Supporting information

Establishing Deep InfoMax as an effective self-supervised learning methodology in materials informatics

Michael Moran,[ab] Michael W. Gaultois,[ab] Vladimir V. Gusev*[c], Dmytro Antypov[ab], Matthew J. Rosseinsky*[ab]

[a] *Department of Chemistry, University of Liverpool, Crown St, Liverpool, L69 7ZD, UK.*
[b] *Leverhulme Research Centre for Functional Materials Design, 51 Oxford Street, Liverpool, L7 3NY, UK*
[c] *Department of Computer Science, University of Liverpool, Ashton Street, Liverpool, L69 3BX, UK*

January 29, 2025

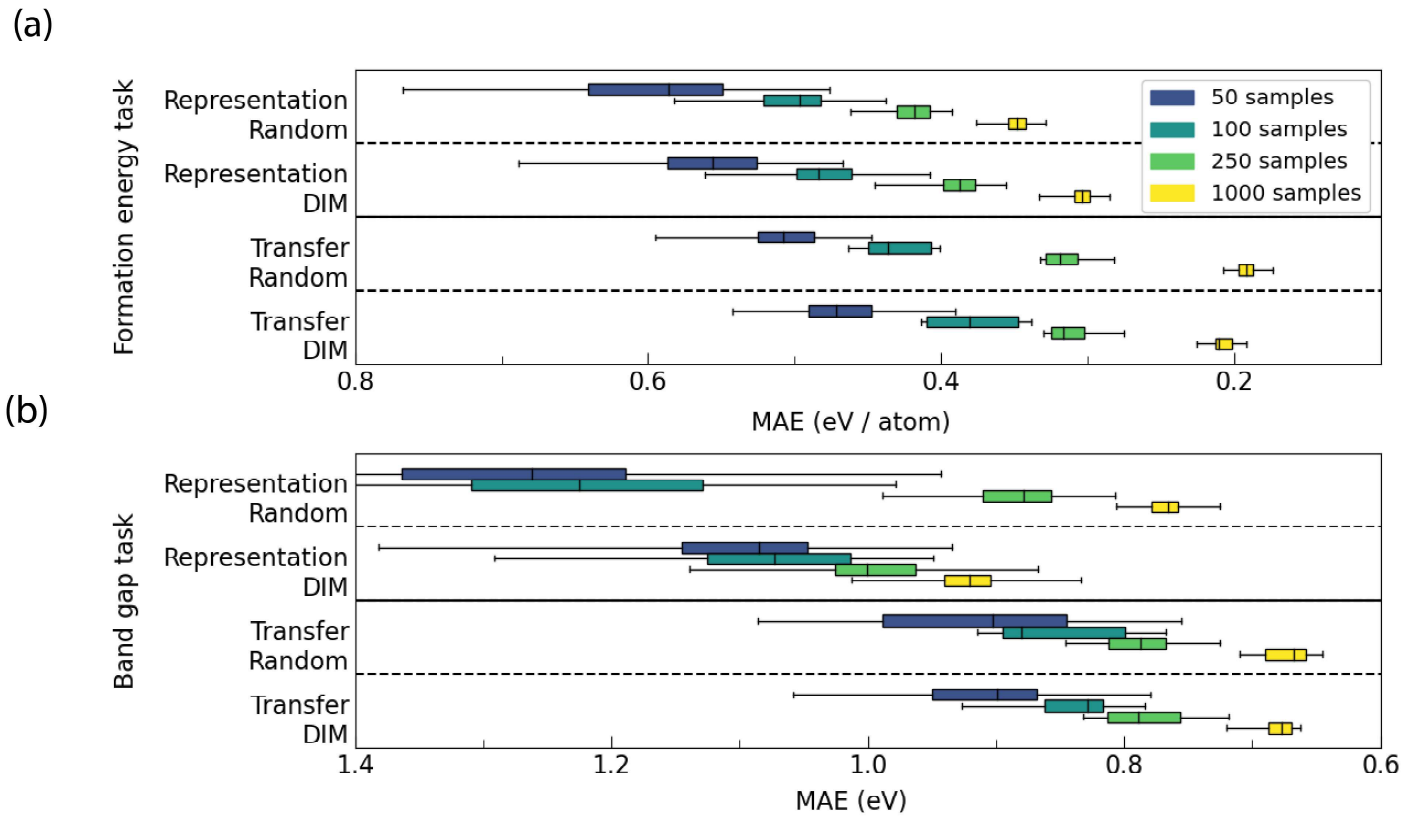

Fig. S1 Training downstream neural networks on the representation produced by a trained Deep InfoMax (DIM) model and an untrained Deep InfoMax model with initial parameters (Representation, fig. 8) is compared to training supervised Site-Net models using both random starting parameters and starting parameters taken from a trained Deep InfoMax model (Transfer, fig. 7). The methodologies are compared for the (a) formation energy task and (b) band gap task with label availability's varying from 50 to 1000. For representation learning, the box plots show 100 models, each trained on distinct randomly sampled property labels from the training dataset. For the Site-Net models, the box plots show 12 models using the same methodology. Results shown are the MAE of the models on the test dataset of ~20,000 samples regardless of the amount of training data. Even with very small amounts of data, training a full Site-Net model results in superior performance for all cases. This is a surprising finding; since the Site-Net model has more than 10^5 parameters and there are only 50 data points in the least data abundant case, we would expect severe overtraining and for training a small model on a frozen representation to be superior. This is not the case. The performance of Site-Nets with such small amounts of data suggests that this is an example of the double descent phenomena, where heavily over parameterised models eventually overcome overtraining and start to perform better than models optimised according to the bias-variance tradeoff. It is noted that both the Site-Net models and the downstream models trained in scikit-learn are using a single 64 node hidden layer in between the global representation and the prediction, and as such, they are directly comparable.

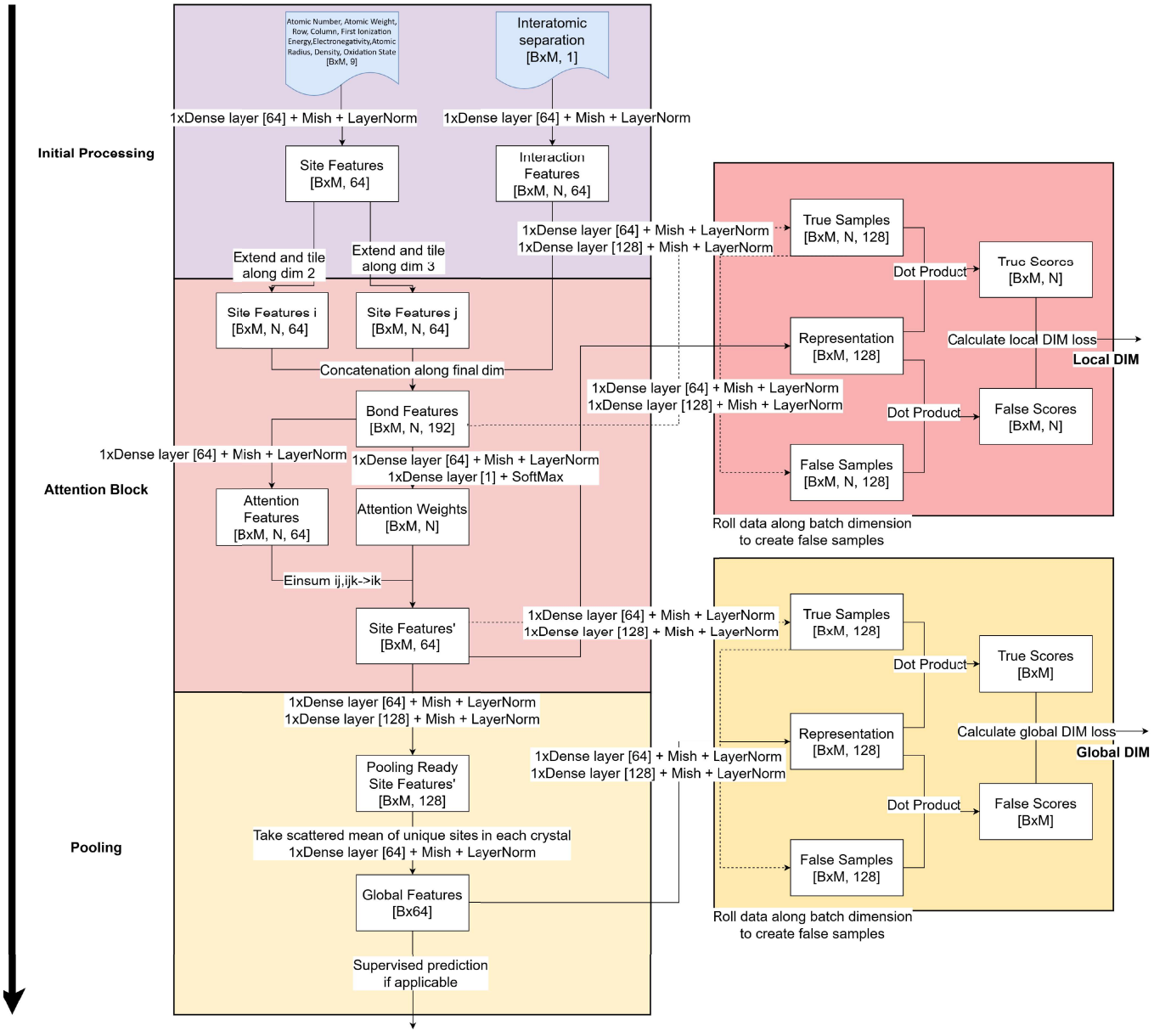

Fig. S2 A full schematic of the Site-Net + Deep InfoMax architecture used for the self-supervised pretraining. The Deep InfoMax blocks are connected across the pooling steps in the Site-Net architecture. The Deep InfoMax blocks are a strict addition to what was originally a supervised architecture demonstrating the flexibility of the approach. B, N, and M represent the batch size, the number of atomic sites in the unit cell, and the number of sites in the primitive unit cell respectively. These are dictated by the data structure during training, and can be arbitrarily set during inference. For simplicity, only the output dimension size for neural net layers is shown. Dense layers are always applied to the rightmost rank of the tensor.