Jang S and Chang H 2021 Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects *npj Comput. Mater.* **7** 1–11
[78] Zhong Y *et al* 2021 Data analytics accelerates the experimental discovery of new thermoelectric materials with extremely high figure of merit (arXiv:2104.08033)
[79] Gan Y, Wang G, Zhou J and Sun Z 2021 Prediction of thermoelectric performance for layered IV-V-VI semiconductors by high-throughput *ab initio* calculations and machine learning *npj Comput. Mater.* **7** 1–10
[80] Jaafreh R, Seong K Y, Kim J-G and Hamad K 2022 A deep learning perspective into the figure-of-merit of thermoelectric materials *Mater. Lett.* **319** 132299
[81] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
[82] Skansi S 2018 *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence* (Berlin: Springer)
[83] Agrawal A and Choudhary A 2019 Deep materials informatics: applications of deep learning in materials science *MRS Commun.* **9** 779–92
[84] Jha D, Ward L, Paul A, Liao W-K, Choudhary A, Wolverton C and Agrawal A 2018 ElemNet: deep learning the chemistry of materials from only elemental composition *Sci. Rep.* **8** 1–13
[85] Jha D *et al* 2019 IRNet: a general purpose deep residual regression framework for materials discovery *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 2385–93
[86] Jha D, Gupta V, Ward L, Yang Z, Wolverton C, Foster I, Liao W-K, Choudhary A and Agrawal A 2021 Enabling deeper learning on big data for materials informatics applications *Sci. Rep.* **11** 1–12
[87] Xie T and Grossman J C 2018 Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties *Phys. Rev. Lett.* **120** 145301
[88] Chen C, Ye W, Zuo Y, Zheng C and Ong S P 2019 Graph networks as a universal machine learning framework for molecules and crystals *Chem. Mater.* **31** 3564–72
[89] Goodall R E A and Lee A A 2020 Predicting materials properties without crystal structure: deep representation learning from stoichiometry *Nat. Commun.* **11** 1–9
[90] Wang A Y-T, Kauwe S K, Murdock R J and Sparks T D 2021 Compositionally restricted attention-based network for materials property predictions *npj Comput. Mater.* **7** 1–10
[91] Parr R G 1980 Density functional theory of atoms and molecules *Horizons of Quantum Chemistry* (Berlin: Springer) pp 5–15
[92] Ricci F, Chen W, Aydemir U, Snyder G J, Rignanese G-M, Jain A and Hautier G 2017 An *ab initio* electronic transport database for inorganic materials *Sci. Data* **4** 170085
[93] Madsen G K H and Singh D J 2006 BoltzTraP. a code for calculating band-structure dependent quantities *Comput. Phys. Commun.* **175** 67–71
[94] Jain A *et al* 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002
[95] Plata J J, Nath P, Sanz J F and Marquez A 2022 In silico modeling of inorganic thermoelectric materials *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* (Amsterdam: Elsevier)
[96] Shi H, Parker D, Du M-H and Singh D J 2015 Connecting thermoelectric performance and topological-insulator behavior: Bi_2Te_3 and Bi_2Te_2Se from first principles *Phys. Rev. Appl.* **3** 014004
[97] Freer R *et al* 2022 Key properties of inorganic thermoelectric materials–tables (version 1) *J. Phys. Energy* **4** 022002
[98] Borchani H, Varando G, Bielza C and Larranaga P 2015 A survey on multi-output regression *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery* **5** 216–33
[99] Xu D, Shi Y, Tsang I W, Ong Y-S, Gong C and Shen X 2019 Survey on Multi-Output Learning *IEEE Trans. Neural Netw. Learn. Syst.* **31** 2409–29
[100] Ho T K 1995 Random decision forests *Proc. 3rd Int. Conf. on Document Analysis and Recognition* vol 1 (IEEE) pp 278–82
[101] Vaswani A *et al* 2017 Attention is all you need *Advances in Neural Information Processing Systems* vol 30 (available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
[102] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8
[103] Ba J L, Kiros J R and Hinton G E 2016 Layer normalization (arXiv:1607.06450)
[104] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
[105] Levine S, Pastor P, Krizhevsky A, Ibarz J and Quillen D 2018 Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection *Int. J. Robot. Res.* **37** 421–36
[106] Nix D A and Weigend A S 1994 Estimating the mean and variance of the target probability distribution *Proc. 1994 IEEE Int. Conf. on Neural Networks (ICNN'94)* vol 1 (IEEE) pp 55–60
[107] Kendall A *et al* 2017 What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems* vol 30, ed I Guyon (Curran Associates, Inc.)
[108] Antunes L M, Grau-Crespo R and Butler K T 2022 Distributed representations of atoms and materials for machine learning *npj Comput. Mater.* **8** 44
[109] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)
[110] Abadi M *et al* 2016 Tensorflow: large-scale machine learning on heterogeneous distributed systems (arXiv:1603.04467)
[111] Chollet F *et al* 2015 Keras (available at: https://github.com/fchollet/keras)
[112] Meredig B, Agrawal A, Kirklin S, Saal J E, Doak J W, Thompson A, Zhang K, Choudhary A and Wolverton C 2014 Combinatorial screening for new materials in unconstrained composition space with machine learning *Phys. Rev.* B **89** 094104
[113] Ward L *et al* 2018 Matminer: an open source toolkit for materials data mining *Comput. Mater. Sci.* **152** 60–69
[114] Pedregosa F *et al* 2011 Scikit-learn: machine learning in python *J. Mach. Learn. Res.* **12** 2825–30
[115] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press) ch 7, pp 245–6
[116] Kresse G and Hafner J 1993 *Ab initio* molecular dynamics for liquid metals *Phys. Rev.* B **47** 558

[117] Kresse G and Furthmüller J 1996 Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set *Phys. Rev.* B **54** 11169
[118] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865
[119] Blöchl P E 1994 Projector augmented-wave method *Phys. Rev.* B **50** 17953
[120] Kresse G and Joubert D 1999 From ultrasoft pseudopotentials to the projector augmented-wave method *Phys. Rev.* B **59** 1758
[121] Madsen G K H, Carrete J and Verstraete M J 2018 BoltzTraP2, a program for interpolating band structures and calculating semi-classical transport coefficients *Comput. Phys. Commun.* **231** 140–5
[122] Wu L, Xiao Y, Ghosh M, Zhou Q and Hao Q 2020 Machine learning prediction for bandgaps of inorganic materials *ES Mater. Manuf.* **9** 34–9
[123] Davies D W, Butler K, Jackson A, Skelton J, Morita K and Walsh A 2019 SMACT: semiconducting materials by analogy and chemical theory *J. Open Source Softw.* **4** 1361
[124] Zahrt A F, Henle J J and Denmark S E 2020 Cautionary guidelines for machine learning studies with combinatorial datasets *ACS Comb. Sci.* **22** 586–91
[125] Lu N, Han G, Sun Y, Feng Y and Lin G 2022 Artificial intelligence assisted thermoelectric materials design and discovery *Research Square Preprint* https://doi.org/10.21203/rs.3.rs-1898309/v1 (posted online 4 August 2022, accessed 1 December 2022)
[126] Wang A Y-T, Mahmoud M S, Czasny M and Gurlo A 2022 CrabNet for explainable deep learning in materials science: bridging the gap between academia and industry *Integr. Mater. Manuf. Innov.* **11** 41–56
[127] Liu G, Lin Z and Yu Y 2009 Multi-output regression on the output manifold *Pattern Recognit.* **42** 2737–43