# Compositionally-Restricted Attention-Based Network for Materials Property Predictions

Anthony Yu-Tung Wang,[†,¶] Steven K. Kauwe,[‡,¶] Ryan J. Murdock,[‡] and Taylor D. Sparks[*,‡]

†*Technische Universität Berlin, Fachgebiet Keramische Werkstoffe / Chair of Advanced Ceramic Materials, 10623 Berlin, Germany*

‡*Department of Materials Science & Engineering, University of Utah, Salt Lake City, UT, 84112, USA*

¶*Contributed equally to this work*

E-mail: sparks@eng.utah.edu

**Abstract**

In this paper, we demonstrate a novel application of the Transformer self-attention mechanism. Our network, the Compositionally-Restricted Attention-Based network, referred to as `CrabNet`, explores the area of structure-agnostic materials property predictions when only a chemical formula is provided. Our results show that `CrabNet`'s performance matches or exceeds current best practice methods on nearly all of 28 total benchmark datasets. We also demonstrate how `CrabNet`'s architecture lends itself towards model interpretability by showing different visualization approaches that are made possible by `CrabNet`'s design. We feel confident that `CrabNet`, and its attention-based framework, will be of keen interest to future materials informatics researchers.

# Keywords

machine learning, materials informatics, attention, self-attention, transformers, materials discovery, material screening, high-throughput screening, regression, interpretability


# Introduction

Materials scientists constantly strive to have better understanding, and therefore predictions, of materials properties. This began with the collection of empirical evidence through repeated experimentation, resulting in mathematical generalizations, theories, and laws. More recently, computational methods have arisen to solve a large variety of problems that were intractable to analytical approaches alone.[1,2]

As experimental and computational methods have become more efficient, high-quality data has opened up a new avenue to materials understanding. Materials informatics (MI) is the resulting field of research which utilizes statistical and machine learning (ML) approaches in combination with high-throughput computation to analyze the wealth of existing materials information and gain unique insights.[2–4] As this wealth has increased, practitioners of MI have increasingly turned to deep learning techniques to model and represent inorganic chemistry, resulting in approaches such as `ElemNet`, `IRNet`, `CGCNN`, `SchNet` and `Roost`.[5–9] In specific cases,[7,8,10–15] including `CGCNN` and `SchNet`, the compounds are represented using their chemical and structural information.

Modeling approaches based on crystal structure are an excellent tool for MI. Unfortunately, there are many material property datasets that lack suitable structural information. An example of this is the experimental band gap data gathered by Zhou *et al.*[16] Conversely, many databases such as the Inorganic Crystal Structure Database (ICSD) and Pearson's Crystal Data (PCD) contain an abundance of structural information, but lack the associated material properties of the recorded structures. In both cases, the applicability of structure-based learning approaches are limited. This limitation is particularly evident in the discovery of novel materials, since it is not possible to know the structural information of (currently undiscovered) chemical compounds *a priori*. Therefore, the development of structure-agnostic techniques is well-suited to the discovery of novel materials.

A typical approach to structure-agnostic learning is done by representing chemistry as a composition-based feature vector (`CBFV`).[17] This allows for data-driven learning in the absence of structural information. The `CBFV` is a common way to transform chemical compositions into usable features for ML and is generated from the descriptive statistics of a compound's constituent element properties. Researchers have effectively used `CBFV`-based ML techniques to generate materials property predictions.[17–25]

One potential issue with the `CBFV` approach lies in the way the element vectors are combined to form the vector describing the chemical compound. Typically, the individual element vectors of the compound are scaled by the element's prevalence (fractional abundance) in the composition, before being used to form the `CBFV`. This step assumes that the stoichiometric prevalence of constituent elements in a compound dictate their chemical signal, or contribution, to the material's property. However, this is not true in all cases; an extreme example of this is element doping. Dopants can be present in very small amounts in a compound, but can have a significant impact on its electrical,[23,26,27] mechanical,[20,28–30] and thermal properties.[31–34] In the case of a typical `CBFV` approach which uses the weighted average of element properties as a feature, the signal from dopant elements would not significantly change the vector representation of a compound. As a result, the trained ML model would fail to capture a portion of the relevant chemical information available in the full composition.

It is apparent that there is no generally-accepted best way to model materials property behaviors. Different ML approaches lend themselves towards different modeling tasks. `CGCNN` requires access to structural information, `ElemNet` operates within realm of large data, and classical models excel when domain knowledge can be exploited to overcome data scarcity.[35] To address the diversity of learning challenges, in Dunn *et al.*, the Automatminer framework uses computationally-expensive searches to optimize classical modeling techniques. They demonstrate effective learning on some data, but show shortcomings when deep-learning is appropriate.[36]

In a similar spirit, we seek to overcome general challenges in the area of structure-agnostic learning using an approach we refer to as the Compositionally-Restricted Attention-Based network (`CrabNet`). `CrabNet` introduces the self-attention mechanism to the task of materials property predictions, and dynamically learns and updates individual element representations based on their chemical environment. To enable this, we use a unique featurization scheme that represents and preserves individual element identity while sharing information between elements. Self-attention is a procedure by which a neural network learns representations for each item in a system based on the other items that are present. In this context, we treat the chemical composition as the system and the elements as the items within that system. This representation enables `CrabNet` to learn inter-element interactions within a compound and use these interactions to generate property predictions.

To perform self-attention, we use the Transformer architecture, which emerged from natural language processing (NLP) and is based on stacked self-attention layers.[37–44] A typical use of the Transformer architecture in NLP is to encode the meaning of a word given the surrounding words, sentences, and paragraphs. Beyond NLP, other example uses of the Transformer architecture are found in music generation,[45] image generation,[46] image and video restoration,[47–51] game playing agents,[52,53] and drug discovery.[54,55] In this work, we explore how our attention-based architecture, `CrabNet`, performs in predicting materials properties relative to the common modeling techniques `Roost`, `ElemNet`, and random forest (`RF`) for regression-type problems.

# Results and Discussions

The results of this study are described in three subsections. First, we describe the collection of materials property data used for benchmarking `CrabNet`. Second, we highlight the performance of `CrabNet` when compared to other current learning approaches which rely solely on composition. Third, we briefly outline how the self-attention mechanism in `CrabNet` enables visualizations and inspectability unique to Transformer-based modeling.

## Data and Materials Properties Procurement

For this work, we obtained both computational and experimental materials data for benchmarking. Our benchmark data includes materials properties from the Matbench dataset as provided by Dunn *et al.*[36] In addition, materials properties data from a number of works[6,56–59] are collected, which are referred to as the "Extended dataset". We included 28 benchmark datasets in total: 10 from the Matbench and 18 from the Extended datasets ranging from 312 to 341,788 instances of data.

The Matbench datasets were split using five-fold cross-validation following instructions provided in the original publication.[36] Materials properties in the Extended dataset were split into train, validation, and test datasets. The full benchmark dataset, comprising the Matbench and Extended datasets, were then used with `Roost`, `CrabNet`, `ElemNet`, and `RF` models. The training and validation data were used for hyperparameter tuning. The test data were held-out to provide a fair evaluation of performance metrics across all models. A summary of the datasets is shown in Table 1.

All datasets are provided as pre-split csv files to facilitate future comparison to the metrics reported in this paper. Additional data processing and cleaning details can also be seen in the code on the dataset repository "mse_datasets".[60] To maintain consistent and simple benchmark comparisons, we selected data suitable for regression tasks and ignored structural information when present.

Table 1: List of all material properties used to benchmark the ML models in this work, together with the dataset size and the training, validation, and test set proportions. The materials properties listed in the top and bottom halves are Matbench and Extended datasets, respectively.

| Dataset name | Source | Material property | # samples | (train/val/test) % |
|---|---|---|---|---|
| `castelli` | Castelli *et al.*[36,61] | formation enthalpy (perovskites) | 18928 | 5-fold (72/8/20) |
| `dielectric` | MP[36,62–64] | refractive index | 4764 | 5-fold (72/8/20) |
| `elasticity_log10(G_VRH)` | MP[36,62,63,65] | $\log_{10}$(shear modulus (VRH) | 10987 | 5-fold (72/8/20) |
| `elasticity_log10(K_VRH)` | MP[36,62,63,65] | $\log_{10}$(bulk modulus (VRH) | 10987 | 5-fold (72/8/20) |
| `expt_gap` | Experiment[16,36] | experimental band gap | 4764 | 5-fold (72/8/20) |
| `jdft2d` | Experiment[36,66] | exfoliation energy | 636 | 5-fold (72/8/20) |
| `mp_e_form` | MP[36,62,63] | formation energy per atom | 132752 | 5-fold (72/8/20) |
| `mp_gap` | MP[36,62,63] | band gap | 106113 | 5-fold (72/8/20) |
| `phonons` | MP[36,62,63,67] | phonon frequency | 1265 | 5-fold (72/8/20) |
| `steels_yield` | MP[36,68] | steels yield strength | 312 | 5-fold (72/8/20) |
| `aflow__ael_bulk_modulus_vrh` | AFLOW[56] | bulk modulus (VRH) | 4905 | (70/15/15) |
| `aflow__ael_debye_temperature` | AFLOW[56] | Debye temperature | 4905 | (70/15/15) |
| `aflow__ael_shear_modulus_vrh` | AFLOW[56] | shear modulus (VRH) | 4905 | (70/15/15) |
| `aflow__agl_thermal_conductivity_300K` | AFLOW[56] | thermal conductivity | 4896 | (70/15/15) |
| `aflow__agl_thermal_expansion_300K` | AFLOW[56] | thermal expansion | 4895 | (70/15/15) |
| `aflow__Egap` | AFLOW[56] | band gap | 27841 | (70/15/15) |
| `aflow__energy_atom` | AFLOW[56] | energy per atom | 27844 | (70/15/15) |
| `CritExam__Ed` | Bartel *et al.*[57] | decomposition enthalpy | 85014 | (70/15/15) |
| `CritExam__Ef` | Bartel *et al.*[57] | formation enthalpy | 85014 | (70/15/15) |
| `mp_bulk_modulus` | MP (Oct. 2018)[59,62,63,65] | bulk modulus | 7632 | (70/15/15) |
| `mp_elastic_anisotropy` | MP (Oct. 2018)[59] | ratio of elastic anisotropy | 7659 | (70/15/15) |
| `mp_e_hull` | MP (Oct. 2018)[59] | energy above the convex hull | 83983 | (70/15/15) |
| `mp_mu_b` | MP (Oct. 2018)[59] | magnetization of the unit cell | 83973 | (70/15/15) |
| `mp_shear_modulus` | MP (Oct. 2018)[59,62,63,65] | shear modulus | 7437 | (70/15/15) |
| `OQMD_Bandgap` | OQMD[6] | band gap | 341696 | (70/15/15) |
| `OQMD_Energy_per_atom` | OQMD[6] | energy per atom | 341788 | (70/15/15) |
| `OQMD_Formation_Enthalpy` | OQMD[6] | formation enthalpy | 341788 | (70/15/15) |
| `OQMD_Volume_per_atom` | OQMD[6] | volume per atom | 341788 | (70/15/15) |

## Benchmark Comparisons

With the benchmark data described above, we generated materials predictions using the publicly-available code repositories for `Roost`,[9] `CrabNet`,[69] and `ElemNet`.[5]

The performance of these benchmarked models is compared using the mean absolute error between true values ($y$) and predicted values ($\hat{y}$), defined by $\mathrm{MAE} = \sum_{i=1}^{n} |y - \hat{y}|$. This allows for consistent comparison to past works.[5–7,9]

Table 2 shows the performance metrics from training and testing the models on all the benchmark materials properties outlined above. Here we note that the models for `Roost`, `CrabNet`, and `ElemNet` were all trained using the default model parameters provided with their respective repositories. In contrast to `Roost` and `ElemNet`, the default parameters for `CrabNet` were optimized using validation data from some of the same datasets on which we benchmarked. Although it is possible this offers a small advantage to `CrabNet`'s performance, we do not expect this to be significant due to `CrabNet`'s consistently strong performance on all benchmark tasks.

We tested two versions of `CrabNet`. The default `CrabNet` uses a `mat2vec` embedding when representing elements, similar to `Roost`. The second version of `CrabNet` (`HotCrab`) uses one-hot encodings (in the form of atomic numbers) and fractional amounts to represent each element in a composition. This is similar to `ElemNet`, as both models start without any chemical information. The random forest (`RF`) model utilizes a `Magpie`-featurized `CBFV` to represent chemistry. This is included as a performance baseline to match similar works.[5,9,36]

Overall, we see similar performance between `Roost` and the two versions of `CrabNet` tested. Given the different architectures and modelling philosophies of `Roost` and `CrabNet`, it is promising that both approaches converge towards the same performance metrics. We also see that `Roost`, and both `CrabNet` versions, achieve consistent and significant improvements to MAE compared to `ElemNet` and `RF` approaches. Interestingly, Table 2 shows that the use of `mat2vec` instead of `onehot` with `CrabNet` improves prediction performance on all materials properties except those present in the largest dataset (OQMD).

Table 2: MAE scores of `Roost`, `CrabNet`, one-hot encoded `CrabNet` (`HotCrab`), and `ElemNet` on the test dataset, compared with the random forest (`RF`) baseline. Cells are colored according to relative MAE performance within each row (blue is better, and red is worse). A `NaN` (not a number) value is reported for instances where the models failed to converge on a given material property. Here we present model results trained using chemical information (`Roost`, `CrabNet`), no chemical information (`HotCrab`, `ElemNet`), and a standard `CBFV` (`RF`).

| MatBench Properties | Roost | CrabNet | HotCrab | ElemNet | RF |
|---|---|---|---|---|---|
| Castelli perovskites | 0.359 | 0.408 | 0.410 | 0.468 | 0.581 |
| Refractive index | 0.327 | 0.309 | 0.323 | 0.538 | 0.419 |
| Shear modulus (log10) | 0.104 | 0.097 | 0.102 | 0.140 | 0.105 |
| Bulk modulus (log10) | 0.078 | 0.073 | 0.077 | 0.124 | 0.082 |
| Experimental Band gap | 0.374 | 0.339 | 0.353 | 0.450 | 0.444 |
| DFT Exfoliation energy | 47.692 | 45.617 | 48.943 | 57.673 | 50.000 |
| MP Formation energy | 0.082 | 0.083 | 0.085 | 1.029 | 0.121 |
| MP Band gap | 0.252 | 0.258 | 0.265 | 0.337 | 0.328 |
| Phonon peak | 46.158 | 49.868 | 56.258 | nan | 64.924 |
| Steels yield | 155.231 | 91.783 | 92.507 | nan | 104.538 |
| **Extended Properties** | **Roost** | **CrabNet** | **HotCrab** | **ElemNet** | **RF** |
| AFLOW Bulk modulus | 8.988 | 8.861 | 9.271 | 12.242 | 12.006 |
| AFLOW Debye temperature | 37.478 | 33.717 | 36.027 | 47.937 | 36.352 |
| AFLOW Shear modulus | 10.063 | 9.101 | 9.510 | 12.890 | 10.125 |
| AFLOW Thermal conductivity | 2.697 | 2.316 | 2.251 | 3.383 | 2.724 |
| AFLOW Thermal expansion | 3.68e-06 | 3.84e-06 | 3.87e-06 | 6.36e-06 | 5.54e-06 |
| AFLOW Band gap | 0.337 | 0.302 | 0.316 | 0.375 | 0.384 |
| AFLOW Energy per atom | 0.086 | 0.093 | 0.094 | 0.129 | 0.230 |
| Bartel Decomposition (Ed) | 0.067 | 0.063 | 0.066 | 0.081 | 0.090 |
| Bartel Formation (Ed) | 0.055 | 0.059 | 0.059 | 0.071 | 0.100 |
| MP Bulk modulus | 14.950 | 14.264 | 14.982 | 17.236 | 16.754 |
| MP Elastic anisotropy | 9.273 | 9.233 | 9.533 | 9.500 | 11.239 |
| MP Energy above convex hull | 0.089 | 0.086 | 0.089 | 0.100 | 0.113 |
| MP Magnetic moment | 2.583 | 2.395 | 2.433 | 2.761 | 2.664 |
| MP Shear modulus | 12.674 | 12.363 | 13.068 | 15.374 | 13.368 |
| OQMD Band gap | nan | 0.039 | 0.038 | 0.119 | 0.049 |
| OQMD Energy per atom | 0.035 | 0.035 | 0.035 | 0.065 | 0.145 |
| OQMD Formation Enthalpy | 0.033 | 0.033 | 0.032 | 0.062 | 0.085 |
| OQMD Volume per atom | 0.265 | 0.247 | 0.247 | nan | 0.529 |

The Matbench data provided by Dunn *et al.*[36] was benchmarked using the Automatminer tool. Their metrics are not included in Table 2, since all but two (expt_gap, and steels_yield) of Automatminer's models use structural information. Consequently, we focus on these two materials properties when comparing `CrabNet`'s results to those from Automatminer. For these two metrics, CrabNet's structure-agnostic approach outperforms the reported MAE values from Automatminer on the same tasks (expt_gap: 0.416 eV vs. 0.339 eV for CrabNet; steels_yield: 95.2 GPa vs. 91.8 GPa for CrabNet).

The performance of CrabNet on the steels_yield task is particularly interesting. The steels_yield dataset contains compounds with small dopant amounts in large chemical systems (up to 13 elements per composition) and only 312 total data. CrabNet's ability to learn on this data-poor property, and outperform the baseline RF method (which is traditionally better in the data-poor regime), is encouraging. We expected the steels_yield task to be difficult for all deep learning approaches. Nevertheless, repeated training and validation of CrabNet consistently produced error metrics equivalent, or better, than the best result obtained by Automatminer (95.2 GPa).

## Visualizing Self-Attention

CrabNet's modeling and visualization capabilities are enabled by its attention-based learning framework. In statistical machine learning (and many deep learning approaches akin to ElemNet) the chemical composition of a compound is represented as a single CBFV. In contrast, Roost and CrabNet represent a composition as a set of element vectors. Distinct to CrabNet, however, is the Transformer-based attention which learns to update these element vectors using learned attention matrices. In Figure 1, we show example attention matrices for each head of a CrabNet model trained on the property mp_bulk_modulus, using $Al_2O_3$ as the example composition. These matrices contain the information regarding how each element (rows) is influenced by all elements in the system (columns). The values in these attention matrices are used in the Transformer encoder to update the element vectors (see Methods for details). A value of zero means that the element in the column is completely ignored when updating the element in that row. A value of one means that the entire vector update is based solely on that column's element. Our implementation of CrabNet has three layers, each with four attention heads, with each head using the same data to generate its own independent attention matrix.

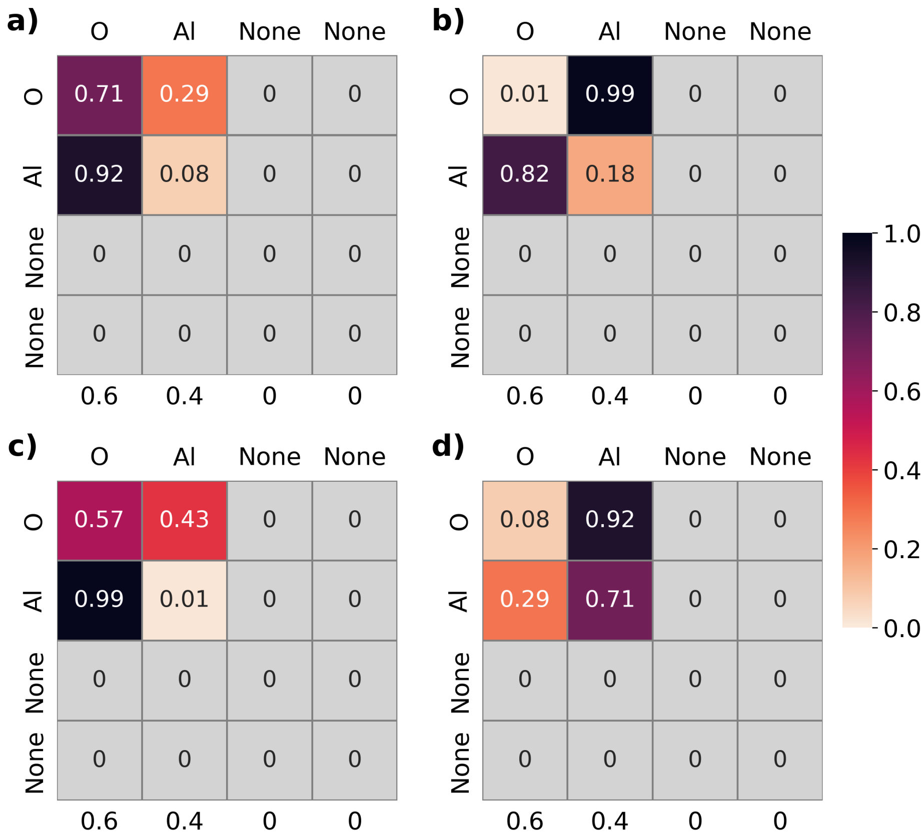


Figure 1: Displayed are the four attention heads (a, b, c, and d) from the first layer of a `CrabNet` model trained on mp_bulk_modulus. Each row represents an element in the system. Each column represents an element being attended to. Each element's fractional amount is shown on the x-axis. The values in the attention matrix are scores representing element-element interactions for the compound. As an example, in head a), $Al_{0.4}$ is attending strongly (with an attention score of 0.92) to $O_{0.6}$.

Shifting our focus to another `CrabNet` model trained on OQMD_bandgap data, we show that in addition to visualization of the individual attention heads, we can also generate a global view of attention from the perspective of individual elements. In Figure 2, we use four periodic tables to visualize, for each attention head, the average attention that silicon dedicates to elements when they are in the same composition. The darker colored elements can be understood as highly influential when updating silicon's vector representation.

Interestingly, each attention head has its own behavior, with some focusing on familiar groups and columns in the periodic table. This behavior lends credibility to `CrabNet` since

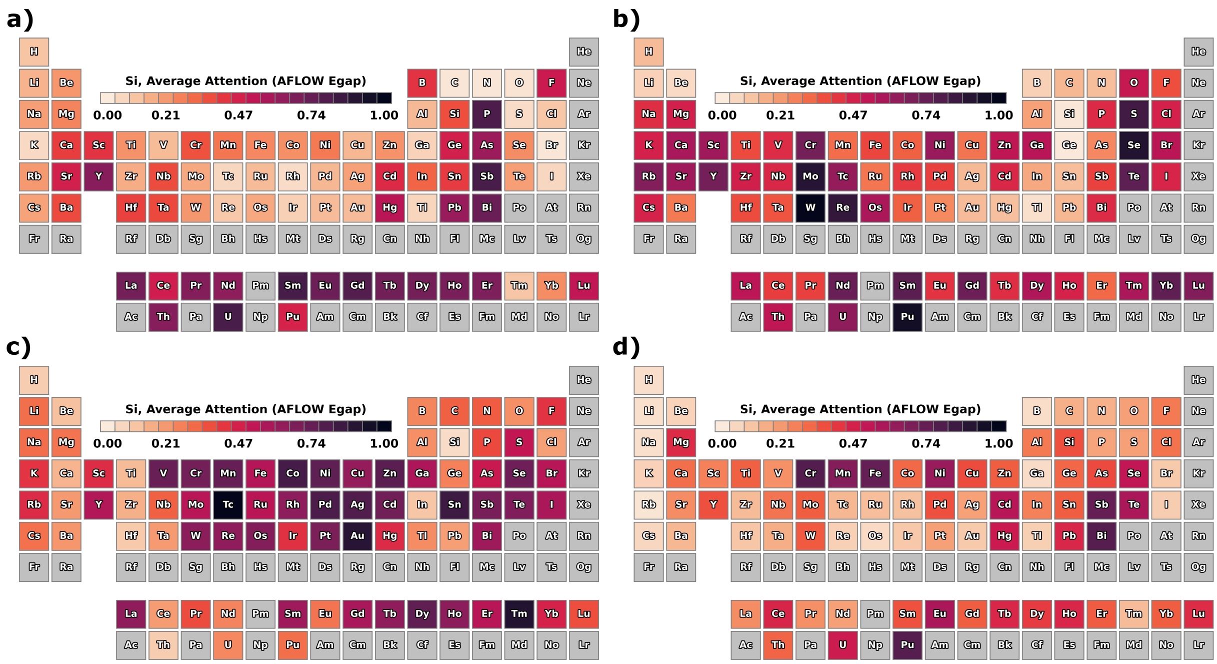


Figure 2: The average attention from each of the four attention heads (a, b, c, d) of a `CrabNet` model trained on the OQMD_bandgap data is shown for systems containing Si. The heatmap shows the average amount of attention that Si dedicates to the other elements in silicon-containing compounds. The darker the coloring, the more strongly Si attends to that element. We can see that each attention head exhibits its own behavior, and attends to different groups of elements. Interestingly, head a) attends to common *n*-type dopants and head c) attends to many transition metals, whereas heads b) and d) have unfamiliar element groupings.

there is no inherent reason that data-driven learning should converge to chemical rules that are familiar to materials scientists.

The preservation of elemental identity within a compound—as a result of the self-attention mechanism—also enables `CrabNet` to generate property predictions in a way that is unique to other approaches. Typically, one would collapse element information into a single vector, and use that to generate the property prediction. In contrast, `CrabNet` uses each element's vector (resulting from the attention process) to directly predict the element's contribution to the property. Figure 3a shows the average contributions from each element for a `CrabNet` model trained on AFLOW_bulk_modulus data. The darker colored elements contribute more towards a compound's bulk modulus value. Alternatively, elements can be visualized individually using their specific per-element contributions. In 3b we show distribution plots for lithium and tungsten's contributions to bulk modulus. From these plots, we can see that `CrabNet` expects lithium to contribute little to overall bulk modulus, whereas it expects tungsten to contribute largely. The visualizations from Figure 3 match closely—and reinforce—expectations regarding which elements most influence bulk modulus. Exploration of data in this manner hints at first steps towards model interpretability of `CrabNet`. We expect these types of property visualizations to be useful for exploring and verifying model and element behavior in detail.

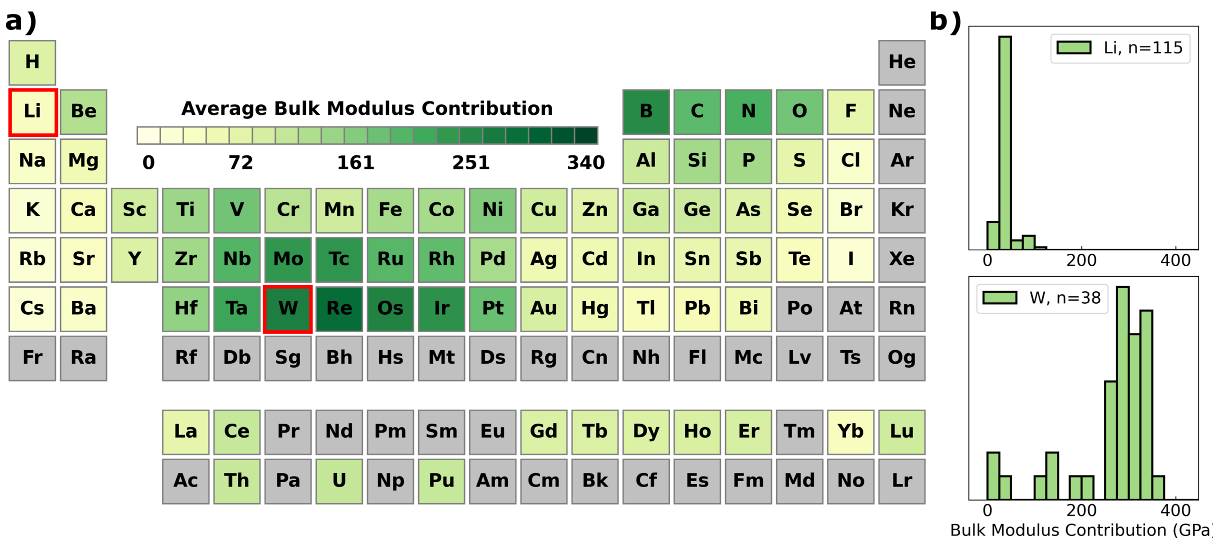


Figure 3: Average contribution of all elements to Bulk Modulus predictions, computed from the AFLOW_bulk_modulus dataset. (a) Plotted on a periodic table. (b) Plotted on histograms showing the per-element contribution amounts of Li and W, respectively. The darker colored elements in the periodic table contribute more towards a compound's bulk modulus value.

Finally, with per-element contributions in mind, we can demonstrate changes to `CrabNet`'s expected material property behavior as a function of composition. To do this, we consider a normalized chemical system consisting of atoms A and B, in the form of $A_xB_{1-x}$. We then generate property predictions for all $x \in \{0.0, 0.02, ..., 1.0\}$. In Figure 4, we visualize the behavior of $Si_xO_{1-x}$ when predicting band gap with a model trained on the OQMD_Bandgap data.

We first observe that the element contributions for both oxygen and silicon are similar

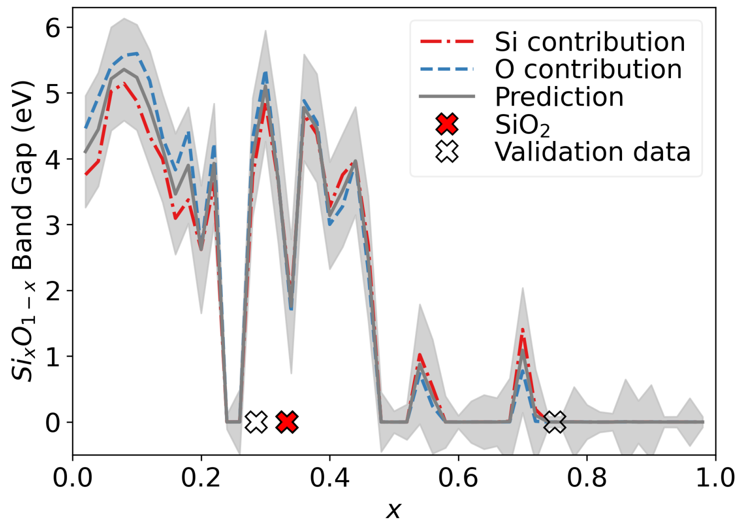


Figure 4: Model predictions over the $Si_xO_{1-x}$ system. The x axis is the fractional amount of silicon. The y axis shows the predicted band gap value at a given composition. The blue and red lines are the individual element contributions to the prediction. The gray shading represents the aleatoric uncertainty for each prediction. Data from the validation set are shown as X's. We highlight, in red, the incorrectly reported ground truth value of $SiO_2$ from the OQMD dataset.

throughout the varied stoichiometry range. Second, we identify that the ground truth labels for some $Si_xO_{1-x}$ compositions in the OQMD_Bandgap dataset do not match `CrabNet`'s predictions. While $SiO_2$ (normalized to $Si_{0.33}O_{0.66}$) is incorrectly labeled as a metal in the OQMD dataset (with a band gap of 0 eV), `CrabNet` returns a non-zero band gap prediction of 2.38 eV. Checking to see if every target is correctly labeled is impracticable given the size of the dataset. The identification of these discrepancies was made possible through our visualization of data. This alone is supportive of the visualization of data and the pursuit of inspectable models.

# Methods

## Self-attention and the CrabNet Architecture

### Representing Composition

Chemical compositions are input using the atomic numbers and fractional amounts of their constituent elements. The atomic numbers are used to retrieve element representations (either `mat2vec` or `onehot`). The fractional amounts are used to obtain fractional embeddings (described below). An element embedding matrix is generated by applying a fully connected network to the element representations. A fractional embedding matrix is created from the fractional embeddings. These matrices are then added together (element-wise) to generate the element derived matrix (`EDM`, see Figure 5). Each row of the `EDM` ($i$-index) represents an element and the columns ($k$-index) contain the element embeddings. We batch each unique chemical composition onto a third dimension (the $i$-index). The resulting three-dimensional tensor contains the input data for the `CrabNet` architecture.

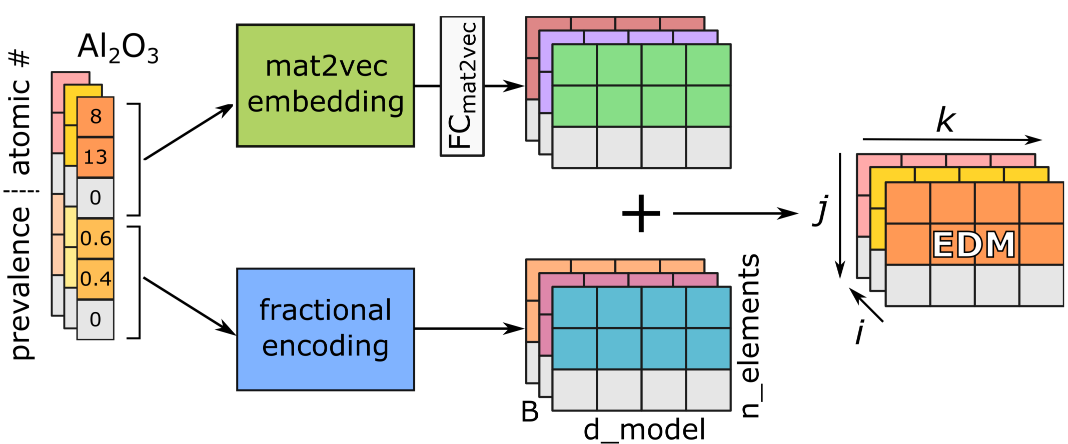


Figure 5: Schematic illustration of the element-derived matrix (`EDM`) representation for $Al_2O_3$, where $B$ represents the batch, $d_{\text{model}}$ is the element features, and $n_{\text{elements}}$ represents the number of elements. Composition slices, when concatenated across batch dimension $i$, form an `EDM` tensor which is then used as the model input to `CrabNet`. When a chemical formula has fewer elements than rows in the `EDM`, the extra data rows are filled with zeros.

We use the `mat2vec` element embeddings[70] as our default source of chemical information for each element (though there are many choices for element properties such as `Jarvis`,[22] Magpie,[71] Oliynyk,[18] or a simple onehot encoding). The mat2vec embedding has the advantage of being pre-scaled and normalized, and having no missing elements or element features. Regardless of the choice of element representation, the representation must be reshaped to fit the the attention input dimensions of ($d_{\text{model}}$). This is done using a learned embedding network; the result is a matrix of size ($n_{\text{elements}}, d_{\text{model}}$). mat2vec.[70] In addition to the default training of CrabNet using the mat2vec embedding, a onehot embedding of the elements was used to train an additional CrabNet model to better facilitate comparison with ElemNet.

The stoichiometric information for each element in the EDM is represented by two fractional embeddings. The fractional embeddings are inspired by the positional encoder as described in the seminal work by Vaswani *et al.*[37] We use *sin* and *cosine* functions of various periods to project the fractional amounts into a high-dimensional space (dimension = $d_{\text{model}}/2$) where smooth interpolation between fractional values is preserved. The first part of the fractional embedding represents the stoichiometry, using the normalized fractional amounts, on a linear scale with a fractional resolution of 0.01. The second part of the embedding maps stoichiometry using a log scale and spans $1 \times 10^{-6}$ to $1 \times 10^{-1}$. This log transformation of the fractional embedding preserves small fractional amounts such as those present in doping. The two parts of the fractional embedding for all elements are concatenated across the embedding dimension to obtain a matrix of size ($n_{\text{elements}}, d_{\text{model}}$).

Once the element and fractional embeddings are calculated and added together, we can batch the finished EDMs. This gives the final input data of shape ($n_{\text{compounds}}, n_{\text{elements}}, d_{\text{model}}$), where $n_{\text{compounds}}$ is the total number of compounds in a given batch, $n_{\text{elements}}$ is the number of rows in the EDM (inferred from the number of elements in the largest composition in a given dataset), and $d_{\text{model}}$ is the size of the embeddings. Here, we also note that the exact ordering of the element rows ($j$) in a compound in the EDM does not influence CrabNet due to the permutation-invariant nature of the self-attention mechanism.

### CrabNet Network Structure

CrabNet contains two primary modules with the default hyperparameters as shown in Table 3. The first module is a Transformer encoder with 3 layers and 4 attention heads in each layer. The second module is a residual network that converts element vectors into element contributions.

Table 3: List of default model parameters of CrabNet.

| Parameter | description | default value |
|---|---|---|
| $in_{\text{dims}}$ | (input) dimension of element embedding | 200 (mat2vec); 118 (onehot) |
| $d_{\text{model}}$ | dimension for EDM and positional encoder | 512 |
| $d_{\text{ff}}$ | feedforward dimension for self-attention mechanism | 2048 |
| $d_{\text{k}}$ | key dimension (equal to $d_{\text{q}}$ in this work) | $d_{\text{model}}/H = 128$ |
| $H$ | number of attention heads per attention block | 4 |
| $N$ | number of stacked self-attention layers | 3 |
| $res_{\text{nodes}}$ | number of nodes at each layer for residual network | [1024, 512, 256, 128] |
| $out_{\text{dims}}$ | (output) dimensions of residual network | 3 |

To understand the Transformer encoder, we first describe the self-attention mechanism. During self attention (Figure 6a), the EDM is operated on by three fully-connected linear networks ($\text{FC}_\text{Q}$, $\text{FC}_\text{K}$, and $\text{FC}_\text{V}$). These networks generate the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ tensors. These tensors can be conceptualized as a learned high-dimensional space where the model stores chemical behavior from the training data.

The $\mathbf{K}$ and $\mathbf{Q}$ tensors contain information regarding the magnitude to which elements interact. The $\mathbf{V}$ tensor stores the information that is used to map from element to property contribution. The dot product of each $\mathbf{Q}$ and $\mathbf{K}^\mathsf{T}$ tensor pair generates the relative element importances in the system (Figure 6b). The importances are scaled using a constant $\sqrt{d_k}$ and then normalized using a softmax function. This results in the self-attention tensor, commonly referred to as the "attention map". We denote this tensor as $\mathbf{A}$. The matrix multiplication of $\mathbf{A}$ with $\mathbf{V}$ updates the element-representations in the compound based on the importance of each element.

Each of the four attention-heads independently performs self-attention with their own $\mathbf{Q}_i$, $\mathbf{K}_i$, $\mathbf{V}_i$, and $\mathbf{Z}_i$ tensors, where $i$ is the head index. As a result, the network generates four

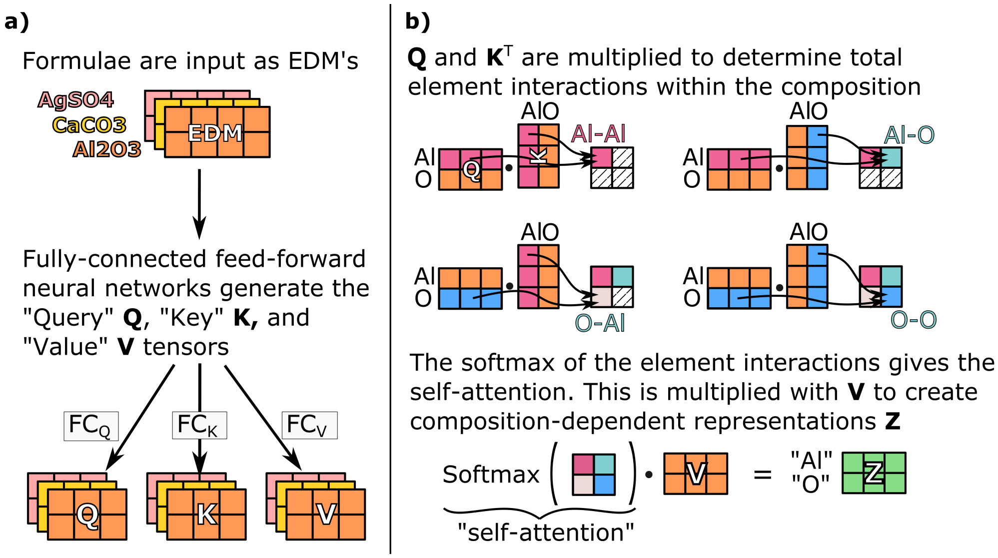


Figure 6: Schematic of an attention block in the `CrabNet` architecture. (a) the initial projection of the input `EDM` into the $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ tensors. (b) the scaled dot-product attention operation obtaining the self-attention matrix and the updated $\mathbf{Z}$ element representation. Here, the batch dimension is not shown to improve legibility.

different element representations at each layer. The individual $\mathbf{Z}_i$ tensors are concatenated across the last dimension to make the $\mathbf{Z}$ tensor (as seen in Figure 7a). The $\mathbf{Z}$ tensor is then passed into a linear FC network which combines the element representations from each head. The output of this FC network is an updated `EDM`′ (for each composition in the batch). This process of converting an `EDM` into an updated `EDM`′ is referred to as a self-attention block. `CrabNet` repeats the process of updating the `EDM` via the self-attention block three times (hence, three layers) resulting in the final updated representations, denoted `EDM`″. This concludes the transformer encoder module.

Once the Transformer encoder has updated the element representations, each `EDM`″ passes through a fully-connected residual network hidden with layer dimensions of $res_{nodes}$. The residual network then transforms the `EDM`s into the shape $(n_{\text{elements}}, n_{\text{elements}}, 3)$. We define these final three vectors as the element-proto-contributions $p'$, element-uncertainties $u'$, and element-logits (see Figure 7a). The element scaling factor $s$ is obtained by taking the sigmoid of the element-logits. The element-contributions are then obtained by multiplying the element-proto-contributions $p'$ by their respective scaling factor $s$. This results in element-

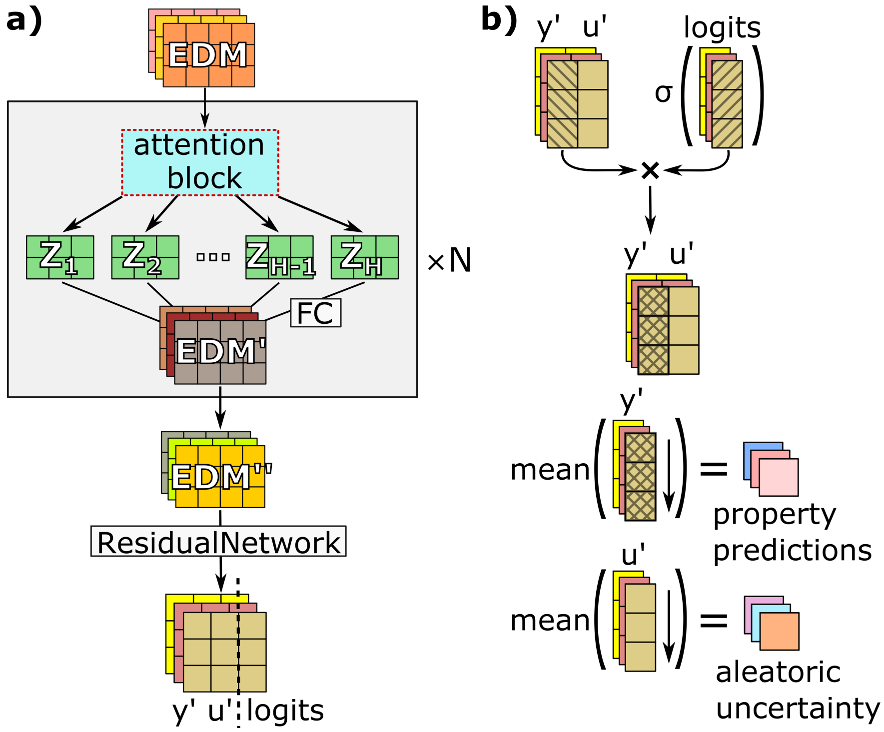


Figure 7: In a), we show a schematic of the `CrabNet` architecture including the input `EDM`, the self-attention layers (repeated $N$ times), the updated and final element representations (`EDM`$'$ and `EDM`$''$), the residual network, and the final model output. In b), we show how element-contributions and prediction of the target and uncertainties are obtained. The $p'$ and $u'$ vectors represent the element-proto-contributions and the element uncertainties, respectively. $y'$ represent the element-contributions. The material property is obtained by taking the mean of element-contributions ($y'$) for each compound. Similarly, the mean of the element-uncertainties ($u'$) gives us the estimated aleatoric uncertainty.

contributions $y'$. Finally, the mean of the element-contributions is taken and output as the predicted property value for each compound (see Figure 7b). Similarly, the mean of the element-uncertainties is used in the aleatoric uncertainty prediction as described by `Roost`.[9]

## Training `CrabNet`

After the featurization of compositions into `EDM`s, the dataset loading and batching is performed with the built-in `Datasets` and `DataLoaders` classes from `PyTorch`. All target values are scaled to zero-centered mean and unit variance for training and inference. The target scaling is then undone for performance evaluation. Batch size during training is dynamically calculated using the training set size for faster training, and limited to be within the range $2^7$ to $2^{12}$. For inference, the batch size was fixed at $2^7$.

Model weights are updated using the look-ahead[72] and Lamb optimizer[73] with a learning rate that is cycled between $1 \times 10^{-4}$ and $6 \times 10^{-3}$ every 4 epochs to achieve consistent model convergence. A robust mean absolute error (MAE)[9] is used as the loss criterion for model performance. The default parameters generalize well when predicting most of the benchmark materials properties. Although we expect that optimization of hyperparameters may improve `CrabNet`'s results for individual materials properties, we believe it is more important that materials scientists be able to use `CrabNet` with little or no adjustments to the underlying code.

It is a known phenomenon that random weight initialization can impact the performance of the Transformer encoder architecture. Thus, to mitigate variance in the performance metrics between different model runs, we trained `CrabNet` using a fixed random seed of 42 for all training runs across all materials properties. We do note that in the case of random model initialization, the run-to-run variation between different trained models is a feature that could be taken advantage of for determining the epistemic uncertainty. Unfortunately, due to the sheer volume of materials properties investigated in this work and the limited compute resources available, we have not investigated this thus far.

Finally, we note that all model training, evaluation and benchmarking (`CrabNet`, `Roost`, `ElemNet`, and `RF`) was conducted on a single workstation PC equipped with an Intel i9-9900K CPU, 32 GB of DDR4 RAM, and two NVIDIA RTX 2080 Ti GPUs with 10 GB VRAM per GPU. The deep learning models were trained on the GPU, while the `RF` models were trained on the CPU.

## Reference Models

### `Roost` predictions

Predictions for all materials properties were generated using code from the `Roost` repository. Minor adaptations were made to the code to allow for automated benchmarking. Overall, `Roost` generates consistently impressive results. In the case of OQMD_Bandgap, `Roost` failed to converge on multiple training runs, hence the reported `NaN` (not a number) value for its metrics. `Roost` relies on a soft-attention mechanism used over a graph representation of the compound. This is in the same spirit of `CrabNet`, and both seek to generate vector representations for the elements in the system without using structure information. The residual network and robust loss function from `Roost` were helpfully adopted into our architecture.[9]

### `ElemNet` predictions

Predictions from `ElemNet` were generated using default parameters. Custom scripts were written to train and evaluate `ElemNet` over all materials properties data. `ElemNet` consistently under-performed compared to `Roost` and `CrabNet`. `ElemNet` also failed to converge for multiple properties resulting in NaN values in the model outputs. An example of this occurring is in the OQMD_Bandgap dataset. Here, we would like to note that `IRNet` could have also been benchmarked and compared in this study. However, due to the prohibitively large computational requirements, we chose to not train and evaluate IRNet. We do however note the OQMD performance reported in the `IRNet` manuscript[6] is consistently lower than both `Roost` and `CrabNet` for the same properties. These following values show the performance of `IRNet` vs. `HotCrab`, respectively, for Formation enthalpy (0.048 eV vs. 0.032 eV), band gap (0.047 eV vs. 0.038 eV), energy per atom (0.070 eV vs. 0.035 eV), and volume per atom (0.394 $Å^3$ vs. 0.247 $Å^3$).

### Random Forest baseline

We generate baseline metrics using a random forest regression with the `Magpie CBFV` as defined by Matminer.[36] This is done using the `scikit-learn` Python package. The `RF` models were trained with $n_{\text{estimators}} = 500$ and default parameters.

## Data Availability

Data is provided in its cleaned and pre-split form to ensure reproducible results and with the hope that other researchers find it useful when benchmarking their own approaches. We also provide detailed instructions for installation, training, and general usage of this open-source tool on GitHub.[69]

Finally, we recommend that readers consult the paper "Machine Learning for Materials Scientists: An introductory guide towards best practices"[4] for a detailed treatment of best practices in machine learning and justification for many of the unmentioned experimental design decisions used in this work.

The following files are available with this publication: (1) `GitHub` repository with the source code, figures, pre-trained weights and example property predictions: `https://github.com/anthony-wang/CrabNet`, and (2) Supplementary Information.

## Conclusions

Unique challenges exists when applying machine learning to materials science. In this paper, we address the limitations of machine learning on chemical composition by introducing `CrabNet`. The `CrabNet` architecture uses a Transformer encoder and the `EDM` representation scheme to perform context-aware learning on materials properties. Using 28 benchmark datasets, we demonstrate `CrabNet`'s performance compared to `Roost`, `ElemNet`, and `RF` baselines. `CrabNet` exhibits consistent predictive accuracy across the full range of materials properties.

Furthermore, we show that a Transformer-based learning technique also provides new methods for visualizing model behavior. We demonstrate the use of attention and per-element prediction capabilities for visualizing common trends in our trained models that match chemical expectations. Given this novel application of self-attention in the context of materials science, we expect that there can be many informative and impactful follow-up works. Specifically, we believe these will largely fall into three thematic categories:

1. **`CrabNet` directly contributing to the community's focus towards improved property predictions.**

   `CrabNet` consistently generates good MAE scores. The performance achieved with the use of self-attention, combined with the innovative use of novel element and composition featurization techniques, will allow researchers to delve deeper into analyzing and predicting materials properties. As a result, we believe that `CrabNet` will be relevant in areas where other ML methods fall short (*e.g.*, dopants, small data, and materials extrapolation tasks). We also note that with minimal changes to `CrabNet` it can also perform classification tasks; we expect `CrabNet` to similarly excel at this.

2. **Attention-based models allow for new ways of thinking about materials-specific problems.**

   In this work we briefly examined the attention mechanism. Attention highlights important interactions and may be used to understand which element-interactions mediate materials properties. Model explainability has thus far been elusive to the traditional materials informatics paradigms. The inclusion of self-attention in this work has introduced new areas of model inspectability that may be a step towards this goal.

3. **Augmentation of `CrabNet` using structural and domain-specific knowledge.**

   This work intentionally used a compositionally-restricted `EDM` representation with no structural information. Structure-agnostic learning is an important task in materials informatics and `CrabNet`demonstrates that accurate learning is achievable using the self-attention mechanism. However the prediction of materials properties using crystal information is also an important task. Integration of structural information could be done by describing elements in their structural and chemical environments. We expect that the self-attention mechanism of `CrabNet` will be able to utilize this additional information to make more accurate predictions. This application of attention-based learning to crystal systems is an exciting and promising direction. We also expect that materials prediction tasks involving processing steps or other non-compositional features could be used in this approach. Both of these changes could easily be implement as extensions to the `EDM`.

While further research is necessary to fully discern the utility of self-attention in materials problems, we believe that this paper highlights a major new direction in its application in materials informatics and suggests exciting new directions for future research.

## Acknowledgements

The authors gratefully acknowledge support from the NSF CAREER Award DMR 1651668. The authors also thank the Berlin International Graduate School in Model and Simulation based Research as well as the German Academic Exchange Service (under program number 57438025) for their financial support. Special thanks is given to Dr. Aleksander Gurlo and Mathias Czasny for advising and supporting Anthony Yu-Tung Wang and for encouraging his collaborative stay at the University of Utah.

The authors thank the creators of AFLOW for the creation of the database and for making the material properties available for this study. In addition, the authors express their gratitude to the open-source software community, for developing the excellent tools used in this research, including but not limited to Python, `Pandas`, `NumPy`, `matplotlib`, `scikit-learn`, and `PyTorch`.

Last but not least, the authors thank OpenAI, the researchers at Hugging Face, and Adam King for their contribution to `TalkToTransformer.com`. The underlying Transformer-powered GPT-2 model was used to generate text for the closing lines of this publication.

## Contributions

AYTW and SKK jointly and in equal amounts conceived, developed the concept, and implemented the algorithms, code and visualizations described in this work. AYTW and SKK analyzed the results.

RJM assisted with developing the architecture and provided insight and guidance during model optimization and training.

All authors discussed the results and contributed to the writing of the manuscript.

## Competing interests

The authors declare no conflicts of interest.

## References

(1) Maier, W. F.; Stöwe, K.; Sieg, S. Combinatorial and high-throughput materials science. *Angewandte Chemie (International ed. in English)* **2007**, *46*, 6016–6067.

(2) Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials* **2016**, *4*, 053208.

(3) Barnard, A. S. Best Practice Leads to the Best Materials Informatics. *Matter* **2020**, *3*, 22–23.

(4) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chemistry of Materials* **2020**, *32*, 4954–4965.

(5) Jha, D.; Ward, L.; Paul, A.; Liao, W.-K.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Scientific Reports* **2018**, *8*, 17593.

(6) Jha, D.; Ward, L.; Yang, Z.; Wolverton, C.; Foster, I.; Liao, W.-K.; Choudhary, A.; Agrawal, A. IRNet: A General Purpose Deep Residual Regression Framework for Materials Discovery. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining – KDD '19. New York, NY, USA, 2019; pp 2385–2393.

(7) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **2018**, *120*, 145301.

(8) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **2018**, *148*, 241722.

(9) Goodall, R. E. A.; Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nature Communications* **2020**, *11*, 6280.

(10) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nature Communications* **2018**, *9*, 2775.

(11) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry* **2015**, *115*, 1094–1101.

(12) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals. *Physical Review Letters* **2016**, *117*, 135502.

(13) Kong, C. S.; Luo, W.; Arapan, S.; Villars, P.; Iwata, S.; Ahuja, R.; Rajan, K. Information-theoretic approach for the discovery of design rules for crystal chemistry. *Journal of Chemical Information and Modeling* **2012**, *52*, 1812–1820.

(14) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nature Materials* **2006**, *5*, 641–646.

(15) Curtarolo, S.; Morgan, D.; Persson, K. A.; Rodgers, J.; Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Physical Review Letters* **2003**, *91*, 135503.

(16) Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *The Journal of Physical Chemistry Letters* **2018**, *9*, 1668–1673.

(17) Kauwe, S. K.; Graser, J.; Vazquez, A.; Sparks, T. D. Machine Learning Prediction of Heat Capacity for Solid Inorganics. *Integrating Materials and Manufacturing Innovation* **2018**, *7*, 43–51.

(18) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chemistry of Materials* **2016**, *28*, 7324–7331.

(19) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding Nature's Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory. *Chemistry of Materials* **2010**, *22*, 3762–3767.

(20) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *Journal of the American Chemical Society* **2018**, *140*, 9844–9853.

(21) Graser, J.; Kauwe, S. K.; Sparks, T. D. Machine Learning and Energy Minimization Approaches for Crystal Structure Predictions: A Review and New Horizons. *Chemistry of Materials* **2018**, *30*, 3601–3612.

(22) Choudhary, K.; DeCost, B.; Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Physical Review Materials* **2018**, *2*, 083801.

(23) Kauwe, S. K.; Graser, J.; Murdock, R. J.; Sparks, T. D. Can machine learning find extraordinary materials? *Computational Materials Science* **2020**, *174*, 109498.

(24) Gaultois, M. W.; Oliynyk, A. O.; Mar, A.; Sparks, T. D.; Mulholland, G. J.; Meredig, B. Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Materials* **2016**, *4*, 053213.

(25) de Jong, M.; Chen, W.; Notestine, R.; Persson, K. A.; Ceder, G.; Jain, A.; Asta, M.; Gamst, A. A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds. *Scientific Reports* **2016**, *6*, 34256.

(26) Glaudell, A. M.; Cochran, J. E.; Patel, S. N.; Chabinyc, M. L. Impact of the Doping Method on Conductivity and Thermopower in Semiconducting Polythiophenes. *Advanced Energy Materials* **2015**, *5*, 1401072.

(27) Zhang, S. B. The microscopic origin of the doping limits in semiconductors and wide-gap materials and recent developments in overcoming these limits: a review. *Journal of Physics: Condensed Matter* **2002**, *14*, R881–R903.

(28) Sheng, L.; Wang, L.; Xi, T.; Zheng, Y.; Ye, H. Microstructure, precipitates and compressive properties of various holmium doped NiAl/Cr(Mo,Hf) eutectic alloys. *Materials & Design* **2011**, *32*, 4810–4817.

(29) Mansouri Tehrani, A.; Oliynyk, A. O.; Rizvi, Z.; Lotfi, S.; Parry, M.; Sparks, T. D.; Brgoch, J. Atomic Substitution to Balance Hardness, Ductility, and Sustainability in Molybdenum Tungsten Borocarbide. *Chemistry of Materials* **2019**, *31*, 7696–7703.

(30) Mihailovich,; Parpia, Low temperature mechanical properties of boron-doped silicon. *Physical Review Letters* **1992**, *68*, 3052–3055.

(31) Qu, Z.; Sparks, T. D.; Pan, W.; Clarke, D. R. Thermal conductivity of the gadolinium calcium silicate apatites: Effect of different point defect types. *Acta Materialia* **2011**, *59*, 3841–3850.

(32) Sparks, T. D.; Fuierer, P. A.; Clarke, D. R. Anisotropic Thermal Diffusivity and Conductivity of La-Doped Strontium Niobate Sr2Nb2O7. *Journal of the American Ceramic Society* **2010**, *93*, 1136–1141.

(33) Grimvall, G. *Thermophysical properties of materials*; Elsevier and North Holland: Amsterdam and Lausanne (Suisse) and New York (N.Y.), 1999.

(34) Gaumé, R.; Viana, B.; Vivien, D.; Roger, J.-P.; Fournier, D. A simple model for the prediction of thermal conductivity in pure and doped insulating crystals. *Applied Physics Letters* **2003**, *83*, 1355–1357.

(35) Murdock, R. J.; Kauwe, S. K.; Wang, A. Y.-T.; Sparks, T. D. Is Domain Knowledge Necessary for Machine Learning Materials Properties? *Integrating Materials and Manufacturing Innovation* **2020**, *9*, 221–227.

(36) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property

prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials* **2020**, *6*.

(37) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. arXiv, 2017-06-12; http://arxiv.org/pdf/1706.03762v5.

(38) Radford, A.; Wu, J.; Amodei, D.; Amodei, D.; Clark, J.; Brundage, M.; Sutskever, I. Better Language Models and Their Implications. 2019; https://openai.com/blog/better-language-models/.

(39) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. 2019.

(40) Tang, G.; Müller, M.; Rios, A.; Sennrich, R. Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. 2018-08-27; http://arxiv.org/pdf/1808.08946v3.

(41) Al-Rfou, R.; Choe, D.; Constant, N.; Guo, M.; Jones, L. Character-Level Language Modeling with Deeper Self-Attention. arXiv, 2018-08-09; http://arxiv.org/pdf/1808.04444v2.

(42) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv, 2018; http://arxiv.org/pdf/1810.04805v2.

(43) Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; Le, Q. V. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. arXiv, 2018; http://arxiv.org/pdf/1804.09541v1.

(44) Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q. V. XLNet:

Generalized Autoregressive Pretraining for Language Understanding. arXiv, 2019; http://arxiv.org/pdf/1906.08237v2.

(45) Huang, C.-Z. A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Simon, I.; Hawthorne, C.; Dai, A. M.; Hoffman, M. D.; Dinculescu, M.; Eck, D. Music Transformer. arXiv, 2018; http://arxiv.org/pdf/1809.04281v3.

(46) Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-Attention Generative Adversarial Networks. arXiv, 2018; http://arxiv.org/pdf/1805.08318v2.

(47) Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; Zhang, L. Second-Order Attention Network for Single Image Super-Resolution. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019-06-15 - 2019-06-20; pp 11057–11066.

(48) Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. In *Computer Vision – ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Vol. 11211; pp 294–310.

(49) Zhang, Y.; Li, K.; Li, K.; Zhong, B.; Fu, Y. Residual Non-local Attention Networks for Image Restoration. arXiv, 2019; http://arxiv.org/pdf/1903.10082v1.

(50) Kim, T. H.; Sajjadi, M. S. M.; Hirsch, M.; Schölkopf, B. In *Computer Vision – ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Vol. 11207; pp 111–127.

(51) Wang, X.; Chan, K. C. K.; Yu, K.; Dong, C.; Loy, C. C. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks. arXiv, 2019-05-07; http://arxiv.org/pdf/1905.02716v1.

(52) Vinyals, O. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **2019**, *575*, 350–354.

(53) Baker, B.; Kanitscheider, I.; Markov, T.; Wu, Y.; Powell, G.; McGrew, B.; Mordatch, I. Emergent Tool Use From Multi-Agent Autocurricula. arXiv, 2019; `http://arxiv.org/pdf/1909.07528v2`.

(54) Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying Structure-Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism. *Journal of Chemical Information and Modeling* **2019**, *59*, 914–923.

(55) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019**, *5*, 1572–1583.

(56) Clement, C. L.; Kauwe, S. K.; Sparks, T. D. Benchmark AFLOW Data Sets for Machine Learning. *Integrating Materials and Manufacturing Innovation* **2020**, Accepted, in press. DOI: `https://doi.org/10.1007/s40192-020-00174-4`.

(57) Bartel, C. J.; Trewartha, A.; Wang, Q.; Dunn, A.; Jain, A.; Ceder, G. A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials* **2020**, *6*, 97.

(58) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **2015**, *1*.

(59) Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **2018**, *152*, 60–69.

(60) Kauwe, S. K. Online GitHub repository for mse_datasets. 2020; `https://github.com/kaaiian/mse_datasets`.

(61) Castelli, I. E.; Olsen, T.; Datta, S.; Landis, D. D.; Dahl, S.; Thygesen, K. S.; Jacobsen, K. W. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy & Environmental Science* **2012**, *5*, 5814–5819.

(62) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

(63) Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; Persson, K. A. The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Computational Materials Science* **2015**, *97*, 209–215.

(64) Petousis, I.; Mrdjenovich, D.; Ballouz, E.; Liu, M.; Winston, D.; Chen, W.; Graf, T.; Schladt, T. D.; Persson, K. A.; Prinz, F. B. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Scientific Data* **2017**, *4*, 160134.

(65) de Jong, M.; Chen, W.; Angsten, T.; Jain, A.; Notestine, R.; Gamst, A.; Sluiter, M.; Krishna Ande, C.; van der Zwaag, S.; Plata, J. J.; Toher, C.; Curtarolo, S.; Ceder, G.; Persson, K. A.; Asta, M. Charting the complete elastic properties of inorganic crystalline compounds. *Scientific Data* **2015**, *2*, 150009.

(66) National Institute of Standards and Technology (NIST), NIST JARVIS-DFT Database. 2017; `https://www.nist.gov/programs-projects/jarvis-dft`, accessed May 5, 2020.

(67) Petretto, G.; Dwaraknath, S.; P C Miranda, H.; Winston, D.; Giantomassi, M.; van Setten, M. J.; Gonze, X.; Persson, K. A.; Hautier, G.; Rignanese, G.-M. High-throughput

density-functional perturbation theory phonons for inorganic materials. *Scientific Data* **2018**, *5*, 180065.

(68) Conduit, G.; Bajaj, S. Mechanical properties of some steels: ID: 153092 - Version 3. 2017; https://citrination.com/datasets/153092/.

(69) Wang, A. Y.-T.; Kauwe, S. K. Online GitHub repository for CrabNet. 2020; https://github.com/anthony-wang/CrabNet.

(70) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98.

(71) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*, 16028.

(72) Zhang, M. R.; Lucas, J.; Hinton, G.; Ba, J. Lookahead Optimizer: k steps forward, 1 step back. arXiv, 2019-07-19; http://arxiv.org/pdf/1907.08610v2.

(73) You, Y.; Li, J.; Reddi, S.; Hseu, J.; Kumar, S.; Bhojanapalli, S.; Song, X.; Demmel, J.; Keutzer, K.; Hsieh, C.-J. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. arXiv, 2019-04-01; http://arxiv.org/pdf/1904.00962v5.