Supplementary Material

Accelerating materials property prediction via a hybrid Transformer Graph framework that leverages four body interactions

Mohammad Madani[1,2], Valentina Lacivita[3], Yongwoo Shin[3*], and Anna Tarakanova[1,4*]

[1]School of Mechanical, Aerospace, and Manufacturing Engineering, University of Connecticut, Storrs, CT, United States

[2]Department of Computer Science & Engineering, University of Connecticut, Storrs, CT, United States

[3]Samsung Semiconductor Inc, Cambridge, MA, United States

[4]Department of Biomedical Engineering, University of Connecticut, Storrs, CT, United States

[*]Corresponding authors: yongwoo.s@samsung.com

anna.tarakanova@uconn.edu

Datasets

Phonons

The dataset contains 1,265 crystal structure files and has been adapted from the work done by Petretto et al[1] and derived from the Materials Project[2]. The target variable for the dataset is the frequency of the highest frequency optical phonon mode peak (1/cm).

Dielectric

The dataset contains 4,764 crystal structure files. The property predicted in this dataset is the refractive index of the material.

G_{VRH}

The dataset consists of 12,086 structural files and the DFT calculated $\log_{10} G_{VRH}$ - shear modulus of the structures.

K_{VRH}

The dataset consists of 12,086 structural files and the DFT calculated $\log_{10} K_{VRH}$ - bulk modulus of the structures.

Fermi Energy

The Fermi Energy dataset consists of 26,447 structural files from the Materials Project and the DFT calculated Fermi energies.

Formation Energy

The Formation Energy (FE) dataset consists of 126,785 structural files from the Materials Project and the DFT calculated Formation Energy.

Band Gap

The Band Gap (BG) dataset consists of 126,728 structural files from the Materials Project and the DFT calculated Band Gap. Most of the structural files in this dataset are possibly a subset of the larger band gap dataset in the MatBench suite as both datasets are derived from the Materials Project repository.

Energy above the convex hull

The dataset has been adapted from Materials Project. It contains 126,352 crystal structures with a target variable of predicting the energy above the convex hull in eV/atom.

Tables

Model	Description
CrysCo	Trained on compositions, human-extracted physical properties and structure information
CoTAN	Trained only on compositions and human-extracted physical properties
CrysGNN	Trained including chemical species and structure information
EGAT	Edge Gated Attention neural network

Table S 1 Summary and description of our model's components. The human-extracted physical properties are listed in Table S .

CrysCo configuration	Value
EGAT layer	10 layers
GCN	2 layers
Attention layers	8 layers
Global attention layers	1 layer
Edge input	120
Node input features	80
Embedding features	64
Input dimension element embedding	200
Dimension for positional encoder	512
Feedforward dimension for self-attention	1024
Number of attention heads per attention layer	6
Hidden features	128

Table S 2 Final model configuration and hyperparameters used for CrysCo and selected via Grid search algorithms.

Dataset name	CrysCo	CrysCo_Egg	CrysCo_3body	ALIGNN-D	ALIGNN
E_f	**0.024±0.005**	0.025±0.015	0.026±0.013	<u>0.024±0.007</u>	0.025±0.008
Elasticity_log10(G_{VRH})	**0.079±0.002**	0.081±0.002	0.080±0.013	**0.079±0.004**	**0.079±0.002**
Elasticity_log10(K_{VRH})	**0.058±0.001**	0.061±0.003	0.063±0.007	<u>0.061±0.005</u>	0.062±0.001
E_{Hull}	**0.023±0.001**	0.026±0.002	<u>0.024±0.006</u>	0.025±0.008	0.026±0.002

Table S 3 MAE and corresponding MSD associated to CrysCo and compared with the same base model but trained with different configurations: CrysCo_Egg = trained replacing the EGAT component with Edge Gated Graph convolution, Crysco_3body = trained including only up to 3-body interactions. Comparisons are made also with the performance of ALIGNN[3] and ALIGNN-D[4] which include 3-body and 4-body interactions, respectively. The best performance is shown in boldface and the second best performance is underlined. MAE and MSD are calculated over 5-fold cross-validations. CrysCo shows the highest accuracy across all considered properties.

Dataset name	CoTAN	Roost	CrabNet
E_f	0.038±0.009	**0.034 ± 0.002**	0.035±0.005
Mp_bgap	0.403±0.009	0.388±0.011	**0.384±0.086**
Phonons	**48.88±11.86**	49.508±12.086	57.45±16.93
Elasticity_log10(G_{VRH})	0.107±0.011	0.108±0.074	**0.097±0.01**
Elasticity_log10(K_{VRH})	0.077±0.0053	0.073±0.016	**0.072±0.002**
E_{Hull}	**0.087±0.027**	0.094±0.092	0.087±0.033
Fe_energy	0.096±0.0051	**0.093±0.088**	0.110±0.079
Dielectric	0.404±0.016	**0.388±0.090**	0.408±0.034

Table S 4 MAE and corresponding MSD associated to CoTAN trained exclusively on compositional features (i.e., without human-extracted features) in comparison to the composition-based baselines Roost[5] and CarbNet[6] on regression benchmarks. The best performance is shown in boldface and the second best performance is underlined. MAE and MSD are calculated over 5-fold cross-validations.

Comparing our model's performance with JMP models

In general, it can be seen from Table S5 that, in addition to being computationally more efficient, although large version of JMP, namely JMP-L (~230M parameters), outperforms our model for all properties significantly.

Given this, it is even more important to underline the superior computational efficiency of CrysCo. Indeed, the pre-training of JMP-L on 128x V100 32GB GPUs for 2 epochs required around 34,400 GPU hours, whereas CrysCo achieved similar performance with only 1,200 GPU hours/epoch. Moreover, the significant difference in the number of trainable parameters is noteworthy, with CrysCo having only 2.2M parameters compared to about 235M parameters in JMP-L. In summary, CrysCo proves to be not only highly accurate but also much more computationally efficient, making it a practical and resource-efficient choice for material property predictions.

Dataset name	CrysCo	JMP-L
E_f	0.024±0.005	**0.020±0.007**
Mp_bgap	0.188±0.009	**0.176±0.004**
Phonons	29.75±9.62	**27.32±2.01**
Elasticity_log10(G)	0.079±0.002	**0.066±0.002**
Elasticity_log10(K)	0.058±0.001	**0.052±0.003**
E_{Hull}	0.023±0.001	**0.021±0.004**
Fe_energy	0.020±0.001	**0.016±0.004**
Dielectric	0.318±0.055	**0.277±0.028**

Table S 5 MAE and corresponding MSD associated to CrysCo in comparison to the latest Joint Multi-domain Pre-training (JMP) strategy[7] using a GemNet-OC backbone model for its effectiveness across various chemical domains and scales. Comparisons are made with JMP model ("large", ~230M parameters) pre-trained models. The best performance is shown in boldface and the second best performance is underlined. MAE and MSD are calculated over 5-fold cross-validations.

Dataset (size of dataset)	CrysCo (MAE)	CrysCoT-MP_eform (MAE)
1. Expt. E_F(1709)	0.1294±0.005	0.0891±0.0083
2. E_{exfol}(636)	54.82±7.12	51.41±8.91
3. d_{33}(941)	0.219±0.062	0.201±0.038
4. PhonDOS-Peak(1265)	0.115±0.017	0.102±0.007
5. $2D-E_F$(633)	0.111±0.038	0.100±0.029
6. $2D-E_{g\prime}.opt$(522)	0.559±0.096	0.538±0.16
7. $2D-E_{g\prime}.Tbmbj$(120)	1.106±0.31	1.07±0.21
8. A^U(1181)	3.232±0.0576	3.11±0.0209
9. Log(∈ ∞)(1296)	0.149±0.028	0.144±0.012
10. Log(∈ total)(1296)	0.219±0.034	0.218±0.0124
11. Poisson ratio(1181)	0.0287±0.006	0.0282±0.007
12. Expt. E_g(2481)	0.376±0.040	0.344±0.034
13. $E_{g\prime}.Tbmbj$(7348)	0.352±0.026	0.348±0.016
14. Steel strength (312)	41.53±12.41	41.22 ± 15.33

Table S 6 Additional benchmarking for 14 additional data scarce material properties. Definitions: 1. Experimental formation enthalpies (eV/atom) from Matminer's expt_formation_enthalpy and expt_formation_enthalpy kingsbury datasets; 2. Exfoliation energies (meV/atom) from Matminer's jarvis_dft_2d dataset; 3. Piezoelectric modulus from Matminer's piezoelectric_tensor dataset; 4. Highest frequency optical phonon mode peak (cm−1) from Matminer's matbench_phonons dataset; 5. Formation energies (eV/atom) from Matminer's jarvis_dft_2d dataset; 6. Band gap of 2D materials (eV) from Matminer's jarvis_dft_2d dataset, calculated with the OptB88vDW DFT functional; 7. Band gap of 2D materials (eV) from Matminer's jarvis_dft_2d dataset, calculated with the TBMBJ DFT functional; 8. Elastic anisotropy index from Matminer's elastic_tensor_2015 dataset; 9. Electronic contribution to dielectric constant from Matminer's phonon_ dielectric_mp dataset; 10. Dielectric constant from Matminer's phonon_dielectric_mp dataset; 11. Poisson ratio from Matminer's elastic_tensor_2015 dataset; 12. Experimental bandgaps (eV) from Matminer's expt_gap_kingsbury dataset; 13. Band gap (eV), calculated with the TBMBJ DFT functional, from Matminer's jarvis_dft_3d dataset.

Property
1. Group Number
2. Period Number
3. Electronegativity
4. Covalent Radius
5. Valence Electron
6. First Ionization energy
7. Electron Affinity
8. Block
9. Atomic Volume

Table S 7 Chemical Features used in the atomic graph of CrysGNN.

Full name of descriptors
1. Mean absolute deviation in relative bond length
2. Maximum relative bond length
3. Minimum relative bond length
4. Maximum neighbor distance variation
5. Minimum neighbor distance variation
6. Range neighbor distance variation
7. Mean neighbor distance variation
8. Standard deviation neighbor distance variation
9. Mean average bond length.
10. Standard deviation average bond length
11. Mean absolute deviation in atomic volume.
12. Mean average bond angle.
13. Mean coordination number determined by the Voronoi method.
14. Standard deviation coordination number determined by the Voronoi method
15. Structural complexity per atom
16. Structural complexity number of atoms per primitive cell
17. The largest lattice length of the primitive cell
18. The second largest lattice length of the primitive cell
19. The smallest lattice length of the primitive cell
20. The largest lattice angle of the primitive cell
21. The second largest lattice angle of the primitive cell
22. The smallest lattice angle of the primitive cell
23. Crystal system

Table S 8 List of human extracted descriptors. The primitive cell was determined using Niggli reduction implemented in the Structure class in pymatgen, while other descriptors were calculated using matminer.

Local atomic environment properties	CGCNN	ALIGNN	CrysCo
Mean absolute deviation in relative bond length	0.97	0.98	**0.98**
Maximum relative bond length	0.75	0.82	**0.84**
Minimum relative bond length	0.91	**0.93**	0.91
Maximum neighbor distance variation	0.92	**0.95**	0.93
Minimum neighbor distance variation	0.94	0.96	**0.96**
Range neighbor distance variation	0.90	**0.96**	0.95
Mean neighbor distance variation	0.98	0.97	**0.98**
Standard deviation neighbor distance variation	0.92	0.95	**0.97**
Mean average bond length	0.88	0.90	**0.94**
Standard deviation average bond length	0.91	**0.94**	0.93
Mean absolute deviation in atomic volume	0.94	0.94	**0.97**
Mean average bond angle	0.83	**0.87**	0.86
Standard deviation average bond angle	0.74	0.81	**0.82**
Mean Voronoi coordination number	0.68	0.83	**0.83**
Standard deviation Voronoi coordination number	0.81	0.89	**0.89**
Structural complexity per atom	**0.91**	0.86	0.90

Table S 9 Performance (R2 scores) of CrysCo on local atomic environments predictions, as compared to CGCNN[8] and ALIGNN[3]. The best performance is highlighted in boldface. (Note: this table is directly obtained from recent work[9])

Global structural descriptors	CGCNN	ALIGNN	MEGNet	Matformer	This work	
					CrysGNN	CrysCo
Structural complexity per primitive cell	0.33	0.67	0.46	**0.79**	0.65	0.76
Density	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
Volume per atom	0.80	0.94	0.85	**0.96**	0.91	0.93
Packing fraction	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
Num. atoms per primitive cell	0.77	0.89	0.82	**0.87**	0.82	0.85
a	0.14	0.43	0.24	**0.72**	0.46	**0.66**
b	0.41	0.67	0.54	0.73	0.69	**0.74**
c	0.54	0.79	0.67	0.83	0.79	**0.86**
α	0.15	0.13	0.11	**0.39**	0.17	0.36
β	0.13	0.15	0.11	**0.28**	0.17	0.25
γ	0.10	0.08	-0.05	0.15	0.12	**0.21**

Table S 10 Performance (R2 scores) of CrysCo and CrysGNN on global structural descriptors calculated with matminer[10] and pymatgen[11]. Comparisons are made with CGCNN[8], ALIGNN[3], MEGNet[12], and Matformer[13]. The best performance is highlighted in boldface. (Note: this table is directly obtained from recent work[9])

Figures

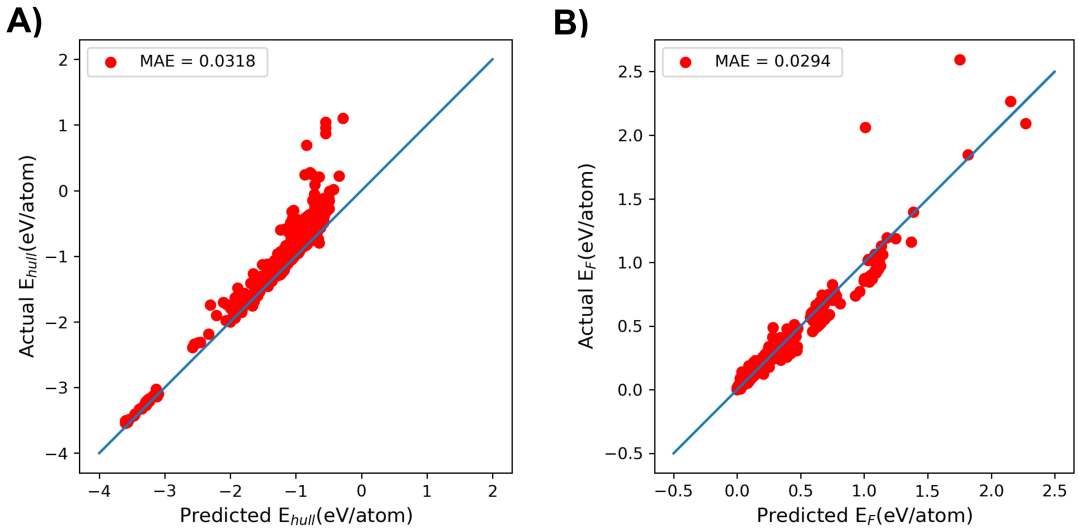

Figure S1 Performance of CrysCo on E_f (A) and E_{Hull} (B) predictions on a novel dataset of in-house generated and DFT-relaxed structures.

Figure S2 shows a comparison of the representations derived from CrysCo (a) and from ALIGNN (b) for the E_{Hull} dataset. Each point in the plot represents a material and is color-coded based on its E_{Hull} value (eV/atom), which is the label that the models were fine-tuned on. We observe that the t-SNE projection from the CrysCo model provides a better clustering of the materials compared to ALIGNN. Specifically, thermodynamically stable or metastable materials, i.e., materials with low E_{Hull} (lower than 0.05 eV/atom), are mainly clustered at the top right of the t-SNE plot (Figure S2a, green-to-blue color region), whereas materials with high E_{Hull} (higher than 0.5 eV/atom) are mainly clustered at the bottom of the plot (Figure S2a, yellow-to-red color region). For instance, widely utilized battery cathode ($LiCoO_2$: mp-22526) and semiconductor materials (GaAs: mp-2534, HfO_2: mp-352, SiO_2: mp-546794) are found in close proximity in the t-SNE projection from CrysCo, while they are spread over the entire t-SNE projection from ALIGNN. This finding suggests that the representations learned by CrysCo are more generalizable than those learned by ALIGNN. As a consequence, we expect the representations generated by our GNN framework to have the potential to be utilized for the characterization of materials from a vast chemical space.

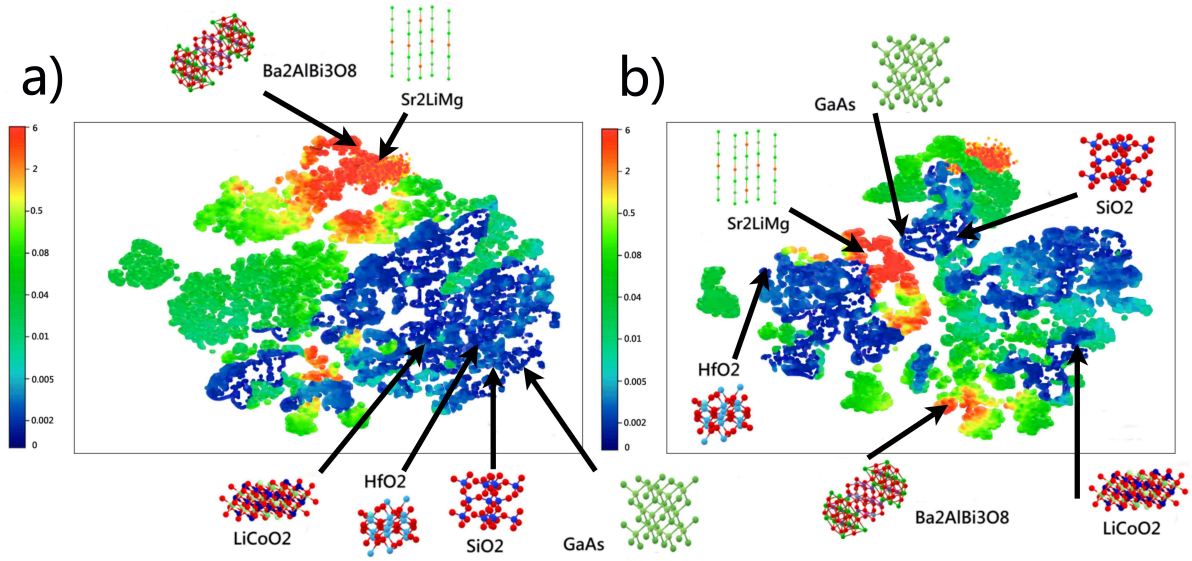

Figure S2 Embedding space for the E_{Hull} dataset visualized using t-SNE. Colors are mapped to E_{Hull} values (eV/atom) as indicated in the color bar. (a) The t-SNE plot for the embedding generated from our CrysCo model. (b) The t-SNE plot for the embedding generated from the ALIGNN model. Six compounds are highlighted to compare the clustering capabilities of the two models.

References

1 Petretto, G. *et al.* High-throughput density-functional perturbation theory phonons for inorganic materials. *Scientific Data* **5**, 180065, doi:10.1038/sdata.2018.65 (2018).

2 Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **1** (2013).

3 Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials* **7**, 185 (2021).

4 Hsu, T. *et al.* Efficient and interpretable graph network representation for angle-dependent properties applied to optical spectroscopy. *npj Computational Materials* **8**, 151 (2022).

5 Goodall, R. E. & Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature communications* **11**, 6280 (2020).

6 Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials* **7**, 77 (2021).

7 Shoghi, N. *et al.* From molecules to materials: Pre-training large generalizable models for atomic property prediction. *arXiv preprint arXiv:2310.16802* (2023).

8 Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **120**, 145301 (2018).

9 Gong, S. *et al.* Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. *Science Advances* **9**, eadi3245 (2023).

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

11 Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314-319 (2013).

12 Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials* **31**, 3564-3572 (2019).

13 Yan, K., Liu, Y., Lin, Y. & Ji, S. Periodic graph transformers for crystal material property prediction. *Advances in Neural Information Processing Systems* **35**, 15066-15080 (2022).