Supplementary Information

Incorporation of density scaling constraint in density functional design via contrastive representation learning

Weiyi Gong [a], Tao Sun [b], Hexin Bai [c], Shah Tanvir ur Rahman Chowdhury [d], Peng Chu [c], Anoj Aryal [a], Jie Yu [e], Haibin Ling [*b], John P. Perdew [*efg], Qimin Yan [*a]

[a] Department of Physics, Northeastern University, Boston, MA 02115, USA
[b] Department of Computer Science, Stony Brook University, Stony Brook, NY 11794, USA
[c] Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA
[d] Department of Material Science, Thayer School of Engineering, Dartmouth College, Hanover, NH 03755, USA
[e] Department of Physics, Temple University, Philadelphia, PA 19122, USA
[f] Department of Chemistry, Temple University, Philadelphia, PA 19122, USA
[g] Present address: Department of Physics and Engineering Physics, Tulane University, New Orleans, LA 70118, USA

Residual neural networks

In this work, we used three-dimensional convolutional neural networks (Conv3d) as building blocks to encode the electron density in three-dimensional space and utilized a residual neural network (ResNet) architecture. ResNet is a type of neural network architecture that has been widely used in image recognition. With deeper and deeper neural networks, effective learning becomes more challenging due to the gradient vanishing or exploding problem,[1, 2] which makes traditional models using convolutional neural network layers reach a limit of performance when the number of layers increases. In 2016, He *et al.*[3] proposed using skip-connection that allows direct connection from the input layer to the output. By skipping intermediate layers, the model is able to learn the identity map even if there is a gradient issue within these layers. Instead of learning the mapping H between input x and target y, residual networks aim to learn the residual F:

$$F(x) := H(x) - x$$

In the worst case, a trivial result is learned such that $F(x) = 0$, the mapping H is the identity mapping $H(x) = x$. This skip-connection architecture enables the learning ability of neural networks that are extremely deep, which is critical for large-scale three-dimensional electron densities.

Derivation of NT-Xent loss

The normalized temperature-scaled cross entropy loss (NT-Xent) loss is defined as:

$$l_{ij} = -\log \frac{\exp\left(z_i \cdot z_j/\tau\right)}{\sum_{k=1,k\neq i}^{2N} \exp\left(z_i \cdot z_k/\tau\right)},$$

We would like to prove that when $z_i \cdot z_j = 1$ and $z_i \cdot z_k = -1\ (k \neq j)$, the loss will converge to 0. Indeed, for $\tau \to 0^+$, $z_i \cdot z_j = 1$, $z_i \cdot z_k = -1\ (k \neq j)$,

$$l_{ij} = \log\left(1 + \frac{\sum_{k=1,k\neq i,j}^{2N} \exp(z_i \cdot z_k/\tau)}{\exp(z_i \cdot z_j/\tau)}\right) = \log\left[1 + (2\mathrm{N} - 2)\exp\left(-\frac{2}{\tau}\right)\right] \to 0$$

For a batch of N molecules, each molecule has a pair of unscaled and scaled data. z_{2k-1} and z_{2k} are the corresponding projected representations of unscaled and scaled densities of the same molecule. Notice that the loss function is asymmetric ($l_{ij} \neq l_{ji}$). To make it symmetric, the total loss is chosen as:

$$L = \frac{1}{2N}\sum_{k=1}^{N}(l_{2k-1,2k} + l_{2k,2k-1})$$

The loss is zero when the projected representations of different molecules are antiparallel to each other while that of the same molecule are parallel to each other, which ensures that dissimilar samples are pushed far apart from each other.

The down-sampling of electron densities

We implemented a down-sampling technique for the (129, 129, 129) shape of grids by applying a simple multi-layer perceptron (MLP) consisting of two linear layers and a ReLU activation function in between. The input shape is changed from (129, 129, 129) to (65, 65, 65) after applying the down-sampling. We compare the MAE of ResNet(16, 32, 64, 128) trained, evaluated and tested on these two datasets with 80%, 10%, and 10% train-evaluate-test split, with and without down-sampling. The testing MAE of the model reduces from 0.565 eV to 0.461 eV. The result shows that although the model performance is restricted to the grid size, it can be partially improved by applying down-sampling on larger grids.

Comparison of ResNet and DoubleConv

We compare the results of ResNet with feature maps (16, 32, 64, 128) and DoubleConv with feature maps (32, 64, 128). The models were trained, evaluated, and tested on different train/validate/test split. As shown in Table S1, ResNet generally performs better than DoubleConv in predicting the exchange energies of electron densities for the same datasets.

Table S1. The MAE of ResNet and DoubleConv trained, evaluated, and tested on different splits.

Train/validate/test split	MAE on test set (eV)	
	ResNet (16, 32, 64, 128)	DoubleConv (32, 64, 128)
Supervised learning		
40k/5k/5k	0.696	1.182
Contrastive + transfer learning		
40k/5k/5k	0.565	1.008
32k/5k/5k	0.652	0.997
24k/5k/5k	0.777	1.080
16k/5k/5k	0.996	1.263
8k/5k/5k	1.262	1.811

Effect of translational symmetry

We investigate the impact of translational symmetry on the model performance by introducing a random translation for the grid-base electron densities during the supervised learning, contrastive learning, and transfer learning stages. As shown in Table S2, additional translational symmetry

generally improves the performance of contrastive and transfer learning, while make the performance of ResNet trained in a supervised manner slightly better.

Table S2. The MAE of ResNet trained evaluated and tested on the dataset without and with translation.

Train/validate/test split	MAE on test set (eV)	
	ResNet (16, 32, 64, 128)	ResNet (16, 32, 64, 128) + translation
Supervised learning		
40k/5k/5k	0.696	0.711
Contrastive + transfer learning		
40k/5k/5k	0.565	0.547
32k/5k/5k	0.652	0.609
24k/5k/5k	0.777	0.681
16k/5k/5k	0.996	0.819
8k/5k/5k	1.262	1.163

Interpolatability of the model

We investigate the interpolatability of the model trained using our proposed approach. The model was pre-trained by contrastive learning and fine-tuned by supervised learning using all the scaled data. Then another dataset (~5000 data) containing electron densities with scales randomly chosen from 0 to 3 was used to test the interpolatability of the trained model. As shown in Figure S1 for the 500 selected molecules, although those randomly chosen scales were not in the training data,

the model still gives reasonable predictions of exchange energies. This demonstrates the potential of our contrastive learning approach for the generalization to other random scales following the uniform scaling constraint.

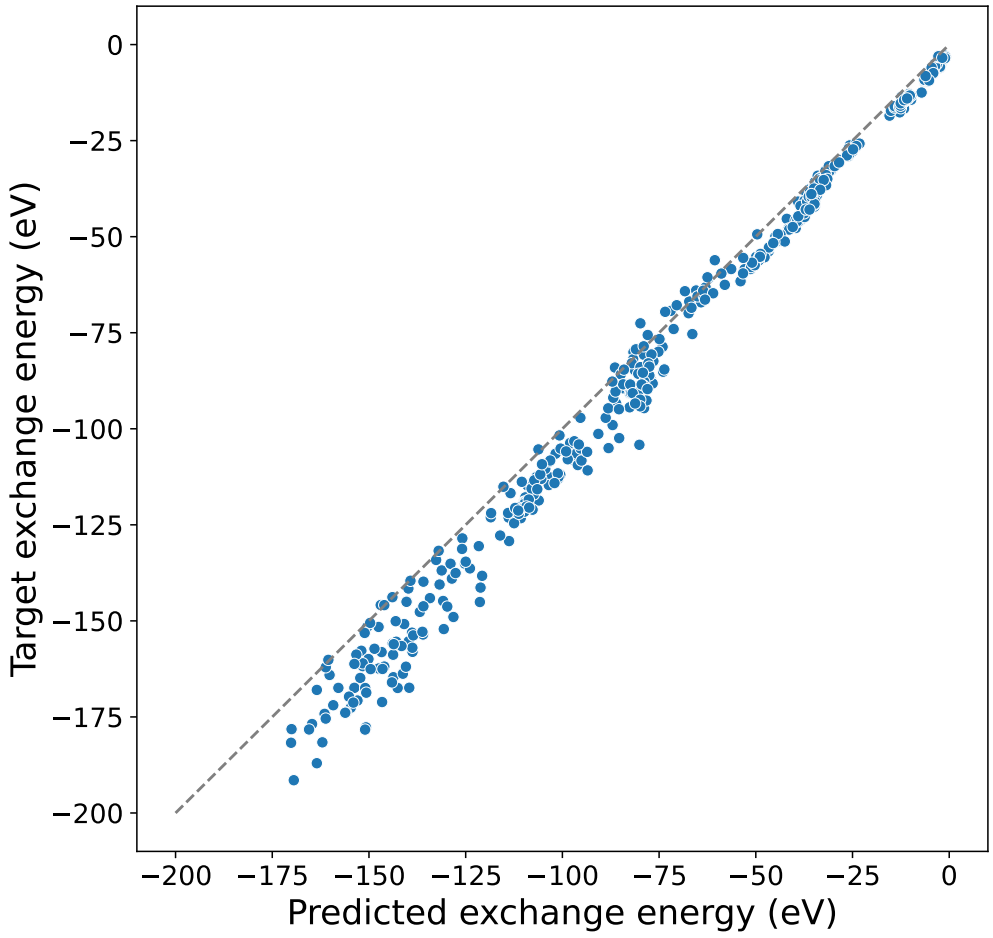

Figure S1. The target vs predicted exchange energies for the scaled densities of 500 molecules with random scales that are not present in the training set.

References:

1. Bengio, Y., Simard, P. and Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* **5** (2), 157-166 (1994).

2. Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics,* 249-256 (Sardinia, Italy, 2010).

3. He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770-778 (Las Vegas, NV, USA, 2016).