Investigating extrapolation and low-data challenges via contrastive learning of chemical compositions

Federico Ottomano
Department of Computer Science
University of Liverpool
Liverpool, L69 3DR, UK
federico.ottomano@liverpool.ac.uk

Giovanni De Felice
Department of Computer Science
University of Liverpool
Liverpool, L69 3DR, UK

Rahul Savani
The Alan Turing Institute
London, NW1 2DB
Department of Computer Science
University of Liverpool
Liverpool, L69 3DR, UK

Vladimir Gusev
Department of Computer Science
University of Liverpool
Liverpool, L69 3DR, UK

Matthew Rosseinsky
Materials Innovation Factory
University of Liverpool
Liverpool, L7 3NY, UK

Abstract

Practical applications of machine learning for materials discovery remain severely limited by the quantity and quality of the available data. Furthermore, little is known about the ability of machine learning models to extrapolate outside of the training distribution, which is essential for the discovery of compounds with extraordinary properties. To address these challenges, we develop a novel deep representation learning framework for chemical compositions. The proposed model, named COmpositional eMBedding NETwork (CombNet), combines recent developments in graph-based encoding of chemical compositions with a supervised contrastive learning approach. This is motivated by the observation that contrastive learning can produce a regularized representation space from raw data, offering empirical benefits for extrapolation in low-data scenarios. Moreover, our method harnesses exclusively the chemical composition of the underlying materials, as crystal structure is generally unavailable before the material is discovered. We demonstrate the effectiveness of CombNet over state-of-the-art methods under a bespoke evaluation scheme that simulates a realistic materials discovery scenario with experimental data.

1 Introduction

Materials discovery is increasingly benefiting from the synergy between recent advancements in Machine Learning (ML) and the growing availability of material databases [Jain et al., 2013, Blokhin and Villars, 2018], revealing interesting perspectives in accelerating the exploration of new materials [Wang et al., 2022, Tewari et al., 2020, Hargreaves et al., 2023]. ML aims to address the limitations imposed by physics-based simulations in *density functional theory* (DFT), which require substantial computational resources and are prone to systematic errors due to numerical approximations [Schleder

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Ai4Mat Workshop.

et al., 2019]. Despite these interesting premises, the desire to discover novel over-performing materials poses unique challenges to ML. First, available data offer a narrow diversity in terms of chemical properties and mainly originate from DFT calculations. This affects the suitability for industrial applications, frequently necessitating specialized materials with unconventional traits and might bias ML models away from modeling real-world experimental conditions. Secondly, the discovery of extraordinary materials requires designing stable approaches to predict material properties outside the known data distribution, which is known as *extrapolation*. However, extrapolation in ML is just in a nascent state [Courtois et al., 2023, Xu et al., 2021]. As a result, the majority of current research predominantly occurs within *interpolation* settings [Zhuo et al., 2018, Wang et al., 2021, Goodall and Lee, 2020], where models' predictions are evaluated on a random subset of the original dataset. We argue that this is not in line with the interest of researchers, which is more often aimed at the discovery of materials with extraordinary properties, e.g., room temperature superconductors [Pickett, 2023], rather than compounds with an average behavior. The interplay between these two primary considerations results in significant limitations for most of the popular approaches in the field. For example, ensemble methods [Breiman, 2001a, Chen and Guestrin, 2016] have shown great robustness in predicting chemical properties in low-data regimes [Chelladurai et al., 2022, Riebesell, 2016, Gaultois et al., 2016], but are unable to extrapolate, as new predictions will simply be generated from averages of instances in the training dataset [Ellis]. On the other hand, modern deep learning architectures dealing either with compositions [Wang et al., 2021, Goodall and Lee, 2020] or structures [Xie and Grossman, 2018, Choudhary and DeCost, 2021, Chen et al., 2019], offer more promises for extrapolation but are still severely limited by the lack of extensive training datasets.

In this work, we tackle both challenges, i.e., low-data regime and extrapolation, by developing a deep representation learning framework for chemical compositions driven by supervised *Contrastive Learning* (CL). CL is a popular representation learning paradigm that clusters together similar data points according to predefined similarity criteria, enhancing the discriminative power of the learned features Le-Khac et al. [2020]. Because of that, it recently achieved broad success in computer vision [Chen et al., 2020, Khosla et al., 2020, Zbontar et al., 2021] and natural language processing [Gao et al., 2022, Zhang et al., 2022a] domains. Our motivation comes from the recently established promises of CL in low-data regimes [Na and Kim, 2022] and extrapolation [Na et al., 2022]. In fact, empirical evidence has shown that CL can be valuable in obtaining meaningful data representations when dealing with limited training datasets [Na and Kim, 2022]. Intuitively, learning the latent space directly by grouping together similar samples can lead to more discriminative features in the learned representations. In contrast, neural networks may face challenges in inferring the representation space when only a relatively small amount of training data is available. Furthermore, recent studies have shown that *multi-layer perceptrons* (MLPs) tend to converge towards linear functions when evaluated outside of the training distribution [Courtois et al., 2023]. On the other hand, a CL approach can make the latent space smoother and more amenable to linear relationships, yielding favorable outcomes in extrapolation tasks [Na et al., 2022] when paired with a downstream MLP. Notably, the proposed approach exclusively relies on the chemical composition of the underlying materials. Models centered on composition hold substantial value for materials discovery, as composition can be defined for materials that may not have been discovered yet, whereas crystal structure is unavailable in this regard. Our main contributions can be summarized as follows:

- We introduce *COmpositional eMBedding NETwork* (CombNet), a novel framework for representation learning of chemical compositions driven by supervised contrastive learning.
- We provide a thorough analysis of state-of-the-art ML models for property prediction in a realistic discovery scenario, i.e., experimental datasets paired with an extrapolation task.
- We demonstrate the effectiveness of CombNet in learning useful representations over state-of-the-art ML models and chemically-informed feature schemes.

2 Related work

Materials property prediction The use of ML to predict material properties has undergone considerable growth in recent years. Two main strands can be highlighted, namely traditional approaches and deep learning architectures. Conventional featurization schemes, whether from crystalline structures [Himanen et al., 2020, isa, 2017] or from stochiometry alone [Ward et al., 2016, Tshitoyan et al., 2019, Oliynyk et al., 2016], provide a detailed description of the abstract materials space and serve as effective means to encapsulate chemical knowledge. These are usually

paired with traditional approaches such as linear models or ensemble methods (e.g., random forests [Breiman, 2001b]). Despite their simplicity, these methods have found widespread application in the literature [Zhuo et al., 2018, Gaultois et al., 2016, Mansouri Tehrani et al., 2018]. However, linear models may fail to model complex chemical phenomena whereas ensemble methods are unsuitable for extrapolation. Moreover, *deep learning* (DL) models have been harnessed for encoding chemical information beyond feature engineering, thereby expanding the accessibility of ML techniques to a wider audience, including those who may not possess in-depth materials science expertise dom [2020]. Such models are commonly classified according to their input, with some designed to accept only chemical compositions and others able to include crystalline structures as well. Notable examples from the former category include *Elemnet* [Jha et al., 2018], which processes vectorized representations of compositions using a one-hot encoding scheme; *Roost* [Goodall and Lee, 2020], an attentional graph neural network that generates and processes stoichiometric graphs from input compositions; *CrabNet* [Wang et al., 2021], a recently proposed transformer-based model that also operates at the stoichiometry level and that has achieved state-of-the-art results in several proposed benchmarks regarding materials property prediction [Dunn et al., 2020]. Notable examples of the second category are *crystal graph convolutional neural network* (CGCNN) [Xie and Grossman, 2018], *MegNet* [Chen et al., 2019] and *ALIGNN* [Choudhary and DeCost, 2021], all graph neural network models which take into account the geometry of crystal lattices. Despite the remarkable level of flexibility offered by DL models, their success is conditioned to a large availability of training data, which is typically limited in realistic experimental scenarios. Furthermore, none of the mentioned DL models have established theoretical or empirical support for their ability to extrapolate beyond the data they have been trained on. We argue that there is a noticeable absence of prior research that specifically addresses realistic discovery scenarios, i.e., the challenge of simultaneous extrapolation with limited experimental data.

Contrastive Learning CL is a representation learning framework that has gained great popularity in recent years [Le-Khac et al., 2020], especially given its effectiveness in self-supervised learning of visual representations [Chen et al., 2020, Kumar et al., 2022]. Moreover, the adoption of such paradigm has yielded promising outcomes within multi-modal domains explored by recent text-to-image models [Zhang et al., 2022b, Radford et al., 2021]. Despite most applications privileging computer vision and natural language processing, CL recently gained considerable attention in the field of materials informatics: Koker et al. [2022] applied a self-supervised CL framework to learn invariant representations of crystal structures under a predefined set of transformations; Kong et al. [2022] incorporated a supervised CL module to facilitate the alignment between feature and label embeddings under a density-of-states prediction task; Magar et al. [2022] derived crystal representations by minimizing the cross-correlation between pairs of distorted samples originating from the same crystal; Na and Kim [2022] provided a framework to map original crystal structures to a smooth latent space shaped according to an initial target property. However, most of such CL applications focus on crystalline structures, which are not functional for materials discovery tasks, as knowledge of the structure is unavailable before the actual material is discovered. To the best of our knowledge, no previous work has directed CL approaches to chemical compositions.

3 COmpositional eMBedding NETwork

We present COmpositional eMBedding NETwork (CombNet) (depicted in Fig. 1), a novel representation learning framework for chemical compositions. The main components of the architecture can be broadly grouped into two stages: an *encoder block* and a *CL module*. In the encoder, chemical compositions are represented as fully connected graphs, allowing for further processing of the information with a graph-based neural encoder. Later, an additional network is used to project representations into a separate space where the CL module learns a mapping to a regularized metric space. At inference, representations learned through CombNet can be used to initialize a separate MLP designated for downstream property prediction. In this section, we delve into the details of each of these components.

3.1 Encoder block

Composition graphs In order to extract meaningful information from raw compositions, it is necessary to translate them into suitable representations. Different approaches have been adopted to

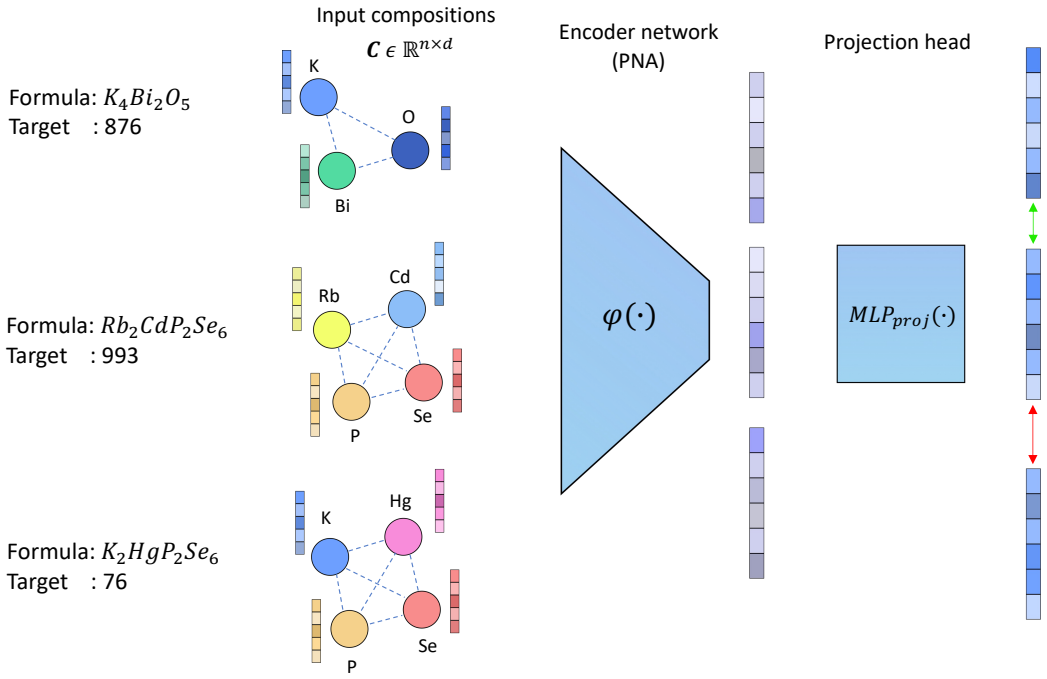

Figure 1: Overview of CombNet. The *encoder network* φ_θ first extracts and processes chemical information from the constituent elements of the composition. Then, the projection head g_ψ further maps representations into a separate space where CL takes place. The embeddings of $Rb_2CdP_2Se_6$ and $K_4Bi_2O_5$ are brought closer together within the embedding space due to their similar target property (indicated below the respective formulas), while simultaneously being pushed apart from the representation of $K_2HgP_2Se_6$.

represent compositions within computational domains. Among those, we choose *mat2vec* [Tshitoyan et al., 2019], which incorporates chemical information by representing each chemical element in a composition with a predefined feature vector $\mathbf{c}_i^k \in \mathbb{R}^d$ $(k = 1, ..., n_i)$, where n_i denotes the number of elements in the $i\text{-}th$ composition and d is the embedding dimensionality. We obtain the representation of the $i\text{-}th$ composition by weighting each $\mathbf{c}_i^k$ by the fractional prevalence of the corresponding element in the chemical formula and stacking them into a matrix $\mathbf{C}_i \in \mathbb{R}^{n \times d}$.

The fundamental constituents of a chemical composition are atoms interconnected by chemical bonds. These are ideal candidates for modeling using Graph Neural Networks (GNNs), which excel at capturing interactions within structured data and exploiting them as architectural bias [Bacciu et al., 2020, Bronstein et al., 2021]. To harness the potential of GNNs for our setting, as in Goodall and Lee [2020], we conceptualize the elements within a composition as nodes and represent chemical bonds as edges in a graph. The node features are given by the rows $\mathbf{c}_i^k$ of $\mathbf{C}_i$, while for the adjacency matrix $\mathbf{A_C}$, we employ a fully connected matrix (with self-loops) with equal unitary edge weights, signifying that every pair of atoms is considered, a priori, equally interconnected. As in Goodall and Lee [2020], for each node (element) in each chemical composition, we project its representation into a learnable space using a linear layer:

$$\mathbf{H}_i = \mathbf{C}_i\mathbf{W} + \mathbf{b} \tag{1}$$

Weights $\mathbf{W}$ and $\mathbf{b}$ are shared across different compositions in the dataset. Fractional amounts of elements are concatenated to the terminal positions of node embeddings in the compositional graph $(\mathbf{H}_i, \mathbf{A_C})$.

Encoder network $\varphi(\cdot)$ A GNN encoder is used to map the nodes (elements) embeddings into compact representation vectors. Specifically, we adopt multiple layers of *Principal Neighbourhood Aggregation* (PNA) [Corso et al., 2020]:

$$\mathbf{H}_i' = \varphi(\mathbf{H}_i, \mathbf{A_C}) = \text{PNA}(\mathbf{H}_i, \mathbf{A_C}) \tag{2}$$

a graph convolution scheme proposed in recent years that demonstrated great performance with continuous node features and outperformed other state-of-the-art message-passing methods like *Graph Attention Network* (GAT) [Veličković et al., 2017, Goodall and Lee, 2020] in various downstream tasks, including molecular-property predictions [Stärk et al., 2022]. In principle, different encoders can be used to process tokenized representations of compositions. In Appendix B we provide a comparison between our chosen PNA encoder with a baseline MLP encoding strategy with no aggregation between neighboring elements. Finally, a mean pooling operation is applied in order to achieve graph-level (composition) representations $\mathbf{m}_i$ from the node (element) representations $\mathbf{h}_i'^k$ (rows of $\mathbf{H}_i'$):

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{k=1}^{n} \mathbf{h}_i'^k \,. \tag{3}$$

3.2 Contrastive learning block

Projection head We employ an MLP, denoted as MLP_{proj}, to project the embeddings produced by the encoder into a separate space, where CL can take place. This is meant to facilitate the following extraction of discriminative features. In fact, it has been empirically shown that this enhances the effectiveness of contrastive training [Chen et al., 2020, He et al., 2020]. As in Chen et al. [2020], we set:

$$\mathbf{z}_i = \text{MLP}_{\text{proj}}(\mathbf{m}_i) = \mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{m}_i) \tag{4}$$

with ReLU activation function. Despite enhancing the training process, it was observed in Chen et al. [2020] that such transformation removes information that might instead be useful for downstream tasks. As a consequence of this evidence, we discard the *projection head* at inference time, i.e., after the training is complete, only the encoder block (Sec. 3.1) is used to create representations of chemical compositions for downstream tasks.

Contrastive learning module: The central idea behind CL is to learn a parametric mapping from an input space (chemical compositions $\mathbf{C}_i$) to an embedding space (representations $\mathbf{z}_i$) by minimizing the distances between representations of similar data points while maximizing the separation between representations of dissimilar ones. Similarity between data points is assessed based on a predefined similarity function. A common declination of this framework is triplet-based representation learning [Schroff et al., 2015]: for each data point , labeled as *anchor* $\mathbf{C}_{\text{anc}}$, both a positive $\mathbf{C}_{\text{pos}}$ and a negative $\mathbf{C}_{\text{neg}}$ are sampled accordingly to a predefined similarity and used to optimize a triplet loss function that minimizes the distance between $\mathbf{z}_{\text{anc}}$ and $\mathbf{z}_{\text{pos}}$, while maximizing the separation between $\mathbf{z}_{\text{anc}}$ and $\mathbf{z}_{\text{neg}}$. While originally applied primarily in unsupervised learning scenarios [Chen et al., 2020, Grill et al., 2020], this approach has witnessed recent extensions to regression tasks Na and Kim [2022], Kim et al. [2019]. In regression, positives and negatives are labeled as such based on their proximity to the anchor's target value: given an anchor point with its target $(\mathbf{C}_{anc}, y_{anc})$, the corresponding positive and negative samples are selected such that $|y_{anc} - y_{pos}| < |y_{anc} - y_{neg}|$, with $y \in \mathbb{R}$ being a one-dimensional scalar target. This approach of matching distances between triplets of data points can be generalized to all the points in the batch and has been introduced as *distance-matching problem* [Na et al., 2022]. In practice, such conceptualization can be expressed through the following objective function:

$$\mathcal{L}(\mathbf{z}_1, ..., \mathbf{z}_N, y_1, ...y_N) = \frac{1}{4N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} (\|\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j\|_2 - |\bar{y}_i - \bar{y}_j|)^2 \,. \tag{5}$$

where $\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j$ and $\bar{y}_i, \bar{y}_j$ are normalized representations and targets to ensure numerical stability during training and N is the number of compositions in a training batch. Specifically to our problem, we define $f_\psi = \text{MLP}_{\text{proj}} \circ \varphi$ as the learnable parametric mapping and $\mathbf{z}_i = f_\psi(\text{C}_i)$, where C_i is the input graph (composition). By employing this approach, the resulting learned representations inherently encode the relative distances of their associated targets.

4 Experiments

In this section, we present our experimental setting aimed at simulating a realistic discovery scenario. In our benchmark, we investigate the performance of different methods (4.3) under an extrapolation task (4.1) within the context of experimental datasets (4.2).

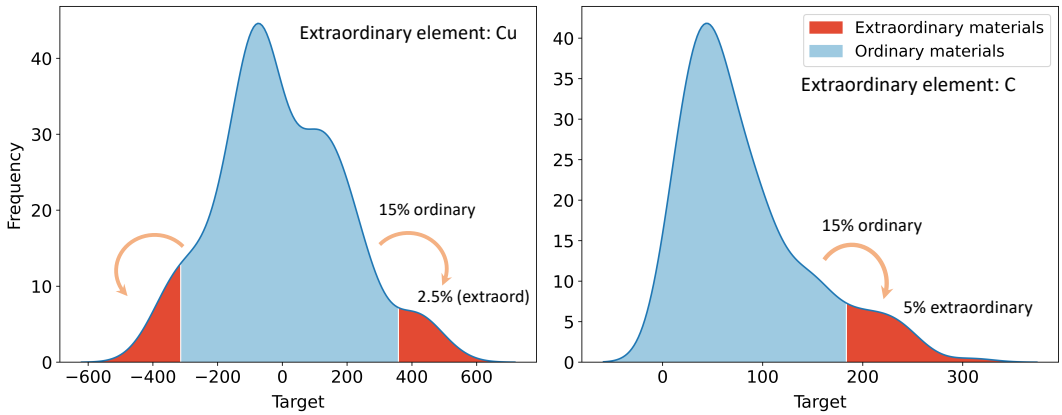

Figure 2: Visual representation of the splitting method applied to the Seebeck coefficient (**left**) and shear modulus (**right**).

4.1 Extrapolation benchmark

In a real materials discovery scenario, the main goal is to correctly identify extraordinary compounds displaying a non-average behavior. This translates into accurately predicting the associated properties of chemical compositions outside the boundaries of the training distribution. To define a suitable benchmark that reflects this purpose, we follow a similar approach to Kauwe et al. [2020]. We consider a regression task and define a biased test set that includes extraordinary compounds based on the following criteria. Figure 2 provides a visual representation of the employed method.

- First, datasets are sorted according to the target property. In situations with one-sided distributions, the top 5% of compositions with the highest (or lowest) associated target are allocated to the test set (Fig. 2, right). In situations involving two-sided distributions, the 5% allocation is divided, with 2.5% representing the highest property values and another 2.5% representing the lowest property values (Fig. 2, left). Secondly, an additional 15% from the rest of the dataset is utilized to populate the test set. This step is aimed at making the extrapolation task less aggressive, while still being representative of the average extrapolation performance of different methods.
- We systematically exclude from the training set all compositions containing the most prevalent element in the test set. For instance, in the case of the Seebeck coefficient, compounds containing Copper (Cu) are removed from the training set. An exception arises when oxygen (O) is the most prevalent element, as it is frequently abundant in several of the considered datasets. In this scenario, we instead remove the second most prevalent element from the training data. The deliberate exclusion of specific chemical elements from the training aims at measuring the ability of different ML models to undertake chemical extrapolation, extending their predictions to materials absent from the training set.

4.2 Datasets

We employ experimental data of thermoelectric properties derived from an available extension of the UCSB repository Gaultois et al. [2013] and experimental data retrieved from the *Materials Platform for Data Science* Blokhin and Villars [2018]. Thermoelectric properties hold pivotal significance in modern materials research and technology, due to their diverse applications across various fields [Finn et al., 2021]. For example, Seebeck coefficient quantifies the material's ability to generate an electric potential difference when subjected to a temperature gradient. An accurate prediction of such quantity could help in energy conversion, waste heat recovery, and innovative heat management solutions across industries [Yuan et al., 2022]. The considered datasets exhibit a relatively small size, which reflects the inherent limitations associated with experimental data acquisition. More details about the datasets can be found in Appendix A

Preprocessing steps are applied to the raw datasets in order to maintain consistency between the various properties under consideration: first, we filter out all the pure elements and measurements

associated with temperatures outside of $\pm 15\,\mathrm{K}$ from room temperature ($298\,\mathrm{K}$). Furthermore, we exclude data points beyond ± 4 standard deviations from their respective medians. This approach aims to ensure an adequate representation of data points within the tail(s) of the distribution, while mitigating the inclusion of extreme values that could potentially introduce excessive noise during the evaluation stage.

4.3 Evaluated models

CombNet Once the *encoder block* has been trained through the CL approach detailed in Sec. 3.2, we evaluate CombNet by means of a separate downstream MLP, named MLP_{pred}, which is separately trained to map representations to the corresponding target properties y_i. This is, for every composition $\mathbf{C}_i$ in the test set:

$$\hat{y}_i = \text{MLP}_{\text{pred}}(\varphi(\mathbf{C}_i))\,, \tag{6}$$

where $\hat{y}_i$ is the predicted property value. The training for this downstream task is performed through the minimization of the Mean Absolute Error (ℓ_1-loss) while all weights of φ are kept frozen. We recall that to generate representations through CombNet, the *projection head* is discarded at this inference stage and only the encoder φ is used.

Baselines We conduct a thorough comparison between CombNet and various ML models that are commonly utilized for predicting material properties. We assess the effectiveness of standard *mat2vec* features [Tshitoyan et al., 2019] paired with traditional ML techniques, i.e., ridge regression (Ridge) [Bishop and Nasrabadi, 2006] and Support Vector Regression (SVR) [Smola and Schölkopf, 2004]. Notably, beyond their prevalent application in materials exploration Zhuo et al. [2018], these approaches have also been studied within an extrapolation setting, to measure their effectiveness in identifying exceptional materials Kauwe et al. [2020]. Additionally, we provide *mat2vec* features directly to a similar MLP_{pred} as the one employed by CombNet. For all the aforementioned models, in order to obtain valid representations at the level of the chemical composition, we average the *mat2vec* features of the constituent elements weighted by the fractional prevalence of each element. Finally, we pair *mat2vec* features with CrabNet [Wang et al., 2021], which we adopt as primary representative of the state-of-the-art in the DL paradigm. Instead, we deliberately avoid any use of ensemble methods, e.g., Random Forest, given the *a priori* inability to extrapolate Ellis. To ensure uniformity in the assessment of different models, *mat2vec* features are also used to initialize compound representations for CombNet (as outlined in Sec. 3.1). Implementation details for CombNet and other baselines can be found in Appendix C.

In order to avoid confusion, we briefly summarize the different MLPs that are utilized in this work, underlining their different purposes: 1) $\textbf{MLP}_{\textbf{proj}}$**:** in CombNet, the *projection head* that is used to transform representations before the application of CL (Eq. 4); 2) $\textbf{MLP}_{\textbf{pred}}$**:** a downstream model that is employed to predict target properties from representations (Eq. 6); in CombNet, this maps the learned representations to the corresponding target properties, while in the employed baseline, this leverages *mat2vec* features instead. 3) $\textbf{MLP}_{\textbf{enc}}$**:** an alternative baseline encoder for updating element representations without an aggregation scheme, investigated in the ablation study proposed in Appendix B).

4.4 Results

In Tables 1, 2 we report the performance of the aforementioned models on different datasets. As evaluation metrics, we consider the *mean absolute error* (MAE) and the *coefficient of determination* (R2), as common choices within regression tasks.

General improvement From the results, we observe that CombNet outperforms the considered baselines on most datasets. For some chemical properties, this leads to a remarkable improvement: for example, when considering Seebeck coefficient, CombNet leads to a reduction in MAE of $\approx$ 20% w.r.t. (MLP_{pred} + *mat2vec*) and $\approx$ 30% w.r.t. CrabNet; similarly, in the case of temperature for congruent melting, we observe a reduction of $\approx 17\%$ w.r.t. (MLP_{pred} + *mat2vec*) and $\approx 21\%$ w.r.t. CrabNet. These improvements are even more pronounced when considering the R2 metric: for example, the performance is almost doubled on Seebeck dataset w.r.t. (MLP_{pred} + *mat2vec*) and a $\approx 45\%$ improvement is observed on thermal conductivity dataset again w.r.t. (MLP_{pred} + *mat2vec*).

Table 1: Extrapolation task: *Mean absolute error* (MAE) for each considered model and dataset. Results are averaged across 5 different random seeds. Best-performing results are shown in green, while second best-performing are shown in yellow, when there is an overlap in the uncertainty bands.

DATASET	SEEBECK	KAPPA	TCONGRMELT	ELECMASS	BMODULUS	SMODULUS	BANDGAP
RIDGE	$171.84_{\pm 3.33}$	$0.4_{\pm 0.01}$	$513.01_{\pm 13.66}$	$0.72_{\pm 0.01}$	$75.1_{\pm 2.57}$	$51.06_{\pm 3.61}$	$2.07_{\pm 0.04}$
SVR	$204.74_{\pm 1.04}$	$0.38_{\pm 0.02}$	$686.33_{\pm 9.21}$	$0.73_{\pm 0.01}$	$97.01_{\pm 2.19}$	$73.00_{\pm 2.39}$	$1.97_{\pm 0.03}$
MLP_{PRED}	$158.36_{\pm 3.92}$	$0.37_{\pm 0.01}$	$341.49_{\pm 40.01}$	$\mathbf{0.65_{\pm 0.01}}$	$\mathbf{57.72_{\pm 1.22}}$	$51.79_{\pm 1.78}$	$1.78_{\pm 0.03}$
CRABNET	$183.63_{\pm 4.71}$	$0.45_{\pm 0.02}$	$358.98_{\pm 12.18}$	$0.83_{\pm 0.01}$	$64.15_{\pm 0.69}$	$69.79_{\pm 3.58}$	$1.88_{\pm 0.03}$
COMBNET (+ MLP_{PRED})	$\mathbf{128.17}_{\pm 7.06}$	$\mathbf{0.36}_{\pm 0.01}$	$\mathbf{282.66}_{\pm 7.82}$	$0.68_{\pm 0.01}$	$61.56_{\pm 3.6}$	$\mathbf{50.64}_{\pm 4.23}$	$\mathbf{1.71}_{\pm 0.01}$

Table 2: Extrapolation task: *Coefficient of determination* (R2) for each considered model and dataset. Results are averaged across 5 different random seeds. Best-performing results are shown in green, while second best-performing are shown in yellow, when there is an overlap in the uncertainty bands. '/' denotes a negative R2 and thus the failure of the corresponding regression task.

DATASET	SEEBECK	KAPPA	TCONGRMELT	ELECMASS	BMODULUS	SMODULUS	BANDGAP
RIDGE	$0.16_{\pm 0.06}$	$0.17_{\pm 0.01}$	$0.38_{\pm 0.03}$	$0.22_{\pm 0.00}$	$0.26_{\pm 0.04}$	$0.3_{\pm 0.05}$	/
SVR	/	$0.21_{\pm 0.02}$	$0.06_{\pm 0.00}$	$0.2_{\pm 0.01}$	/	/	/
MLP_{PRED}	$0.25_{\pm 0.05}$	$0.22_{\pm 0.05}$	$0.64_{\pm 0.07}$	$0.29_{\pm 0.02}$	$\mathbf{0.51_{\pm 0.04}}$	$0.25_{\pm 0.00}$	$0.07_{\pm 0.01}$
CRABNET	$0.14_{\pm 0.02}$	/	$0.67_{\pm 0.04}$	$0.1_{\pm 0.00}$	$0.44_{\pm 0.04}$	/	$0.08_{\pm 0.07}$
COMBNET (+ MLP_{PRED})	$\mathbf{0.47}_{\pm 0.09}$	$\mathbf{0.32_{\pm 0.05}}$	$\mathbf{0.74_{\pm 0.02}}$	$\mathbf{0.33_{\pm 0.02}}$	$0.46_{\pm 0.04}$	$\mathbf{0.35_{\pm 0.05}}$	$\mathbf{0.10}_{\pm 0.01}$

We argue that CombNet may be particularly valuable for thermoelectric applications, especially when dealing with *doped* materials. Doping is a widely practiced technique where the introduction of a low level of distinct atomic species into a parent material radically changes its properties, despite apparently negligible change in composition. Consequently, conventional featurization methods or supervised ML models may struggle to capture these important distinctions. On the other hand, supervised CL could recalibrate the relationships between chemical compositions and their corresponding properties to address this important situation. In fact, given a parent composition, previous approaches may create similar representations for small variations of the original material, while CombNet aligns the representations with the observed differences in the target property. We plan to expand on this work by deepening this consideration.

Comparison with mat2vec The MLP_{pred} functions as both a baseline with standard *mat2vec* features, and as a downstream model for CombNet. This choice allows a fair evaluation of representations deriving from CombNet against chemically informed predefined features. Our results suggest that CombNet can be used to generate more informative representations of chemical compositions in the context of an extrapolation setting.

5 Limitations and future work

Our study addresses two primary challenges in materials informatics: low-data availability and extrapolation. We propose using CL as a primary tool to ameliorate both situations, inspired by the promises established in the recent literature [Na and Kim, 2022, Na et al., 2022]. Nevertheless, the interpretability of our results faces limitations due to the inherent complexity of the task at hand: evaluating different approaches in a pure extrapolation setting may introduce excessive noise, hindering meaningful comparisons between models. So, as a first avenue for future work, it would be interesting to undertake a more extensive study to capture distinct aspects of interpolation and extrapolation, by also including a broader variety of datasets.

A second avenue for further work is motivated by the ablation study reported in Appendix B. Although the main results in Tables 1 and 2 identify CombNet as the top-performing model, the ablation study reported in B reveals some interesting findings. Notably, the model featuring a PNA encoder without CL (trained in an end-to-end manner) exhibits a remarkable performance. As further work, we plan to enhance the comparisons of different models by investigating the performance of various encoders with different levels of complexity, with an without CL, for an improved analysis. As a possible hypothesis in this direction, it seems likely that we would find that the benefit of including CL goes down as one increases the complexity of the encoder, ceteris paribus.

A final important direction for further research is the nature of data utilized as input. In this study, we exclusively rely on the stoichiometry of materials. We recall that this is actually a strength, in the sense that, when trying to discover new materials, we will normally not know much about the underlying crystalline structure. Thus, a stoichiometry-only baseline, as studied here, is a natural and important one. On the other hand, when extra domain knowledge is accessible, it becomes crucial to integrate it appropriately. Therefore, as future work, we envision the integration of additional prior knowledge into the generated material embeddings, e.g. by leveraging recent advancements in Large Language Models (LLMs) for capturing chemistry domain knowledge Xie et al. [2023], Jablonka et al. [2023], or by integrating structural information when it is accessible, e.g., if we have available crystal structures corresponding to the chemical compositions in the datasets under examination.

6 Conclusions

Motivated by the recent surge in the application of CL within the realm of materials science, especially in the domain of crystalline structures, our work introduces CombNet, a novel representation learning framework grounded in supervised CL that relies solely on materials' chemical composition. Contrary to predefined feature schemes, CombNet centers on creating material representations that closely align with the specific chemical property under examination. Additionally, the alignment of representations with their corresponding targets induces a regularized space that becomes more amenable to linear relationships, ultimately simplifying the extrapolation task even when handling complex chemical properties. We have evaluated CombNet in a realistic materials discovery scenario, i.e., low data regime and extrapolation, which we believe to be crucial when assessing the performance of different ML models in materials science. By addressing what we believe are fundamental issues in materials informatics, we believe that our work will contribute to the ongoing exploration of harnessing ML techniques to accelerate the discovery of novel functional materials. As a future research direction, we envision the integration of our approach with additional prior knowledge deriving from crystal structures or pretrainined LLMs, in order to produce more informative material representations. Additionally, we believe it would be interesting to deepen the application of CL techniques, like CombNet, within the task of predicting chemical properties of doped materials for thermoelectric applications.

Data availability

The original UCSB repository can be found at the following link, under the name 'ucsb_thermoelectrics': `https://hackingmaterials.lbl.gov/matminer/dataset_summary.html` An updated version of the UCSB datasets utilized in this work will be made available upon request. All MPDS data can be obtained through accessing the corresponding API. More information can be found at: `https://mpds.io/developer/`.

Acknowledgments and Disclosure of Funding

F.O. acknowledges Pilkington (NSG Group) for funding this research. M.R. acknowledges the EPSRC Programme Grant EP/V026887.

References

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN

2166532X. doi: 10.1063/1.4812323. URL http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi.

Evgeny Blokhin and Pierre Villars. *The PAULING FILE Project and Materials Platform for Data Science: From Big Data Toward Materials Genome*, pages 1–26. Springer International Publishing, Cham, 2018.

Teng Wang, Kefei Zhang, Jesse Thé, and Hesheng Yu. Accurate prediction of band gap of materials using stacking machine learning model. *Computational Materials Science*, 201:110899, 2022. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2021.110899. URL https://www.sciencedirect.com/science/article/pii/S0927025621006078.

Abhishek Tewari, Siddharth Dixit, Niteesh Sahni, and Stéphane P.A. Bordas. Machine learning approaches to identify and design low thermal conductivity oxides for thermoelectric applications. *Data-Centric Engineering*, 1:e8, 2020. doi: 10.1017/dce.2020.7.

Cameron J. Hargreaves, Michael W. Gaultois, Luke M. Daniels, Emma J. Watts, Vitaliy A. Kurlin, Michael Moran, Yun Dang, Rhun Morris, Alexandra Morscher, Kate Thompson, Matthew A. Wright, Beluvalli-Eshwarappa Prasad, Frédéric Blanc, Chris M. Collins, Catriona A. Crawford, Benjamin B. Duff, Jae Evans, Jacinthe Gamon, Guopeng Han, Bernhard T. Leube, Hongjun Niu, Arnaud J. Perez, Aris Robinson, Oliver Rogan, Paul M. Sharp, Elvis Shoko, Manel Sonni, William J. Thomas, Andrij Vasylenko, Lu Wang, Matthew J. Rosseinsky, and Matthew S. Dyer. A database of experimentally measured lithium solid electrolyte conductivities evaluated with machine learning. *npj Computational Materials*, 9(1):9, 2023. doi: 10.1038/s41524-022-00951-z. URL https://doi.org/10.1038/s41524-022-00951-z.

Gabriel R Schleder, Antonio C M Padilha, Carlos Mera Acosta, Marcio Costa, and Adalberto Fazzio. From dft to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3): 032001, may 2019. doi: 10.1088/2515-7639/ab084b. URL https://dx.doi.org/10.1088/2515-7639/ab084b.

Adrien Courtois, Jean-Michel Morel, and Pablo Arias. Can neural networks extrapolate? discussion of a theorem by pedro domingos. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas*, 117(2):79, 2023. doi: 10.1007/s13398-023-01411-z.

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S. Du, Ken ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks, 2021.

Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The Journal of Physical Chemistry Letters*, 9(7):1668–1673, 2018. doi: 10.1021/acs.jpclett.8b00124. URL https://doi.org/10.1021/acs.jpclett.8b00124. PMID: 29532658.

Anthony Yu-Tung Wang, Steven K. Kauwe, Ryan J. Murdock, and Taylor D. Sparks. Compositionally restricted attention-based network for materials property predictions. *npj Computational Materials*, 7(1):77, May 2021. ISSN 2057-3960. doi: 10.1038/s41524-021-00545-1. URL https://doi.org/10.1038/s41524-021-00545-1.

Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nature Communications*, 11(1):6280, Dec 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19964-7. URL https://doi.org/10.1038/s41467-020-19964-7.

Warren E. Pickett. Colloquium: Room temperature superconductivity: The roles of theory and materials design. *Rev. Mod. Phys.*, 95:021001, Apr 2023. doi: 10.1103/RevModPhys.95.021001.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001a. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Samson Jerold Samuel Chelladurai, Kamal Upreti, Manvendra Verma, Meena Agrawal, Jatinder Garg, Rekha Kaushik, Chinmay Agrawal, Divakar Singh, and Rajamani Narayanasamy. Prediction of mechanical strength by using an artificial neural network and random forest algorithm. *Journal of Nanomaterials*, 2022:7791582, 2022. doi: 10.1155/2022/7791582. URL https://doi.org/10.1155/2022/7791582.

Janosh Riebesell. Probabilistic Data-Driven Discovery of Thermoelectric Materials. *University of Cambridge*, 2016.

Michael W. Gaultois, Anton O. Oliynyk, Arthur Mar, Taylor D. Sparks, Gregory J. Mulholland, and Bryce Meredig. Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Materials*, 4(5):053213, 05 2016. ISSN 2166-532X. doi: 10.1063/1.4952607.

Carl McBride Ellis. Extrapolation: Do not stray out of the forest! `https://www.kaggle.com/code/carlmcbrideellis/extrapolation-do-not-stray-out-of-the-forest`.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14), apr 2018. doi: 10.1103/physrevlett.120.145301. URL `https://doi.org/10.1103%2Fphysrevlett.120.145301`.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021. doi: 10.1038/s41524-021-00650-1. URL `https://doi.org/10.1038/s41524-021-00650-1`.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, apr 2019. doi: 10.1021/acs.chemmater.9b01294. URL `https://doi.org/10.1021%2Facs.chemmater.9b01294`.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. doi: 10.1109/access.2020.3031549. URL `https://doi.org/10.1109%2Faccess.2020.3031549`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL `https://arxiv.org/abs/2002.05709`.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2020. URL `https://arxiv.org/abs/2004.11362`.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. URL `https://arxiv.org/abs/2103.03230`.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings, 2022.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.336. URL `https://aclanthology.org/2022.acl-long.336`.

Gyoung S. Na and Hyun Woo Kim. Contrastive representation learning of inorganic materials to overcome lack of training datasets. *Chem. Commun.*, 58:6729–6732, 2022. doi: 10.1039/D2CC01764D. URL `http://dx.doi.org/10.1039/D2CC01764D`.

Gyoung S. Na, Seunghun Jang, and Hyunju Chang. Nonlinearity encoding to improve extrapolation capabilities for unobserved physical states. *Phys. Chem. Chem. Phys.*, 24:1300–1304, 2022. doi: 10.1039/D1CP04450H. URL `http://dx.doi.org/10.1039/D1CP04450H`.

Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2019.106949. URL `https://www.sciencedirect.com/science/article/pii/S0010465519303042`.

Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications*, 8(1):15679, 2017. doi: 10.1038/ncomms15679. URL `https://doi.org/10.1038/ncomms15679`.

Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):16028, Aug 2016. ISSN 2057-3960. doi: 10.1038/npjcompumats.2016.28.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019. doi: 10.1038/s41586-019-1335-8.

Anton O. Oliynyk, Erin Antono, Taylor D. Sparks, Leila Ghadbeigi, Michael W. Gaultois, Bryce Meredig, and Arthur Mar. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chemistry of Materials*, 28(20):7324–7331, 2016. doi: 10.1021/acs.chemmater.6b02724. URL https://doi.org/10.1021/acs.chemmater.6b02724.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001b.

Aria Mansouri Tehrani, Anton O. Oliynyk, Marcus Parry, Zeshan Rizvi, Samantha Couper, Feng Lin, Lowell Miyagi, Taylor D. Sparks, and Jakoah Brgoch. Machine learning directed search for ultraincompressible, superhard materials. *Journal of the American Chemical Society*, 140(31):9844–9853, 08 2018. doi: 10.1021/jacs.8b02717. URL https://doi.org/10.1021/jacs.8b02717.

Is domain knowledge necessary for machine learning materials properties? *Integrating Materials and Manufacturing Innovation*, 9(3):221–227, 2020. doi: 10.1007/s40192-020-00179-z. URL https://doi.org/10.1007/s40192-020-00179-z.

Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific Reports*, 8(1):17593, 2018. doi: 10.1038/s41598-018-35934-y. URL https://doi.org/10.1038/s41598-018-35934-y.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020. doi: 10.1038/s41524-020-00406-3. URL https://doi.org/10.1038/s41524-020-00406-3.

Pranjal Kumar, Piyush Rawat, and Siddhartha Chauhan. Contrastive self-supervised learning: review, progress, challenges and future research directions. *International Journal of Multimedia Information Retrieval*, 11(4):461–488, 2022. doi: 10.1007/s13735-022-00245-6. URL https://doi.org/10.1007/s13735-022-00245-6.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation, 2022b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Teddy Koker, Keegan Quigley, Will Spaeth, Nathan C. Frey, and Lin Li. Graph contrastive learning for materials, 2022. URL https://arxiv.org/abs/2211.13408.

Shufeng Kong, Francesco Ricci, Dan Guevarra, Jeffrey B. Neaton, Carla P. Gomes, and John M. Gregoire. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nature Communications*, 13(1):949, 2022. doi: 10.1038/s41467-022-28543-x. URL https://doi.org/10.1038/s41467-022-28543-x.

Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Crystal twins: self-supervised learning for crystalline material property prediction. *npj Computational Materials*, 8(1):231, 2022. doi: 10.1038/s41524-022-00921-5. URL https://doi.org/10.1038/s41524-022-00921-5.

Davide Bacciu, Federico Errica, Alessio Micheli, and Marco Podda. A gentle introduction to deep learning for graphs. *Neural Networks*, 129:203–221, 2020.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Velickovic. Principal neighbourhood aggregation for graph nets. *CoRR*, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3D infomax improves GNNs for molecular property prediction. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20479–20502. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/stark22a.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL `https://arxiv.org/abs/2006.07733`.

Sungyeon Kim, Minkyo Seo, Ivan Laptev, Minsu Cho, and Suha Kwak. Deep metric learning beyond binary supervision, 2019. URL `https://arxiv.org/abs/1904.09626`.

Steven K. Kauwe, Jake Graser, Ryan Murdock, and Taylor D. Sparks. Can machine learning find extraordinary materials? *Computational Materials Science*, 2020.

Michael W. Gaultois, Taylor D. Sparks, Christopher K. H. Borg, Ram Seshadri, William D. Bonificio, and David R. Clarke. Data-driven review of thermoelectric materials: Performance and resource considerations. *Chemistry of Materials*, Aug 2013.

Peter A. Finn, Ceyla Asker, Kening Wan, Emiliano Bilotti, Oliver Fenwick, and Christian B. Nielsen. Thermoelectric materials: Current status and future challenges. *Frontiers in Electronic Materials*, 2021.

H.M. Yuan, S.H. Han, R. Hu, W.Y. Jiao, M.K. Li, H.J. Liu, and Y. Fang. Machine learning for accelerated prediction of the seebeck coefficient at arbitrary carrier concentration. *Materials Today Physics*, 2022.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14: 199–222, 2004.

Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, and Bram Hoex. Large language models as master key: Unlocking the secrets of materials science with gpt, 2023.

KM Jablonka, P Schwaller, A Ortega-Guerrero, and Smit B. Leveraging large language models for predictive chemistry. 2023. URL `https://chemrxiv.org/engage/chemrxiv/article-details/652e50b98bab5d2055852dde`. This content is a preprint and has not been peer-reviewed.

Pytorch: An imperative style, high-performance deep learning library, 2019.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

Appendix

A Datasets

In table 3 we report a summary of different datasets of chemical properties examined in this study. The stated dataset sizes correspond to the application of the preprocessing procedure described in section 4.1 of the paper.

Table 3: Utilized datasets to benchmark the proposed approach.

Dataset name	Property	units	size	original source
seebeck	Seebeck coefficient	μV/K	403	UCSB
kappa	Thermal conductivity	W/mK (log10)	319	UCSB
tcongrmelt	Temperature for congruent melting	K	3674	MPDS
elecmass	Effective mass of electrons	m_0 (log10)	320	MPDS
bmodulus	Bulk modulus	GPa	1432	MPDS
smodulus	Shear modulus	GPa	317	MPDS
bandgap	Band gap	eV	2728	MPDS

B Ablation study

We have proposed a supervised contrastive learning method for chemical compositions in a joint context of low-data and extrapolation, common situations in materials informatics. The main results are shown in tables 1, 2 and clearly demonstrate the effectiveness of the proposed model over several state-of-the-art methods. To further investigate the impact deriving from various components in the proposed architecture, we provide a comparative analysis of the results across different design choices. In table 4 we report the obtained results under various investigated configurations. We assess the impact deriving from the presence of three main modules: contrastive learning (CL), Projection head (P), and PNA encoder (PNA). Additionally, we compare the results obtained when using exclusively the PNA encoder trained in an end-to-end fashion to directly predict the property of interest (4^{th} row in table 4). Clearly, this last case will lack both CL and Projection head. Therefore, out of the 2^3 potential settings, two are consistently omitted and therefore only 6 configurations are reported. In scenarios where CL is applied without utilizing PNA encoder (1^{st} and 2^{nd} rows in table 4), we employ a baseline encoder given by a simple MLP, that we denote as MLP_{enc}. Our goal is to explore the benefits that the more sophisticated encoder (PNA) can offer in terms of aggregating information from neighboring elements, as opposed to the absence of an aggregation scheme encountered with a straightforward MLP. Overall, we observe a similar performance between the different examined configurations. Interestingly, we note that in some cases the performance of PNA encoder alone trained in an end-to-end manner is comparable with that of the proposed method. We argue that the level of sophistication of PNA is such that it effectively extracts and processes all available information, acting as a bottleneck with respect to the subsequent CL module. In general, we expect different degrees of improvement depending on the different modules utilized as encoder (e.g., GAT [Veličković et al., 2017]). We leave as future work a thorough comparison among various encoders for the input materials.

Table 4: Ablation study for various considerations: *Mean absolute error* (MAE) and *Coefficient of determination* (R2) for each considered model and dataset. Results are averaged across 5 different random seeds. Best-performing results are shown in boldface, while second best-performing are underlined.

Module			SEEBECK	KAPPA	TCONGRMELT	ELECMASS	BMODULUS	SMODULUS	BANDGAP
CL	PNA	P							
✓		✓	$134.49_{\pm 9.78}$	$0.36_{\pm 0.02}$	$260.21_{\pm 15.86}$	$0.72_{\pm 0.03}$	$59.96_{\pm 5.44}$	$49.47_{\pm 4.29}$	$1.76_{\pm 0.00}$
✓			$145.05_{\pm 13.83}$	$0.35_{\pm 0.02}$	$270.88_{\pm 24.29}$	$0.72_{\pm 0.02}$	$60.92_{\pm 3.29}$	$50.58_{\pm 4.41}$	$1.72_{\pm 0.05}$
✓	✓	✓	$129.59_{\pm 4.98}$	$0.35_{\pm 0.02}$	$287.06_{\pm 9.15}$	$0.68_{\pm 0.01}$	$62.73_{\pm 3.19}$	$52.63_{\pm 7.27}$	$1.69_{\pm 0.05}$
	✓		$137.55_{\pm 3.59}$	$0.38_{\pm 0.01}$	$294.74_{\pm 23.88}$	$0.68_{\pm 0.01}$	$59.99_{\pm 1.9}$	$47.85_{\pm 2.05}$	$1.72_{\pm 0.01}$
✓	✓		$128.17_{\pm 11.01}$	$0.36_{\pm 0.02}$	$282.66_{\pm 7.82}$	$0.68_{\pm 0.01}$	$61.56_{\pm 3.6}$	$50.64_{\pm 4.23}$	$1.71_{\pm 0.01}$

C Implementation details

All neural network-based models have been implemented utilizing PyTorch [tor, 2019] and PyTorch Geometric [Fey and Lenssen, 2019]. CombNet's encoder module is configured with 128 input channels, 256 hidden channels, and 256 output channels, employing 3 message-passing layers. The projection head MLP_{proj} is designed as a single-layer MLP with a hidden dimension set to 512. The separate MLP_{pred} model, emloyed both as baseline and for fine-tuning contrastive-learned representations, adopts hidden dimensions [512, 256, 128, 64]. All neural networks utilize ReLU as activation function. CrabNet model is utilized with its default settings, while Ridge and SVR are implemented using the sci-kit learn package [Pedregosa et al., 2011], also with default settings.