Electronic Supplementary Material (ESI) for Molecular Systems Design & Engineering.

Supplementary Information

Application of Transfer Learning to Predict Diffusion Property in Metal-Organic Frameworks

*Yunsung Lim, Jihan Kim**

Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

* Correspondence to: Jihan Kim (jihankim@kaist.ac.kr)

Contents

S1. Parity plots

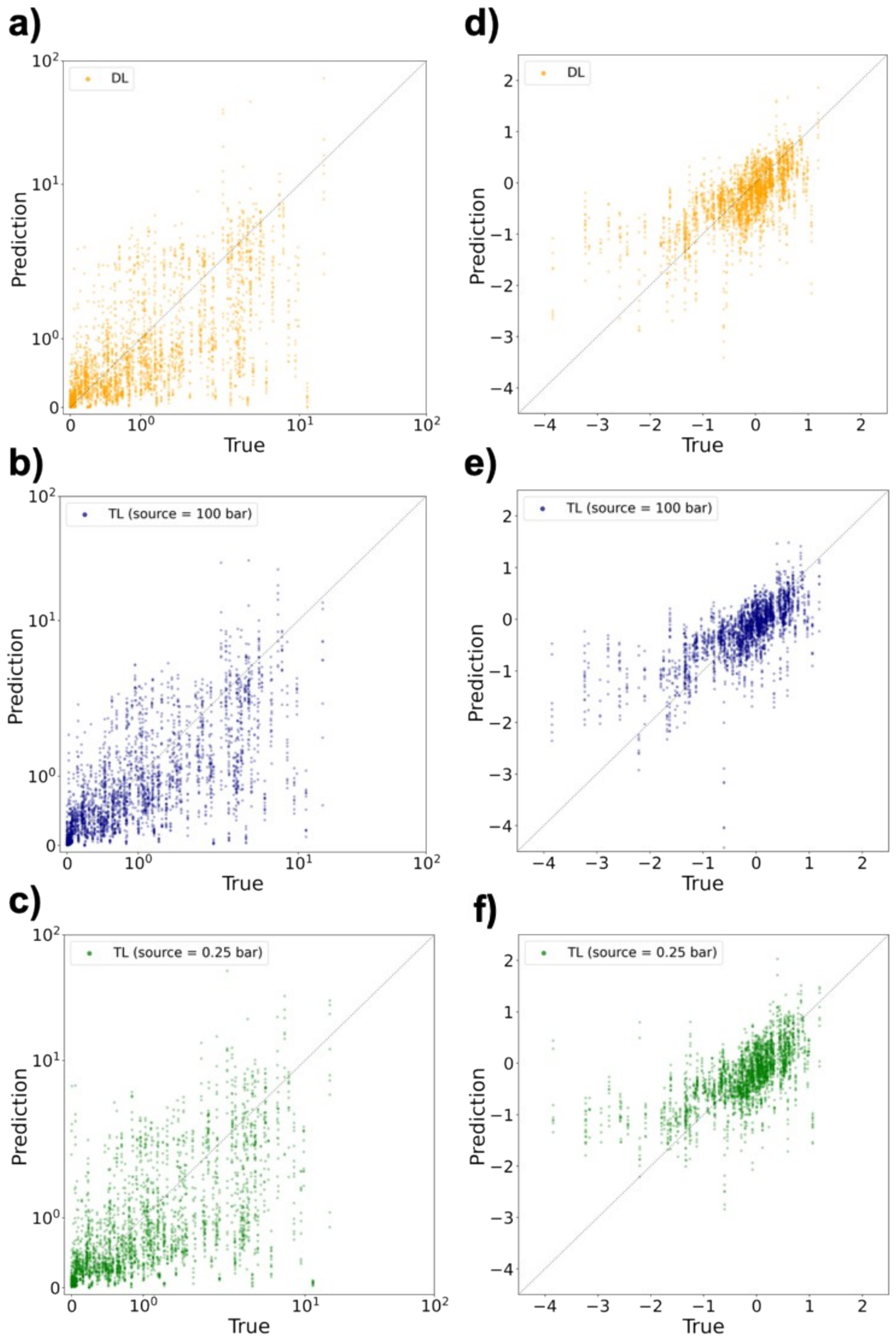

Fig. S1 Parity plot of the test set from every 50 sets for the direct learning (DL) model, the best model (*300-TL-100*) and the worst model (*300-TL-0.25*). Actual value of self-diffusion coefficients (D_s / 10^{-8}) of a) DL model, b) 300-TL-100, and c) 300-TL-0.25. Logarithmic value of self-diffusion coefficients (log $D_s/10^{-8}$) of d) DL model, e) 300-TL-100, and f) 300-TL-0.25.

S2. Performance evaluation with RMSE and MAE metrics

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \ (y_i : true\ value, \hat{y}_i : prediction)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i-1}^{n}|y_i - \hat{y}_i| \ (y_i : true\ value, \hat{y}_i : prediction)$$

$y_i\ and, \hat{y}_i$ are self-diffusion coefficients that converted into log scale and divided by 10^{-8} (m^2/s). For example, $y_i = log\frac{D_{S,\,true}\,(m^2/s)}{10^{-8}\,(m^2/s)}$.

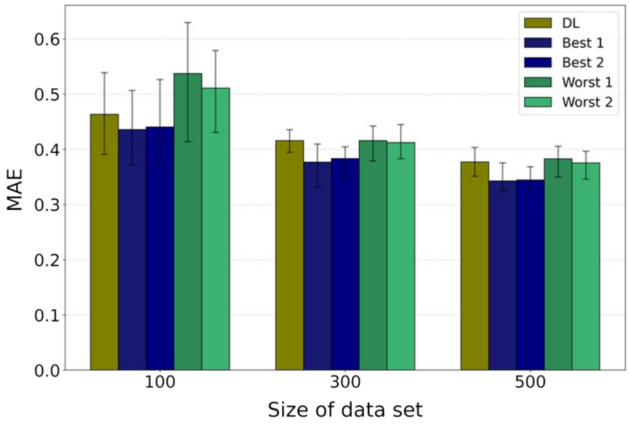

Fig. S2 Results of transfer learning task from adsorption (gas uptake) to diffusion property (self-diffusion coefficient). Top 2 cases and bottom 2 cases among the 12 TL models for each data size (100, 300, and 500) in respect to MAE.

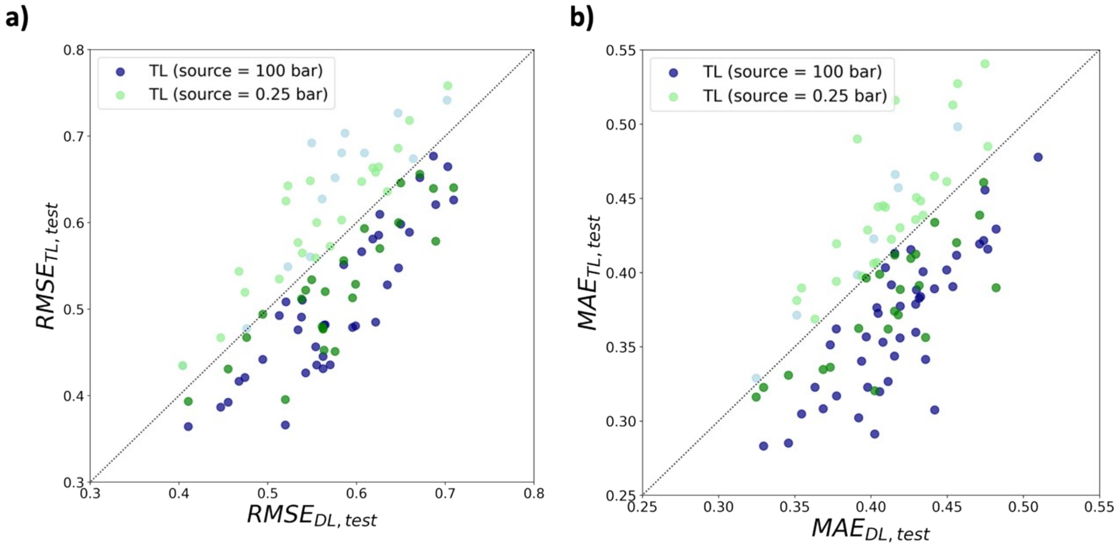

Fig. S3 Scatter plot between TL and DL of every 50 sets for the best source (100 bar) and the worst source (0.25 bar). a) RMSE and b) MAE

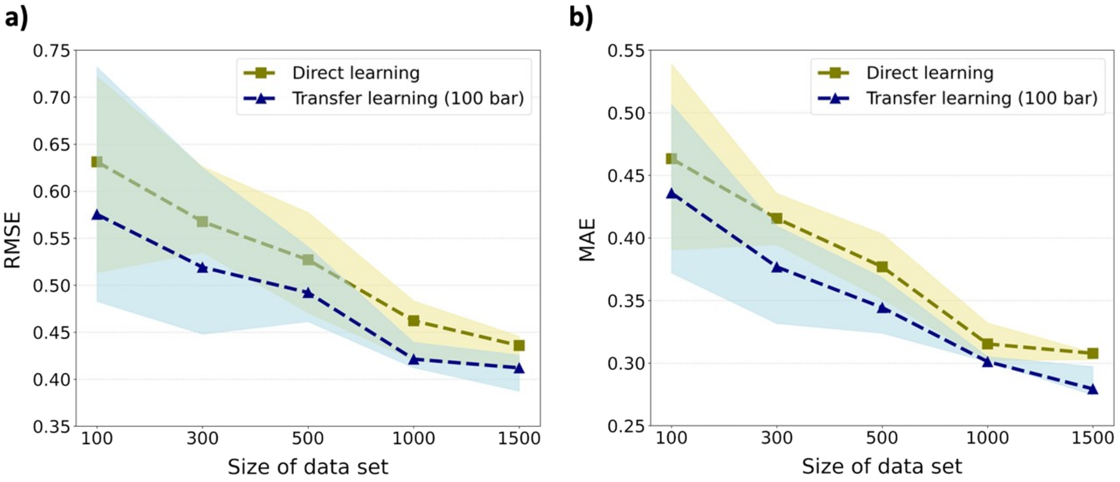

Fig. S4 Aspect of change in performance improvement as data size increases in respect to a) RMSE and b) MAE as an evaluation metric. The comparison was held with the TL model with the pre-trained model at 100 bar (*i-TL-100*) that generally performs well regardless of the data size. The markers denote the median value and the blurred region denotes the range between 1st and 3rd quartile.

S3. Details on simulation time

Simulation time for self-diffusion calculation (molecular dynamics simulation with 8 cpu cores) and gas uptake calculation (monte carlo simulation with single GPU core) are considered for 1563 CoRE MOF structures that we used in this work. Due to 8 cpu cores were used for self-diffusion coefficient calculations, total simulation time was obtained with 8 * total wall time. From the calculations, the averaged simulation time of self-diffusion coefficients calculations is 63628.79 seconds per structure. Meanwhile, the averaged simulation time of the gas uptake calculations is 35.01 seconds per structure. Therefore, self-diffusion calculation need 1817 times more time than gas uptake calculations. Comparing the case of size 300 with pre-trained model (pre-trained with 23,845 gas uptake results) and size 500 without pre-trained model, the computational can be saved as 62.62% in the size 300 with pre-trained model (**Table S1**). Given that the pre-training time is very small compared to the computational time, the training time was not considered.

	Case 1 **(size 300 with pre-trained model)**	**Case 2** **(size 500 without pre-trained model)**
Self-diffusion coefficient calculations	300 x 63628.79s = 19088637s	500 x 63628.79 = 31814395s
Gas uptake calculations	23,845 x 35.01s = 834813.45s	0 x 35.01s = 0s
Total time	19923450.45s	31814395s

Table S1 Simulation time calculation for size 300 with pre-trained model and size 500 without TL (DL)

S4. Details on finding an optimal pressure candidate for transfer learning

We tried to find a new optimal pressure candidate for pre-training in the context that the higher similarity in the PFI, the higher performance in the TL can be guaranteed. As such, we tried to find a meaningful relationship between the similarity of PFI and pressures. As we mentioned in manuscript, the similarity of PFI was measured as an Euclidean distance, so we assume that there is an arbitrary function that takes pressure related values as the domain and similarities as the range (y = f(x), y = similarity of PFI, x = logarithmic value of the pressure), there are two local minima in the function (see **Fig. S5a**). Given that at least quartic function is required to have two local minima, we obtained the polynomial regression model of degree 4 from our results using Python library, scikit-learn[1] (lightblue line in **Fig. S5a**). The root that gives the global minimum is 93.6324 bar, so 93.6324 bar was selected as new optimal pressure candidate. Thus, a new pre-trained model with the methane uptakes at 93.6324 bar was prepared and *300-TL-93.6324* model was constructed. Since the gap between the Euclidean distance between 93.6324 bar and 100 bar is only 0.002 (93.6324 bar: 0.303, 100 bar: 0.305), there is no drastic improvement in performance, but still there was a slight improvement and this can demonstrate that the Euclidean distance between PFI vectors can be an alternative metric to estimate whether the source property is proper for TL tasks to predict the target property. As shown in **Fig. S5b**, although the median of R^2 score of *300-TL-93.6324* is almost the same as *300-TL-100* that was regarded as the best model, the minimum value (lowest bar) of *300-TL-93.6324* is much higher than that of *300-TL-100* and even the 3rd quartile (highest bar) of *300-TL-93.6324* is slightly higher than that of *300-TL-100.* In addition, if the evaluation indicator is changed as RMSE, the median of *300-TL-93.6324* was lower than that of *300-TL-100* (see **Fig. S5c**).

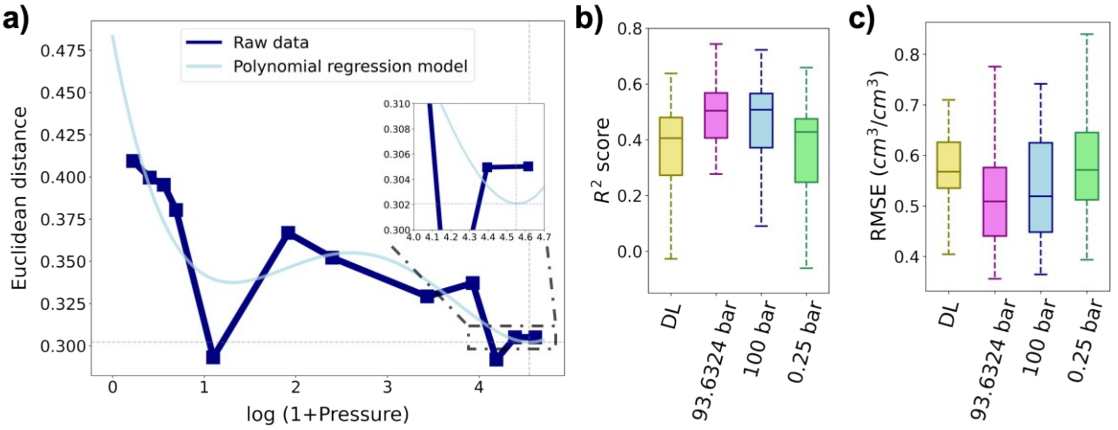

Fig. S5 Suggestion of a new pressure candidate for high performance in the TL task from the relationship between similarity of feature importance and pressure condition of simulation to compute gas uptakes. a) Polynomial regression model and the data that consists of euclidean distances and pressure. The global minimum appeared between 80 bar and 100 bar. Box plot of b) R^2 score and c) RMSE of 50 sets with the DL model, *300-TL-93.6324* (new candidate), *300-TL-100* (best case), and *300-TL-0.25* (worst case).

S5. Details on energy descriptors

Energy descriptors were generated by two steps, generation of energy grids and converting energy values of each grid point as a histogram.

First, the interaction energy of each grid point was computed using the Lennard-Jones (LJ) 12-6 potential model (Equation 1). r is the distance between the guest molecule (CH4 in this case) and the atoms of the host frameworks (MOFs in this case). The force field parameters ε and σ for both of the guest molecule and the atoms of the host frameworks are obtained from Universal Force Field (UFF)[2] and the Lorentz-Berthelot mixing rule was used as a combining rule.

$$U_{LJ}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \text{ (Equation 1)}$$

Given that large space between grid points can be generated if the grid sizes were selected as a fixed value (such as 24 x 24 x 24) as previous works did,[3] we computed energy values for every 1 Angstrom. For example, the energy values of 3000 grid points were computed for a framework with cell size 10 x 15 x 20. The energy for each grid point was computed as Kelvin units.

Second, we obtained the distribution of the energy values from the histogram. To determine a consistent energy range for every framework, we set the cutoff for energy values as -4000K to 5000K. The energy below -4000K is converted into -4000K and the energy above 5000K is converted into 5000K. Then, we obtained a histogram with the converted results and the number of bins as 50. The counts for each bin were normalized as the sum of counts becomes 1 because the number of grid points was different for every framework.

References

1. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *The journal of machine learning research*, 2011, **12**, 2825-2830.
2. A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *Journal of the American Chemical Society*, 1992, **114**, 10024-10035.
3. B. Kim, S. Lee and J. Kim, *Science Advances*, 2020, **6**, eaax9324.