Supplementary Information

Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals

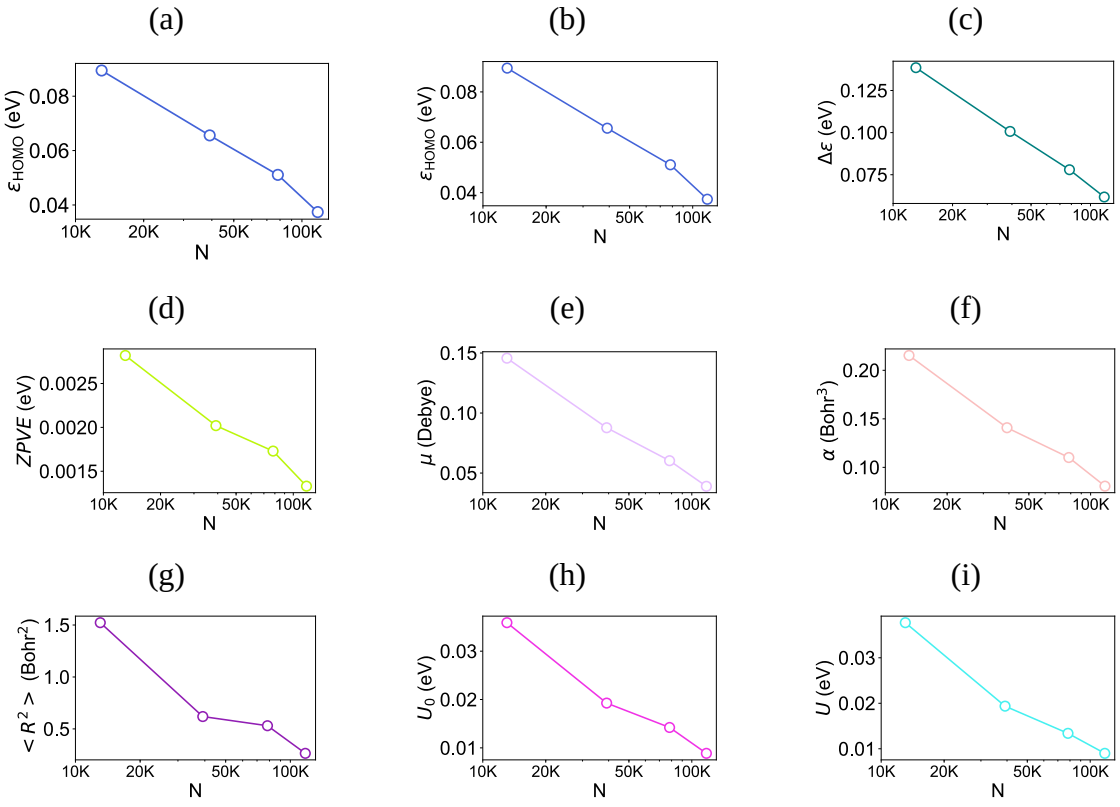
Chi Chen[1], Weike Ye[1], Yunxing Zuo[1], Chen Zheng[1] & Shyue Ping Ong[1]*

[1] Department of NanoEngineering, University of California San Diego, 9500 Gilman Dr, Mail Code, 0448, La Jolla, CA 92093-0448, United States. Correspondence and requests for materials should be addressed to S.P.O. (email: ongsp@eng.ucsd.edu)

1. Molecules

Convergence of 13 targets

The convergence plots are obtained by changing the number of training data points and keeping the validation and test data set fixed.



[1] Department of NanoEngineering, University of California San Diego, 9500 Gilman Dr, Mail Code, 0448, La Jolla, CA 92093-0448, United States

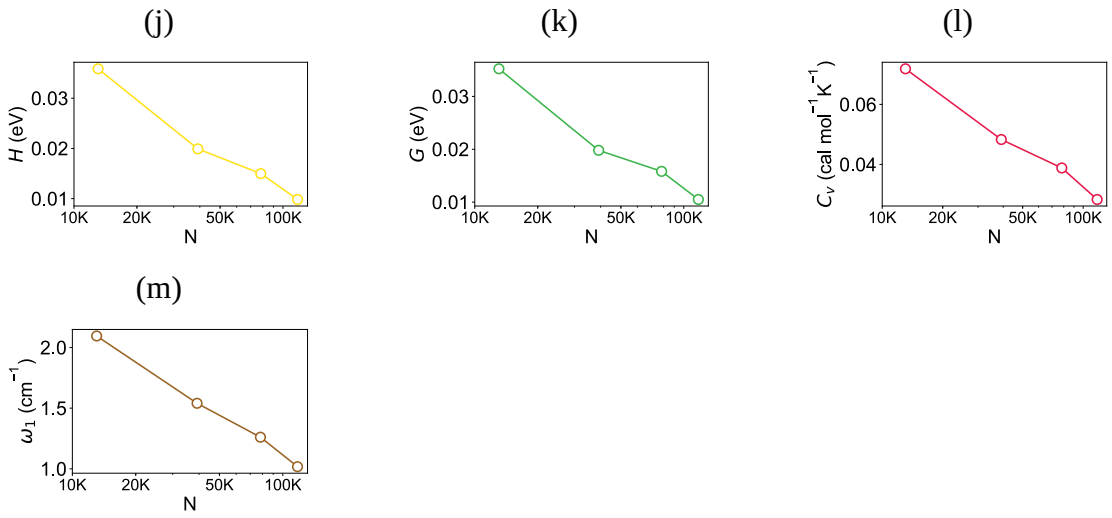
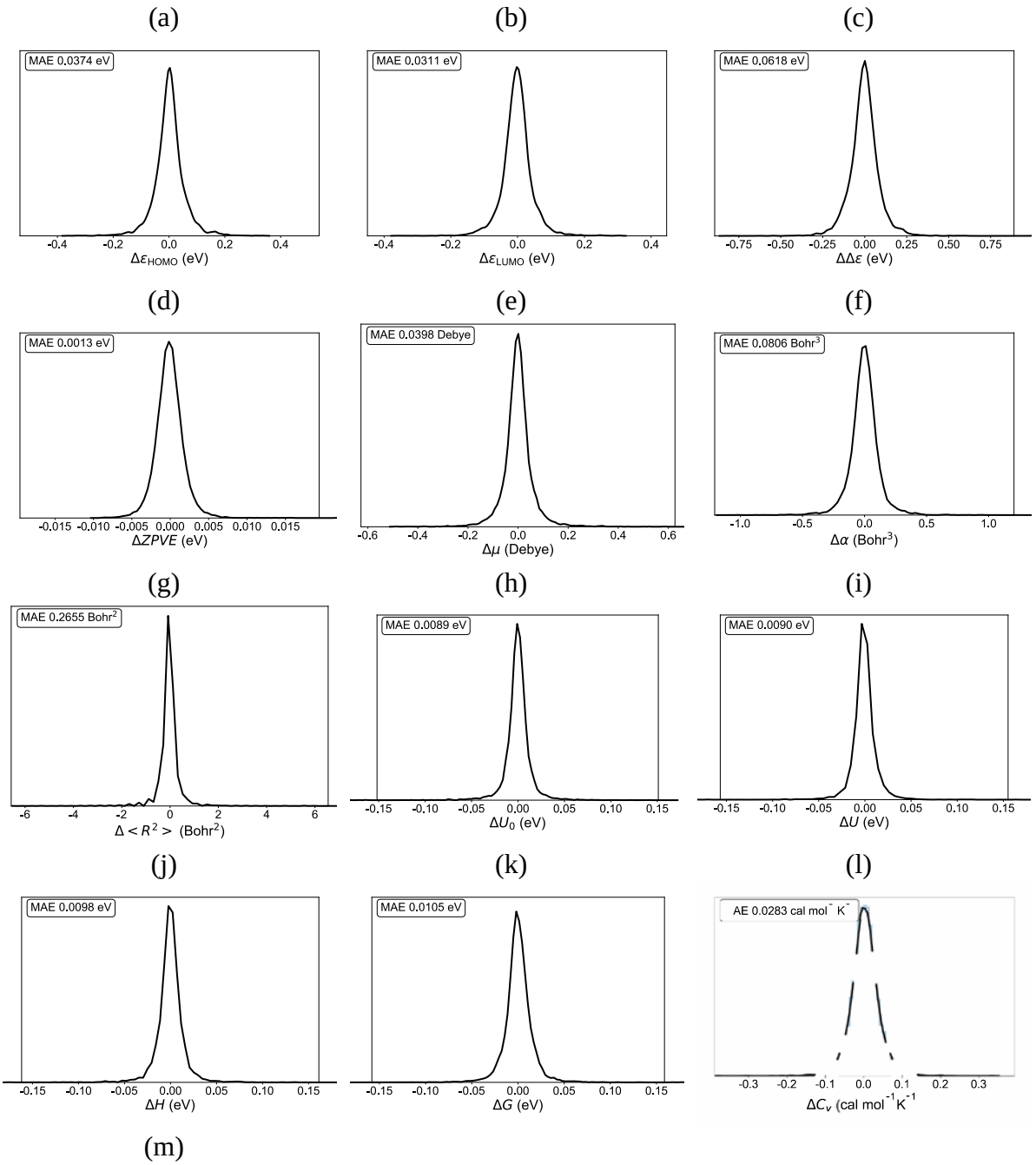


Figure S1. Convergence plots of MEGNet models on the 13 targets in QM9 dataset.

Test error distribution

The distributions of various targets in the QM9[1] test data set are shown in Figure S2. One consistent outlier (qm9: 079919) with relatively large error is found in predicting the energy-related quantities. Overall, the errors for the 13 targets are homoscedastic and exhibit Gaussian-like distributions (Figure S2).



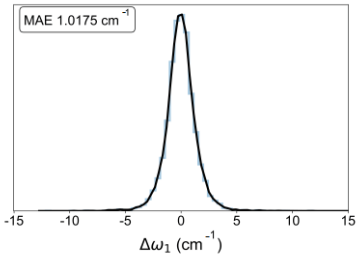


Figure S2. Test error distributions on 13 targets of the QM9 datasets.

Extrapolation of the single-task molecule free energy model at different temperatures

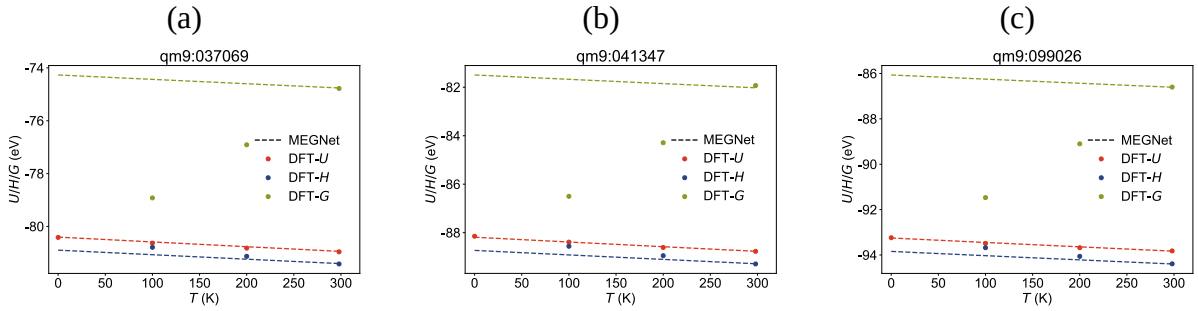


Figure S3. Extrapolation of the single-task model for predicting U, H and G at various temperatures.

The single-task molecule free energy model is trained on *U* at 0 K, *U* at 298 *K*, *H* at 298 K and *G* at 298 K. In Figure S3, we randomly select three molecules from our test dataset and calculate the *U*, *H* and *G* at 100, 200 and 298 K. We also show the model prediction results within the range [0 K, 298 K]. The predictions on *U* match well with the DFT calculations in the [0 K, 298 K] range. However, the predicted *H* and *G* deviates significantly from the DFT values at 100 K and 200 K, and exhibit a linear relationship from 0 K to 298 K, similar to the trend for *U*. This is due to the lack of training data for H and G at these intermediate temperatures, and the model only learns shifts *H*(298)-*U*(298) and *G*(298)-*U*(298) for *H* and *G* respectively.

2. Crystalline materials

Formation energy convergence

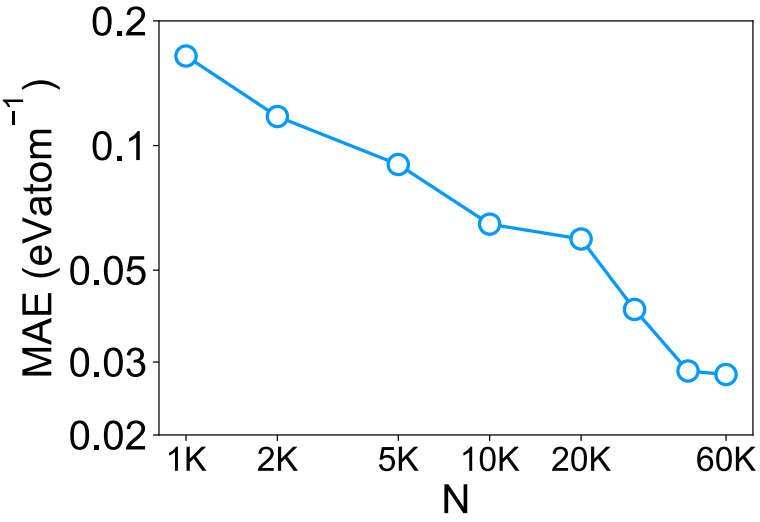


Figure S4. Convergence of formation energy MEGNet model.

Elemental embeddings

Figure S5 show the Pearson correlations between elemental embedding vectors obtained from the (a) one MEGNet block model and (b) five MEGNet block model. The contrast of the embeddings decreases with more MEGNet blocks due to the stronger influence of long-ranged geometry with stacking.

(a) (b)

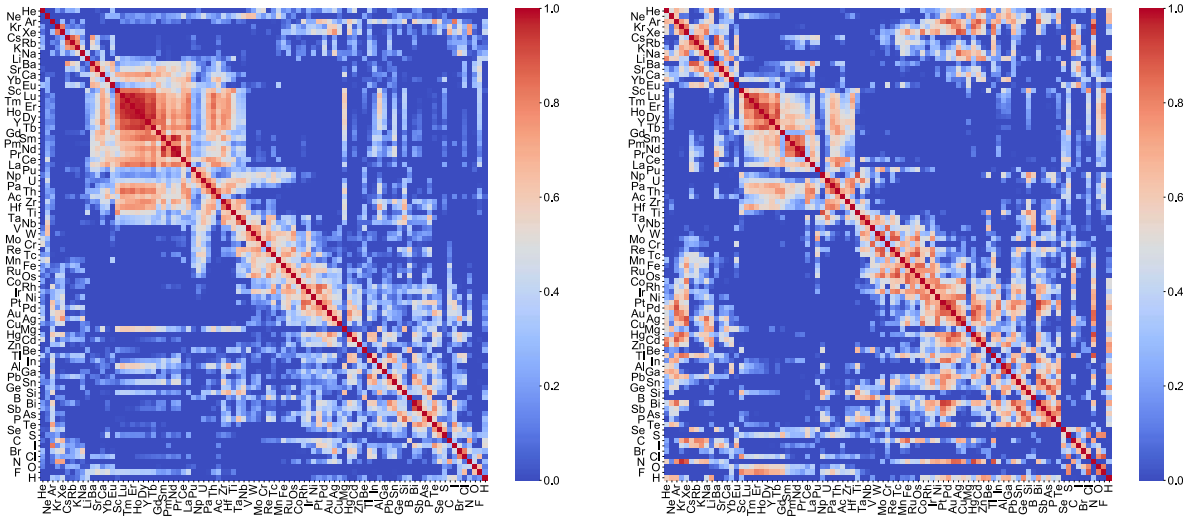


Figure S5. Elemental embedding correlations using (a) one MEGNet block and (b) five MEGNet blocks , sorted by the Mendeleev number.[4]

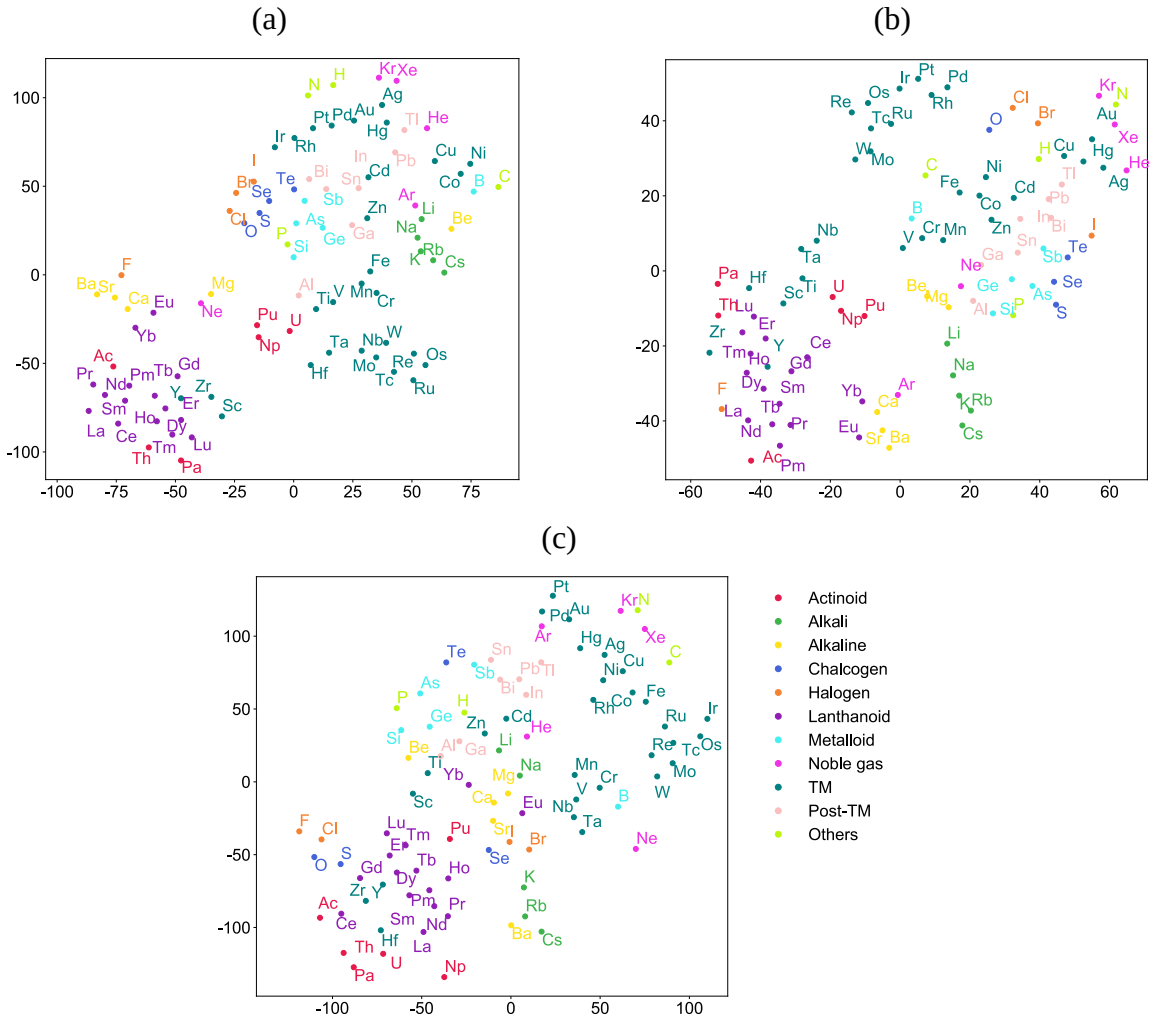


Figure S6. Two-dimensional T-SNE plots of elemental embeddings for (a) one MEGNet block, (b) three MEGNet block and (c) five MEGNet block models.

Figure S6 shows the T-SNE[5] two dimensional plots of elemental embeddings for 1-block, 3-block and 5-block models. Several runs of the T-SNE algorithm were performed to ensure that the resulting graphs are consistent and the error is lowest. One notable observation is that consistent with the Mendeleev number and structural graphs by Pettifor[4], Eu and Yb do not follow closely with other lanthanoids, but are closer to the alkaline earth elements such as Ca and Sr. We also show that Mg is not close to Ca as in periodic table, and this is also consistent with the Mendeleev number.

References

1. R. Ramakrishnan, P.O. Dral, M. Rupp, and O.A. Von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, *Sci. Data* **1**, 1–7 (2014).
2. A. Jain, S.P. Ong, G. Hautier, *et al.*, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.* **1**(1), 11002 (2013).
3. S.P. Ong, W.D. Richards, A. Jain, *et al.*, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, *Comput. Mater. Sci.* **68**, 314–319 (2013).
4. D.G. Pettifor, Structure maps for. Pseudobinary and ternary phases, *Mater. Sci. Technol.* **4**(8), 675–691 (1988).
5. L. van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* **9**(Nov), 2579–2605 (2008).