MTENCODER: A Multi-task Pretrained Transformer Encoder for Materials Representation Learning

Thorben Prein[1,3,4,*], Elton Pan[2,*], Tom Dörr[1], Elsa Olivetti[2], Jennifer L.M. Rupp[1,2,3]

[1] Technische Universität München, [2] Massachusetts Institute of Technology
[3] TUMint. Energy Research GmbH, [4] Munich Data Science Institute
{eltonpan, elsao} @mit.edu, {t.prein, tom.doerr, jrupp} @tum.de
[*]Equal contribution

Abstract

Given the vast spectrum of material properties characterizing each compound, learning representations for inorganic materials is intricate. The prevailing trend within the materials informatics community leans towards designing specialized models that predict single properties. We introduce a *multi-task* learning framework, wherein a transformer-based encoder is co-trained across diverse materials properties and a denoising objective, resulting in robust and generalizable materials representations. Our method not only improves over the performance observed in single-dataset pretraining but also showcases scalability and adaptability toward multi-dataset pretraining. Experiments demonstrate that the trained encoder MTENCODER captures chemically meaningful representations, surpassing the performance of currrent structure-agnostic materials encoders. This approach paves the way to improvements in a multitude of materials informatics tasks, prominently including materials property prediction and synthesis planning for materials discovery.

1 Introduction

Training performant encoders has been the pivotal catalyst of recent advances in deep learning [12, 6, 5, 29].

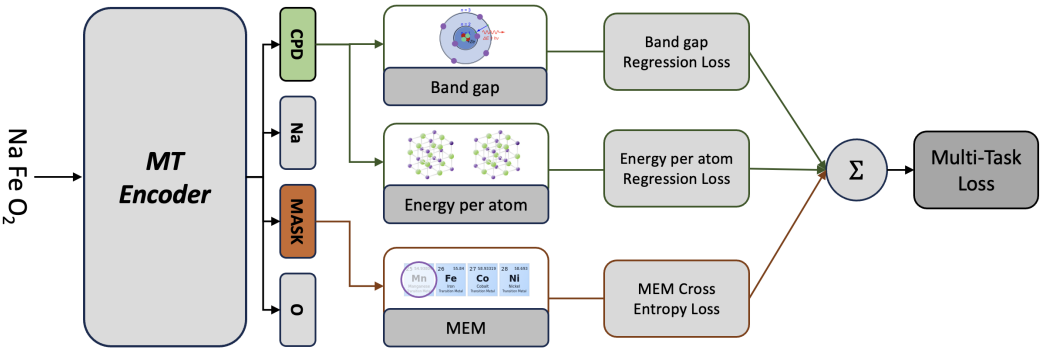

Figure 1: Combining multi-task learning with a denoising objective. MTEncoder leverages the key idea that joint, *multi-task* representation learning across datasets can overcome the issue of data fragmentation in materials science. MEM = Masked Element Modeling (details in section A.4).

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

Contemporary approaches in model development have increasingly moved away from a reliance on manually engineered features. Instead, there is a growing emphasis on extensive pretraining to learn strong representations [6, 5, 4]. This paradigm shift results in models with wide applicability across diverse downstream applications, notably illustrated by the shift from convolutional architectures to vision transformers in feature extraction for computer vision tasks. In materials science, there has been increasing interest in integrating deep learning for inorganic materials. Research can be categorized into 1) materials property prediction, which makes use of computational and experimental databases [8, 34, 25, 16, 4] and synthesis condition suggestions [23, 17, 13, 19].

Modalities Broadly, two modalities of encoding inorganic materials have emerged, structure-aware and structure-agnostic encoding. The former necessitates knowledge about the geometric arrangement of atoms within a crystal structure, while the latter relies solely on the compounds stoichiometry [34]. Given its large application space, and the inavailability of structural data in synthesis datasets, the structure-agnostic encoding remains the method of choice when constructing models for synthesis-related tasks [17, 27, 19].

Transfer learning In the central domains of deep learning, pretraining and transfer learning of large foundational models such as BERT and GPT has resulted in new state-of-the art levels of performance, coupled with eased model usability [5, 26]. On the contrary, transfer learning efforts in materials informatics remains limited due to the relatively modest size of available datasets. Typically comprising no more than a few hundred thousand datapoints [22, 14]. This has limited models to be task-specific, with each covering one or a few, task(s) [34, 8, 15]. However, while the amount of labeled data for specific tasks (eg. band gap) is limited, there exists *much more labeled data when different tasks are aggregated* (eg. band gap, shear modulus, formation energy etc). For example, there are distinct property labels per material available in databases such as OQMD [30], hence increasing the amount of labeled data by $10\times$ compared to training on single property. As such, the question lies in: *How do we leverage these disparate datasets (each with a different task) to learn strong materials representations?*

This work outlines strategies to answer the above question by leveraging multiple properties in a multi-task loss for structure-agnostic materials representation learning. This is achieved by jointly learning multiple tasks and an unsupervised objective shown in Fig. 1. Our main contributions are:

- We introduce a novel multi-task pretraining strategy for structure-agnostic material encoders
- We show that such a pretraining strategy results in strong performance on downstream tasks, outperforming state-of-the-art structure-agnostic encoders across all Matbench tasks [7]
- We open-source the pretrained models for structure-agnostic materials informatics tasks

Table 1: Average $\%$ improvement in performance on the Matbench benchmark (13 tasks) improvement using various pretraining strategies (compared to no pretraining) on the respective pretraining dataset compared to a baseline model, which is an MTEncoder with random initialization. SELF-SUPERVISED refers to masked element masking pretraining before finetuning/evaluating on Matbench. SUPERVISED refers to supervised training on a OQMD dataset before finetuning/evaluating on Matbench (details in A.3). JOINT refers to a combined SUPERVISED + SELF-SUPERVISED pretraining via a *multi-task loss function* before finetuning/evaluating on Matbench pretraining). Self-supervised performances differ due to different random splits. Best performance gain in **boldface**.

Pretraining dataset	SELF-SUPERVISED	SUPERVISED	JOINT
Formation enthalpy	+7.15	+8.50	**+11.26**
Energy per atom	+7.11	+7.48	**+10.34**
Volume per atom	+7.00	+5.72	**+8.28**
Band gap	+7.05	+4.16	**+8.10**

2 Approach

To ensure our approach is broadly applicable, especially in synthesis-related tasks, we focus on learning representations by encoding material compositions, a modality that is agnostic to structure (Fig. 1, left). Our method builds on Wang et al.'s work on encoding material compositions [34]. Notably, we significantly modify the architecture by inserting a `[CPD]` token (Fig. 1) to the start of each sequence of elements. Analogous to the `[CLS]` token introduced in [5] this token undergoes

contextualization by aggregating information from the compound constituent elements, accumulating a holistic representation for the specific composition. We find this approach to boost performance compared to alternatives such as the weighted contributions introduced in CrabNet (Table 2). In our approach, we train the model with compact, task-specific two-layered MLP heads analyzing the computed `[CPD]` representations. This encourages the model to refine its compound representations, ensuring they remain informative across tasks.

2.1 Multi-task pretraining

Multi-task learning has been shown to improve performance in several domains including natural language processing and computer vision [2, 36, 18]. However, conflicting gradients remain a problem when training on multiple objectives.

Loss weighting To overcome conflicting gradients, we use RLW (Random Loss Weighting) [24, 37]. This strategy randomly samples task-specific loss weights from a constrained Bernoulli distribution during each parameter update [24]. This has been shown to be a robust alternative to more expensive gradient surgery methods. Considering all tasks T, each with a corresponding dataset D_t. The overall loss is calculated by summing over the task-specific losses l_t, weighted by weights λ_l (equation 1). Finally an update for the model parameters θ is calculated. Most importantly during each update of model parameters, the task-specific loss weights are resampled from a constrained Bernoulli distribution $p(\tilde{\lambda})$. When training the self-supervised objective with the supervised tasks we apply a constant scaling factor to align loss amplitudes between L1 and cross-entropy loss.

$$L(\theta) = \sum_{t=1}^{T} \lambda_l l_t(D_t; \theta) \tag{1}$$

Evaluation To validate pretraining approaches, various domains have established robust benchmarking frameworks, as previously exemplified by the GLUE benchmark [33] for natural language processing. Central to these benchmarks is their ability to analyze a model's performance across a diverse set of pivotal tasks within the target domain, essentially investigating its versatility. In materials science, the Matbench task-suite has been introduced [7].

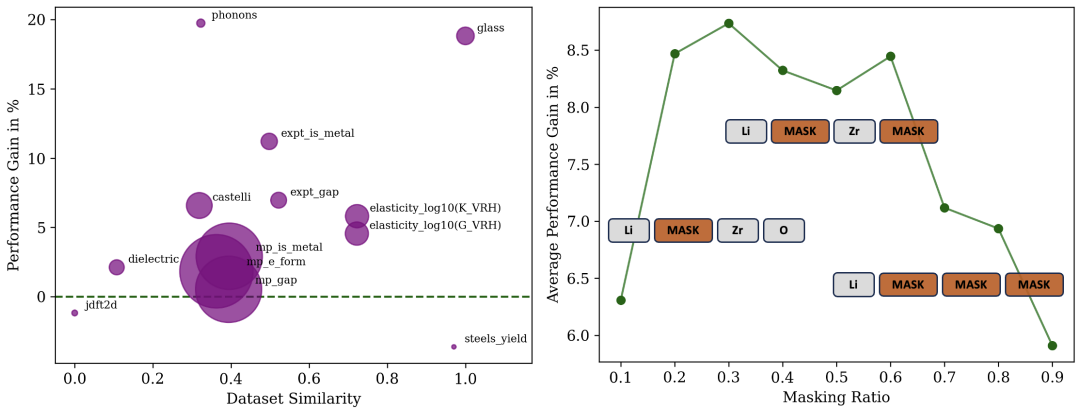

Figure 2: *left:* % improvement in performance (pretraining on energy per atom) after training on the Matbench datasets shown in the plot. As shown, an increase in dataset similarity between pretraining dataset and finetuning data is correlated with increase in performance gain. Dataset similarity is measured using the earth mover's distance [10]. *right:* % improvement in performance after masked element modeling (MEM). Optimal masking ratio is around 0.3, which constitutes roughly 1 masked element in a ternary compound.

Matbench encompasses a varied range of tasks, spanning from regression to classification objectives, and includes both computed and experimental data. In line with methodologies from the previously mentioned domains, we evaluate our pretraining strategies by training on all 13 Matbench tasks. We compute the performance gain (PG) across tasks t by comparing a model's performance l_t^{MD} over a

baseline model's performance l_t^{BL} for each task: $PG = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{l_t^{\text{BL}} - l_t^{\text{MD}}}{l_t^{\text{BL}}}\right)$. Here, we compare our models to CrabNet [34], a state-of-the-art architecture for structure-agnostic materials property tasks, with random initialization as our baseline model.

3 Results

3.1 Comparison of pretraining strategies

Setup We compare 3 distinct pretraining strategies (SUPERVISED, UNSUPERVISED, JOINT) on specific OQMD properties (1st column in Table 1). SELF-SUPERVISED refers to masked element modeling (refer to Section A.4). SUPERVISED refers to supervised training on OQMD properties (refer to Section A.3). JOINT indicates a combined SELF-SUPERVISED + SUPERVISED pretraining via a *multi-task loss*. After pretraining, we evaluate the resultant pretrained models by finetuning and testing on 13 property prediction tasks on Matbench as described above (refer to Section 2.1).

Results Table 1 shows the outcome and average % improvement in performance on 13 Matbench tasks compared to no pretraining. Overall, all 3 pretraining strategies give marked improvements in performance ranging from 4 – 11 % vs. no pretraining. SELF-SUPERVISED (i.e. MEM) results in more consistent performance gains (lower variance across pretraining datasets) compared to SUPERVISED. Interestingly, JOINT takes pretraining a step further by combining both pretraining strategies via a *multi-task loss* (Fig. 1), resulting in the improvements across all pretraining datasets (in **bold**) over the former 2 stand-alone strategies. This empirically shows that multi-task pretraining is beneficial to model performance in material property prediction.

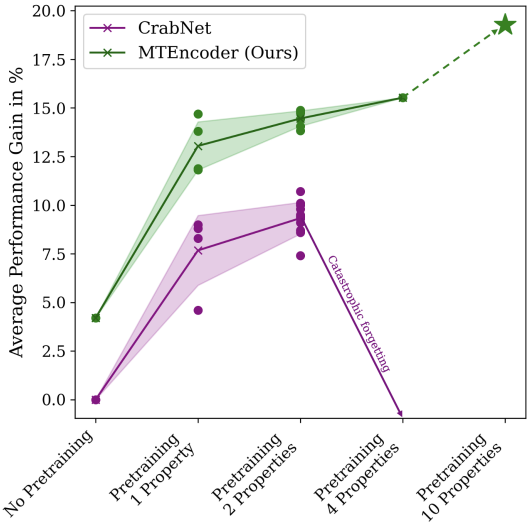

Figure 3: Average % improvement in performance on the Matbench benchmark (13 tasks) improvement as the number of pretraining properties increases (compared to no pretraining) using our JOINT pretraining strategy. CrabNet = Current state-of-the-art model. MTENCODER significantly outperforms CrabNet at all settings up to 4 datasets. Note: The green star refers to a larger MTENCODER model with 2× the number of transformer encoder layers and attention heads.

3.2 Scaling

Setup We investigate the performance of MTENCODER as we scale the number of pretraining datasets. Here, we use the 4 OQMD properties shown in Section 3.1. For pretraining on ≥2 datasets, we consider all possible combinations. For instance, for pretraining on 2 (out of 4 OQMD) datasets, we perform ${}^4C_2 = 6$ different runs (each with a unique set of 2 pretraining datasets) as shown in Fig. 3. In the case of 4 datasets with CrabNet, 3 out of the 24 possible combinations for sequential pretraining are sampled at random. It is worth noting that since the multi-task loss is permutation invariant with respect to tasks, we consider combinations instead of permutations (of pretraining datasets). Additionally, we introduce a form of extreme dataset scaling regimen where MTENCODER is pretrained with 10 OQMD materials properties (star in Fig. 3). Similar to Section 3.1, we evaluate on the 13 Matbench tasks after each pretraining.

Results We benchmark our model against CrabNet, a state-of-the-art structure-agnostic materials encoder in Fig. 3 left. For MTENCODER performance gains increase as the number of pretraining datasets increases, achieving up to 14 % gain (vs. baseline CrabNet) after pretraining with 2 datasets.

MTENCODER exhibits notable performance in transfer learning across all regimes. Its performance increase starts at 4 % without pretraining, and the gap widens further with pretraining. The performance of pretrained CrabNet seems to plateau, with suspected negative transfer upon integrating four datasets, decreasing performance below the baseline — a phenomenon we attribute to potential catastrophic

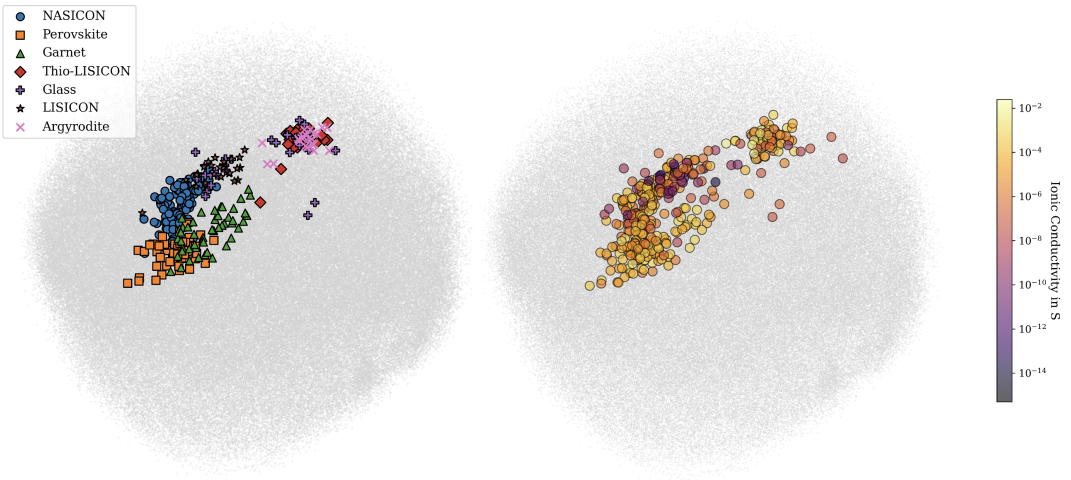

Figure 4: *left:* Visualization (PCA) of the chemical space learned by MTENCODER. Here, we focus on most prominent solid-state electrolyte materials families which show distinct clusters. Grey points refer to all materials from the OQMD database. *right:* Learned MTENCODER embeddings capture trends in ionic conductivity (continuous space showing regions of high and low values) for solid-state electrolytes. Note: The model has not been pretrained or finetuned on compounds in the ionic conductivity dataset.

forgetting [21, 9]. These results underscore the tailored architecture of MTENCODER, as elaborated in Section A.2. Specifically, it highlights its capacity to effectively leverage extensive pretraining datasets and learn better representations relative to current models. Furthermore, scaling up both the number of pretraining datasets to 10 *and* model size, results in further increase in performance of close to 20% over the baseline (star in Fig. 3), indicating that future work should involve scale-up in both aspects.

3.3 Visualization of embedding space

In Fig. 4 left, we visualize the learned representations of materials by MTENCODER pretrained on 4 OQMD properties using the JOINT approach. Distinct clusters corresponding to different solid electrolytes can be observed, as catalogued in the Liverpool ionics dataset [11], which is a representative database assembling more than 400 described solid electrolyte materials. We observe that MTENCODER is able to capture chemically meaningful representations. For instance, in the upper left corner, NASICON type materials are identified, while in the lower left, perovskites are grouped. Finally, we visualize the embeddings of the same compounds with respect to their corresponding ionic conductivities in Fig. 4 right. The observed smooth, continuous space with respect to the property of interest indicates that MTENCODER representations may have potential applications in materials property optimization.

4 Conclusion

We introduce MTENCODER, a model pretrained with a novel pretraining strategy that employs a multi-task loss to simultaneously leverage multiple disparate materials properties, augmented with a denoising objective. Clearly, our experiments suggest that pretraining with larger amounts of labeled materials data from a diverse set of tasks will likely further enhance the performance on downstream material property prediction tasks. Following this strategy MTENCODER has shown to systematically improve pretraining compared to state-of-the-art models on established benchmarks. Importantly, our approach potentially enhances sample efficiency, reducing the domain-specific data needed for fine-tuning — a significant challenge in applying deep learning to materials discovery. Furthermore, MTENCODER learns chemically meaningful representations of materials, both in structure and property space. It is hoped that this work contributes to inorganic materials discovery, particularly in terms of predicting the properties and synthesis routes of new energy materials [20, 3]. Future work will focus on scaling up pretraining, and evaluating the potential of MTENCODER in informing composition-property-synthesis relationships of inorganic materials.

References

[1] Matbench: A benchmarking platform for materials science. https://matbench.materialsproject.org/, 2023. Accessed: 2023-09-30.

[2] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*, 2021.

[3] Moran Balaish, Juan Carlos Gonzalez-Rosillo, Kun Joong Kim, Yuntong Zhu, Zachary D Hood, and Jennifer LM Rupp. Processing thin but robust electrolytes for solid-state batteries. *Nature Energy*, 6(3):227–239, 2021.

[4] Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: self-supervised transformer model for metal–organic framework property prediction. *Journal of the American Chemical Society*, 145(5):2958–2967, 2023.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010.

[7] 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.

[8] Rhys EA Goodall and Alpha A Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature communications*, 11(1):6280, 2020.

[9] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[10] Cameron J Hargreaves, Matthew S Dyer, Michael W Gaultois, Vitaliy A Kurlin, and Matthew J Rosseinsky. The earth mover's distance as a metric for the space of inorganic compositions. *Chemistry of Materials*, 32(24):10610–10620, 2020.

[11] 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, et al. A database of experimentally measured lithium solid electrolyte conductivities evaluated with machine learning. *npj Computational Materials*, 9(1):9, 2023.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[13] Haoyan Huo, Christopher J Bartel, Tanjin He, Amalie Trewartha, Alexander Dunn, Bin Ouyang, Anubhav Jain, and Gerbrand Ceder. Machine-learning rationalization and prediction of solid-state synthesis conditions. *Chemistry of Materials*, 34(16):7323–7336, 2022.

[14] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[15] Dipendra Jha, Kamal Choudhary, Francesca Tavazza, Wei-keng Liao, Alok Choudhary, Carelyn Campbell, and Ankit Agrawal. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature communications*, 10(1):5316, 2019.

[16] 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.

[17] Christopher Karpovich, Elton Pan, Zach Jensen, and Elsa Olivetti. Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction. *Chemistry of Materials*, 35(3):1062–1079, 2023.

[18] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[19] Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, et al. Inorganic materials synthesis planning with literature-trained neural networks. *Journal of chemical information and modeling*, 60(3):1194–1201, 2020.

[20] Kun Joong Kim, Moran Balaish, Masaki Wadaguchi, Lingping Kong, and Jennifer LM Rupp. Solid-state li–metal batteries: challenges and horizons of oxide and sulfide solid electrolytes and their interfaces. *Advanced Energy Materials*, 11(1):2002689, 2021.

[21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[22] Olga Kononova, Tanjin He, Haoyan Huo, Amalie Trewartha, Elsa A Olivetti, and Gerbrand Ceder. Opportunities and challenges of text mining in materials research. *Iscience*, 24(3), 2021.

[23] Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*, 6(1):203, 2019.

[24] Baijiong Lin, Feiyang Ye, Yu Zhang, and Ivor W Tsang. Reasonable effectiveness of random weighting: A litmus test for multi-task learning. *arXiv preprint arXiv:2111.10603*, 2021.

[25] Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Liu, and Jianjun Hu. Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22(32):18141–18148, 2020.

[26] OpenAI. Introducing chatgpt. `https://openai.com/blog/chatgpt`, November 2022. Accessed: 2023-12-03.

[27] Elton Pan, Christopher Karpovich, and Elsa Olivetti. Deep reinforcement learning for inverse inorganic materials design. *arXiv preprint arXiv:2210.11931*, 2022.

[28] PyTorch Development Team. torch.nn.transformerencoder — pytorch 1.11.0 documentation, 2023. Accessed: 2023-12-02.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[30] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *JOM*, 65:1501–1509, 2013. Cited over 100 times.

[31] 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.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[34] 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, 2021.

[35] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

[36] Joseph Worsham and Jugal Kalita. Multi-task learning for natural language processing in the 2020s: where are we going? *Pattern Recognition Letters*, 136:120–126, 2020.

[37] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

[38] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124, 2021.