Electronic Supplementary Information (ESI)

A Neural Compact Model Based-on Transfer Learning for Organic FETs with Gaussian Disorder

Minsun Cho,[a] Marin Franot,[a,b] O-Joun Lee,[c] Sungyeop Jung[a,*]

[a]Advanced Institute of Convergence Technology, Seoul National University, 145, Gwanggyo-ro, Yeongtong-gu, Suwon-si, 16229, Gyeonggi-do, Republic of Korea

[b]École Nationale Supérieure d'Électrotechnique, d' Électronique, d'Informatique, d'Hydraulique et des Télécommunications, Toulouse INP, 2 Rue Charles Camichel, Toulouse, 31000, Occitanie, France

[c]The Catholic University of Korea, 43, Jibong-ro, Bucheon-si, 14662, Gyeonggi-do, Republic of Korea

E-mail : sungyeop.jung@snu.ac.kr

S1. Principal Component Analysis (PCA) of Experiment Data

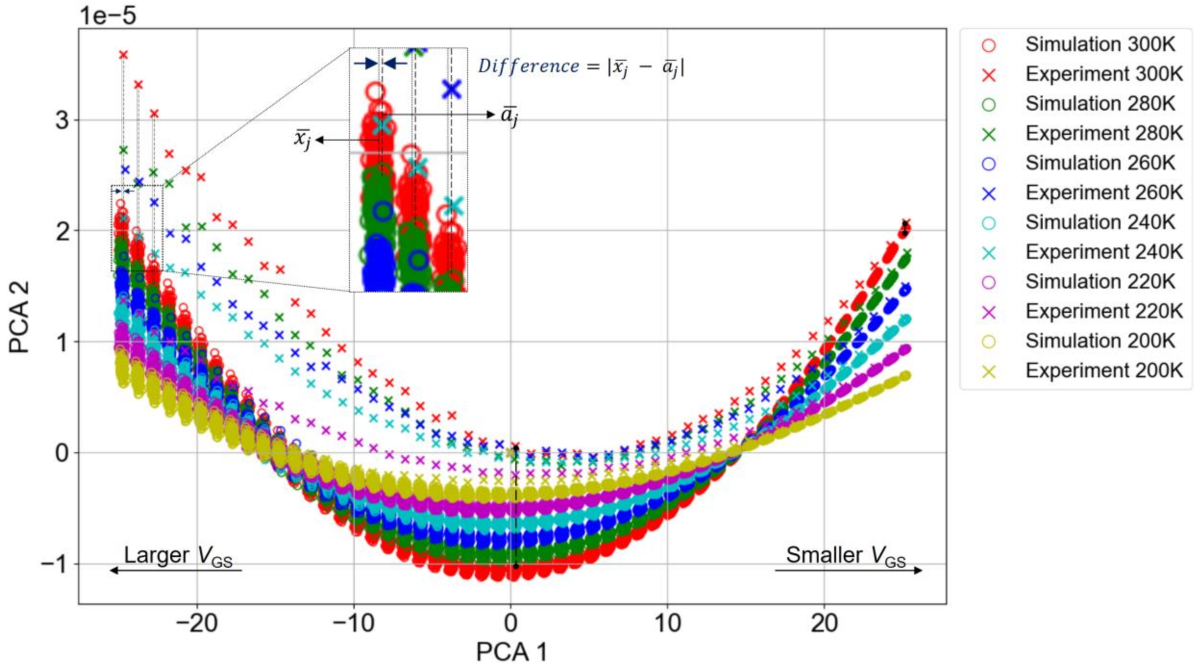

Figure S1. The PCA results under the conditions of V_{DS} = -60V, V_{GS} = -60 to -10V by 1V. The analysis is based on all temperature conditions (200 to 300K by 20K), encompassing all data points of the I_D-V_{GS} curves.

The PCA was conducted to evaluate the similarity between TCAD simulation data and experimental data. For a fair comparison, a dataset at V_{GS} from -60 to -10V by 1V (51 points) was taken from both TCAD and experimental data across all temperature conditions from 200 to 300K by 20K and at a single V_{DS} condition of -60V. In Figure S1, each point (open circle and cross symbol) represents an I_D-V_{GS}-V_{DS} data point, where a total of 51 points along PCA1 constructing an I_D-V_{GS} curve. In the simulation data case, under the same conditions, 100 I_D-V_{GS} curves exist for each temperature condition.

On the PCA1 axis, there was small difference between the simulation and experimental data, indicating a high degree of similarity in terms of V_{GS}. In detail, the differences in the x-values of the simulation ($\bar{x}$) and experimental data ($\bar{a}$) on the PCA1 axis were in average 0.19. This value represents the average difference in x-coordinate values at each V_{GS}, with the average x-coordinate values for the 100 data points used for the simulation data. The detailed formula is as follows, where n = (the number of data points along PCA1 axis) = 51:

$$\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i \tag{S1}$$

$$\text{Mean Difference} = \frac{1}{n}\cdot\sum_{j=1}^{n}|\bar{x}_j - \bar{a}_j| \tag{S2}$$

This value is smaller than the differences in PCA2 axis. On the PCA2 axis, the differences were negligible, which were on the order of 10^{-5}. The difference on the PCA2 was more pronounced at higher V_{GS} values (negative region on the PCA1 axis) due to larger I_D at higher V_{GS}.

S2. Effect of Optimizer Algorithm

In deep learning, optimizers are algorithms that adjust the model's parameters during training to minimize a loss function. Thereby, a correct choice of an optimizer is important to build an accurate model. Among various optimizers, the RMSprop (root mean square propagation), first proposed by Hinton in the Coursera course, is a gradient-based optimization algorithm [S1]. It stores only a certain number of past gradient information instead of keeping track of all previous gradients. To reduce the influence of gradient information, it uses the decaying average of squared gradients [S2]. The update formula for RMSprop is as follows:

$$g_t = \beta_2 g_{t-1} + (1-\beta_2)(\bigtriangledown f(x_{t-1}))^2. \qquad \text{(S3)}$$

On the other hand, Adam, which is the one of the most used optimizers, is a stochastic gradient descent method that is based on adaptive estimation of first order and second order moments [S3]. The Adam optimizer combines the advantages of momentum and RMSprop optimizers. Thereby, it enhances both the direction and magnitude of learning [S4]. The formula for Adam is as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) \bigtriangledown f(x_{t-1}), \qquad \text{(S4)}$$

$$\hat{m} = \frac{m_t}{1-\beta_1^t}, \hat{g}_t = \frac{g_t}{1-\beta_2^t}, \qquad \text{(S5)}$$

$$x_t = x_{t-1} - \frac{\eta}{\sqrt{g^t + \varepsilon}} \hat{m}_t, \qquad \text{(S6)}$$

where β1 represents the exponential moving average of momentum, and β2 represents the exponential moving average of RMSprop. $\hat{m}$ and $\hat{g}$ are bias-corrected values to prevent mt and g_t from being initialized as 0 in the beginning of training. ε is a small value added in the denominator to prevent division by zero. η denotes the learning rate.

As shown in Fig. S2 (a-d), we observed good performance in terms of accuracy and loss evaluation for both optimizer algorithms. In the meantime, for Adam optimizer, we noticed that the non-linear $I_D - V_{GS}$ curves shape near and above threshold region (V_{GS} = -20V to 5V) was not accurately modelled. By changing the optimizer to RMSprop, we improved model accuracy (Fig. S2 (a) v.s. Fig. S2 (b)) without a significant difference in loss values or the epoch-loss graph (Fig. S2 (c) and (d)). Thereby, we opted for an optimization algorithm such as RMSprop that excludes momentum for further experiment.

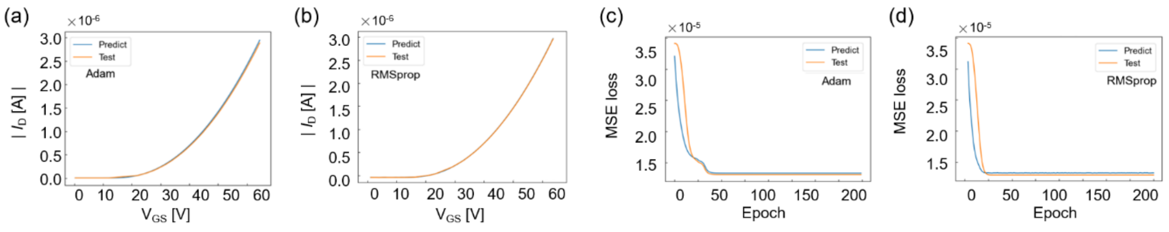

Figure S2. Predict and test results by changing optimizer algorithm. Predict and test $I_D - V_{GS}$ curves using (a) RMSprop and (b) Adam optimizer. The epoch-loss graph using (c) Adam and (d) RMSprop optimizer.

S3. The number of epochs and overfitting problem

To determine the optimal epoch for our model, with a focus on mitigating overfitting, we systematically conducted experiments by comparing the MSE loss of train and test data at various epochs (Fig. S3). A set of experiments had been conducted for 100 sets for transfer learning. The trends could be generalized among base learning and transfer learning in temperature and drain voltage domain. At epoch 100, it became apparent that the learning process was insufficiently progressed, as evidenced by relatively higher training and test loss values compared to epochs 200 and 500. Regarding epochs 200 and 500, both exhibited low test loss values accompanied by appropriately aligned training loss values. However, the epoch 200 configuration is the most suitable epoch for our model, since the epoch 200 configuration demonstrated not only a smaller discrepancy between training and test loss but also superior temporal efficiency. Furthermore, in the pursuit of identifying an optimal learning rate, we executed experiments with 1000 iterations. Although the training loss reached remarkably low values (0.009, 0.006, and 0.001, respectively for base learning and transfer learning in temperature and drain voltage domain), the test loss showed a comparatively higher magnitude, indicating a potential susceptibility to overfitting beyond 1000 iterations. In summary, by setting epoch 200 for the model configuration with the early stopping option and maintaining the minimal difference between training and test loss, we guaranteed that the results are free from overfitting.

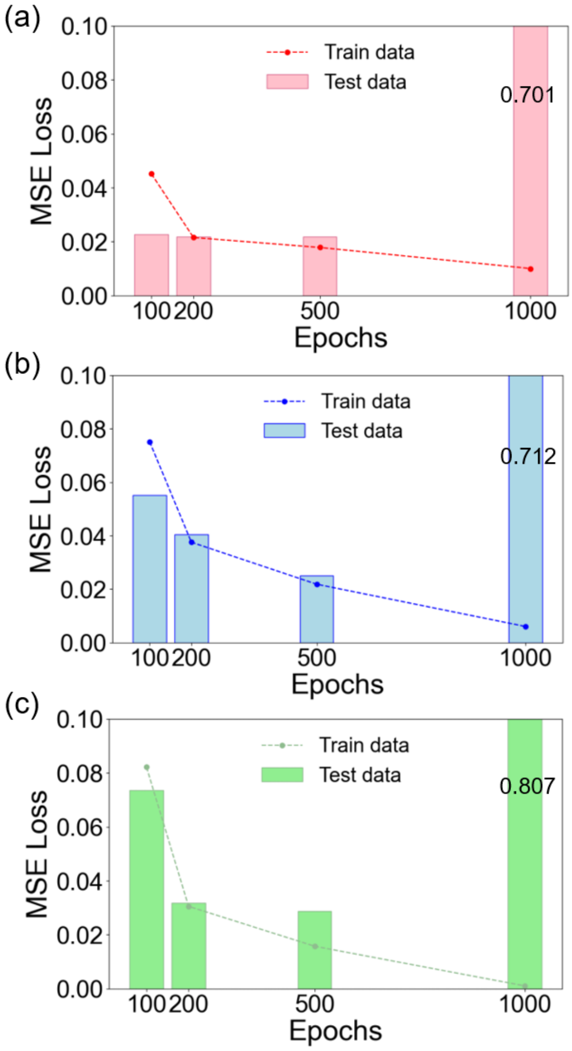

Figure S3. The MSE loss of train and test data with respect to the number of epochs in (a) base learning, and in transfer learning in (b) temperature and (c) V_{DS} domain. The MSE loss value of 1000 epoch is out of y-axis range.

S3. Effect of data scarcity on transfer learning

We conducted transfer learning to assess the performance of our model more accurately by training it on varying amounts of target data. The results on transfer learning on drain voltage and temperature domain are summarized in figure S4. For the drain voltage domain (Fig. S4(a)), the initial attempt with 100 sets demonstrated excellent evaluation metrics, showing an MSE loss of 0.0226 and an R-squared of 0.995. Through fine-tuning with 50 sets, 20 sets, and 10 sets, the model consistently exhibited similar loss values and R-squared to the initial attempt, thus validating the superiority of our transfer learning model. Notably, a more abrupt degradation in the Mean Absolute Percentage Error (MAPE) was observed from 10 sets onwards, suggesting that 20 sets could be the most appropriate number of data samples. For the temperature domain (Fig. S4(b)), the initial attempt with 100 sets provided MSE loss of 0.0318 and R-squared of 0.998. Fine-tuning with 50 sets, 20 sets, and 10 sets, the model showed a similar degradation pattern to that in the drain voltage domain with a pronounced increase of the MAPE from 10 sets onwards. The values of the figure-of-merits are summarized in Table S1 and S2.

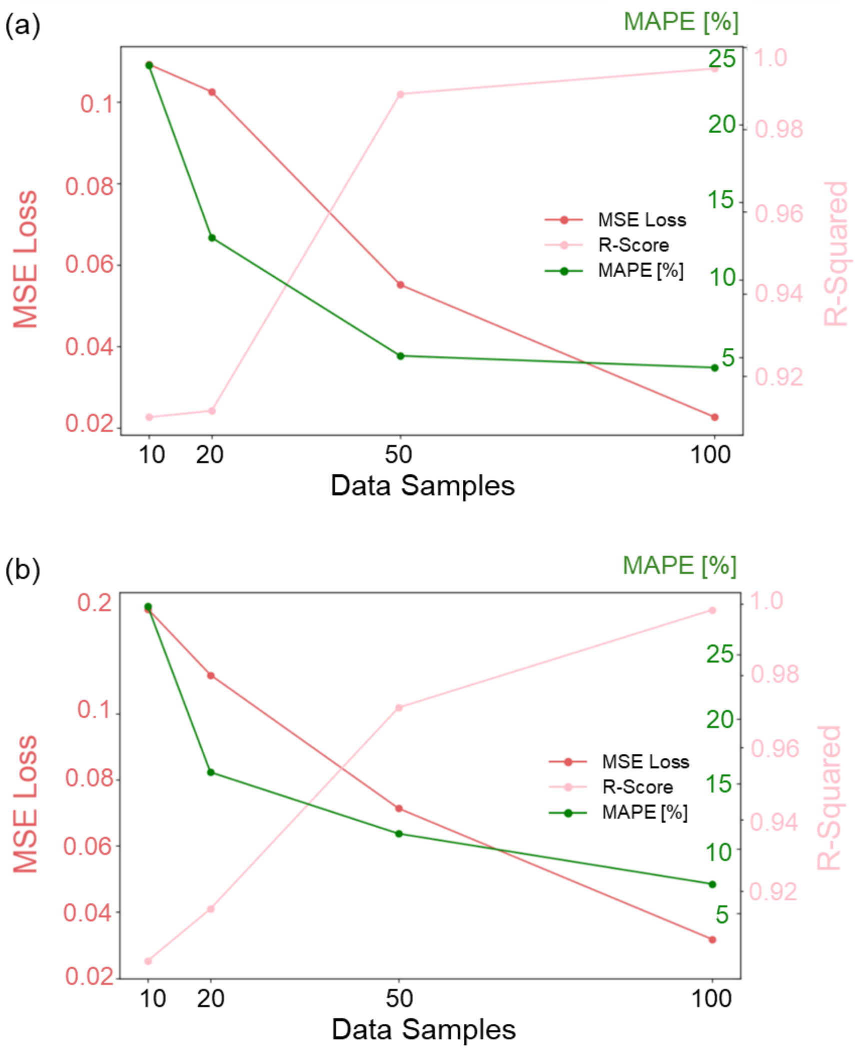

Figure S4. The effect of the number of data samples on the accuracy of the neural compact model represented by means of MSE (in red), R-squared (in pink) and MAPE (in green): for transfer learning in (a) the temperature domain, and (b) in the drain voltage domain.

To investigate the accuracy of our model, we analyzed the current-voltage plots, i.e. the transfer curve I_D-V_{GS} (in both linear and semi-logarithmic scale) as well as the first and second derivative of the transfer curve (see Fig. S5-S13). For the visualization purpose, representative conditions were chosen: V_{DS} = -60 and -5 V for transfer learning in temperature domain and 300 and 200K for transfer learning V_{DS} domain. The first and second derivative plots confirm that out model successfully models the current-voltage characteristics manifested by the complex Gaussian mobility model. The degradation becomes noticeable at low temperature and high V_{DS} conditions in fine-tuning with 10 sets in the prediction of the unseen data (Fig. S9(e, g)).

These experimental results show that our model's excellence can be established with a reduced number of required data sets, presenting a time and cost advantage. The findings highlight the potential of minimizing the demanded data quantity for achieving optimal model performance, thereby offering significant advantages in terms of time and cost efficiency.

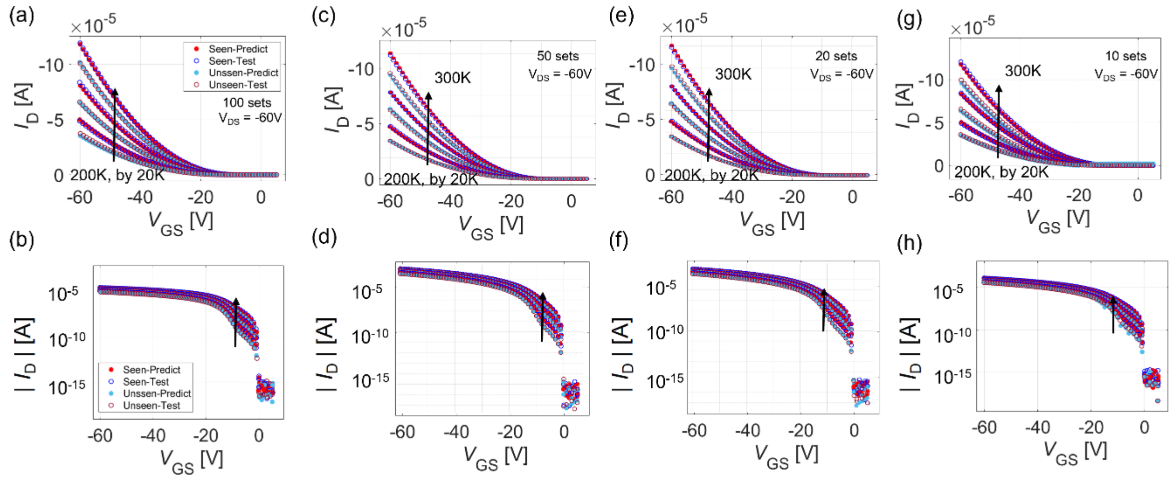

Figure S5. I_D – V_{GS} characteristics saturation (V_{DS} = -60V) regime in linear and semilog scale obtained by transfer learning of 100 sets (a) and (b), 50 sets (c) and (d), 20 sets (e) and (f), and 10 sets (g) and (h). Result of transfer learning: Seen-Prediction data (red filled circle), Seen-Test data (blue open-circle), Unseen-Prediction data (blue filled circle), and Unseen-Test data (red open circle). The temperature ranges from 300 to 200K in decrements of 20K, where the seen data are 300, 260, and 220K while the unseen data are 280, 240, and 200K.

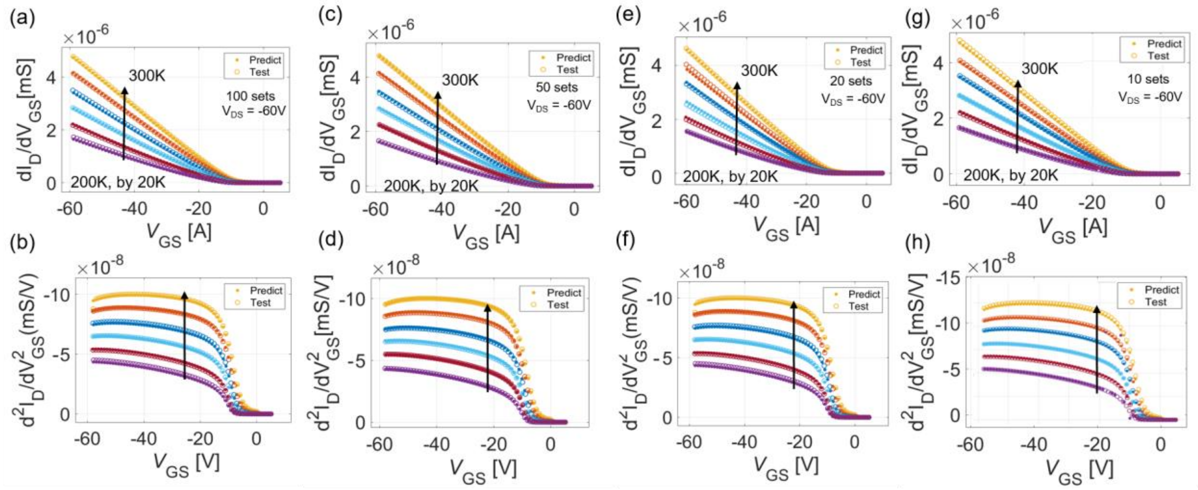

Figure S6. First and second derivative of I_D–V_{GS} characteristics under saturation regime (V_{DS} = -60V) obtained by transfer learning of 100 sets (a) and (b), 50 sets (c) and (d), 20 sets (e) and (f), and 10 sets (g) and (h).

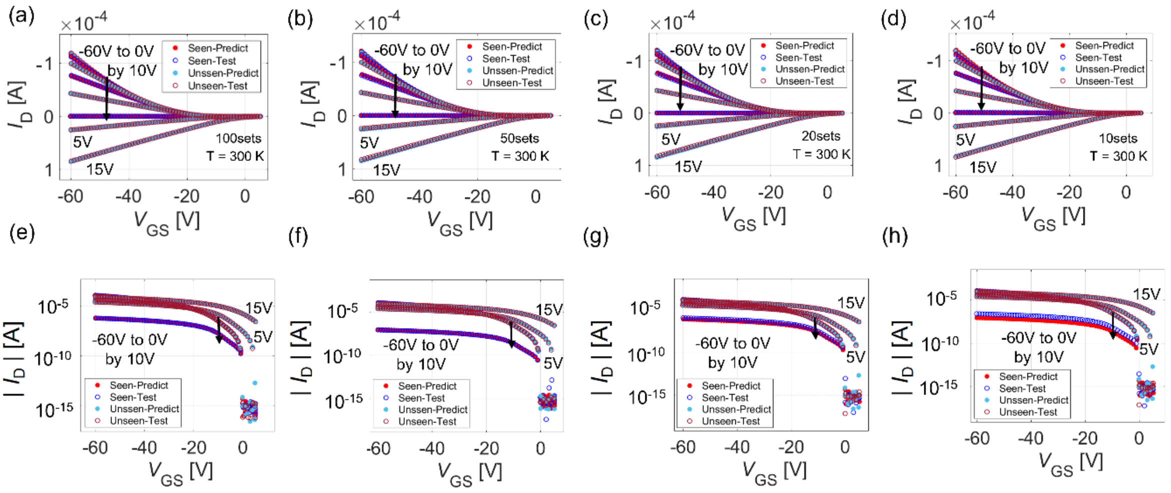

Figure S7. I_D−V_{GS} characteristics under linear regime (V_{DS} = -5V) in linear and semilog scale obtained by transfer learning of 100 sets (a) and (b), 50 sets (c) and (d), 20 sets (e) and (f), and 10 sets (g) and (h): Seen-Prediction data (red filled circle), Seen-Test data (blue open-circle), Unseen-Prediction data (blue filled circle), and Unseen-Test data (red open circle). The temperature ranges from 300 to 200K in decrements of 20K, where the seen data are 300, 260, and 220K while the unseen data are 280, 240, and 200K.

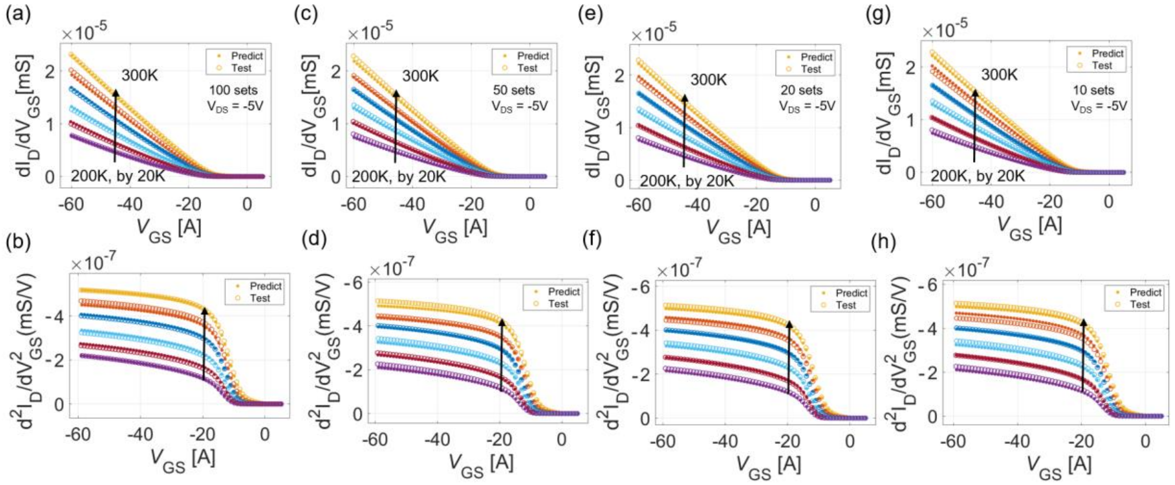

Figure S8. First and second derivative of I_D−V_{GS} characteristics under linear regime (V_{DS} = -5V) obtained by transfer learning of 100 sets (a) and (b), 50 sets (c) and (d), 20 sets (e) and (f), and 10 sets (g) and (h).

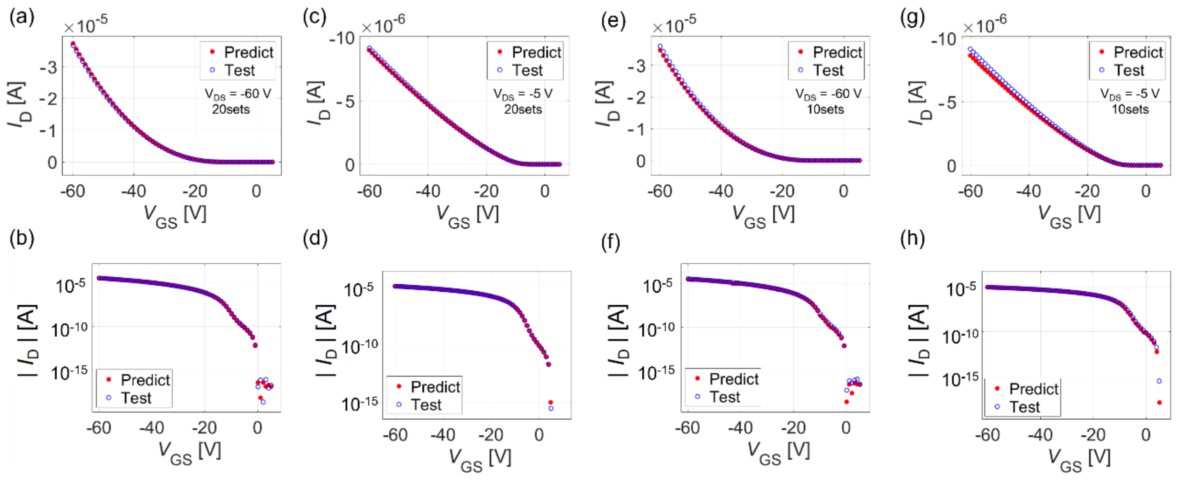

Figure S9. Degradation of accuracy when the data is scarce for the transfer learning in temperature domain. I_D−V_{GS} characteristics under saturation (V_{DS} = -60V) and linear (V_{DS} = -5V) regime: (a) and (b) are for 20 sets at V_{DS} = -60V, (c) and (d) are for 20 sets at V_{DS} = -5V, (e) and (f) are for 10 sets at V_{DS} = -60V (g) and (h) are for 10 sets at V_{DS} = -5V. Only data for 200K are shown.

Table S1. Time experiment result of each data samples and MSE loss, R squared and MAPE (percent, of the on region). Transfer learning in the temperature domain.

	100 SETS	50 SETS	20 SETS	10 SETS
PRE-TRAIN TIME	10h 27m 52s	10h 27m 52s	10h 27m 52s	10h 27m 52s
FINE-TUNE TIME	5h 31m 22s	3h 08m 51s	2h 48m 17s	2h 14m 43s
TOTAL TIME	5h 31m 22s	3h 08m 51s	2h 48m 17s	2h 14m 43s
MSE LOSS	0.0226	0.0551	0.102	0.109
R-SQUARED	0.995	0.988	0.912	0.910
MAPE (%)	4.32	5.08	12.7	23.8

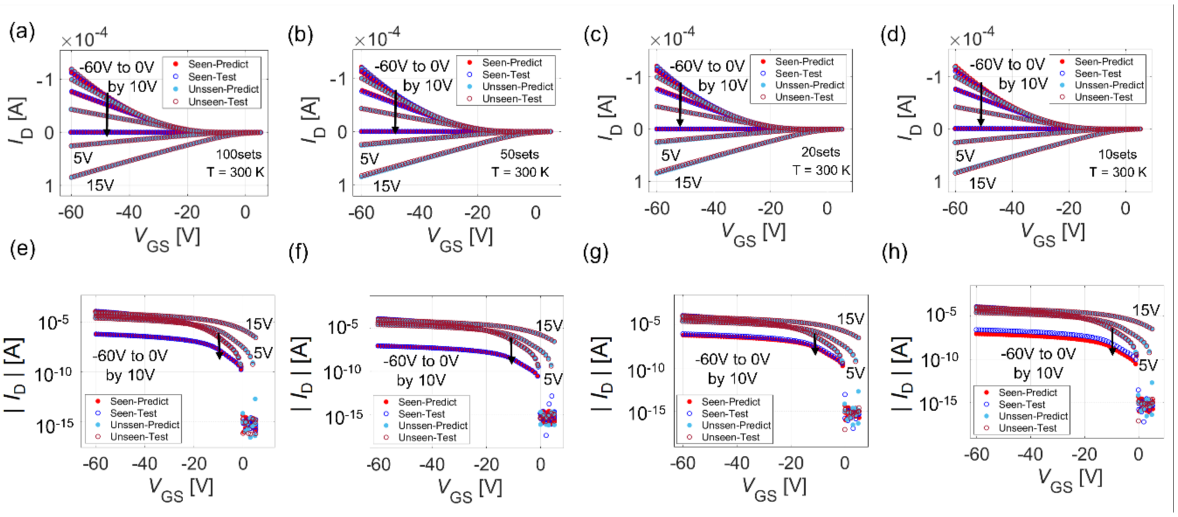

Figure S10. I_D – V_{GS} characteristics under linear (Temperature = 300K) in linear and semilog scale obtained by transfer learning of 100 sets (a) and (b), 50 sets (c) and (d), 20 sets (e) and (f), and 10 sets (g) and (h): Seen-Prediction data (red filled circle), Seen-Test data (blue open-circle), Unseen-Prediction data (blue filled circle), and Unseen-Test data (red open circle). The V_{DS} conditions of the data range from -60V to 15V in decrements of 10V. For transfer learning, the (V_{DS} conditions of the seen data are -60V, -40V, -20V and 0V. And the Unseen data are -50V, -30V, -10V, 5V and 15V.

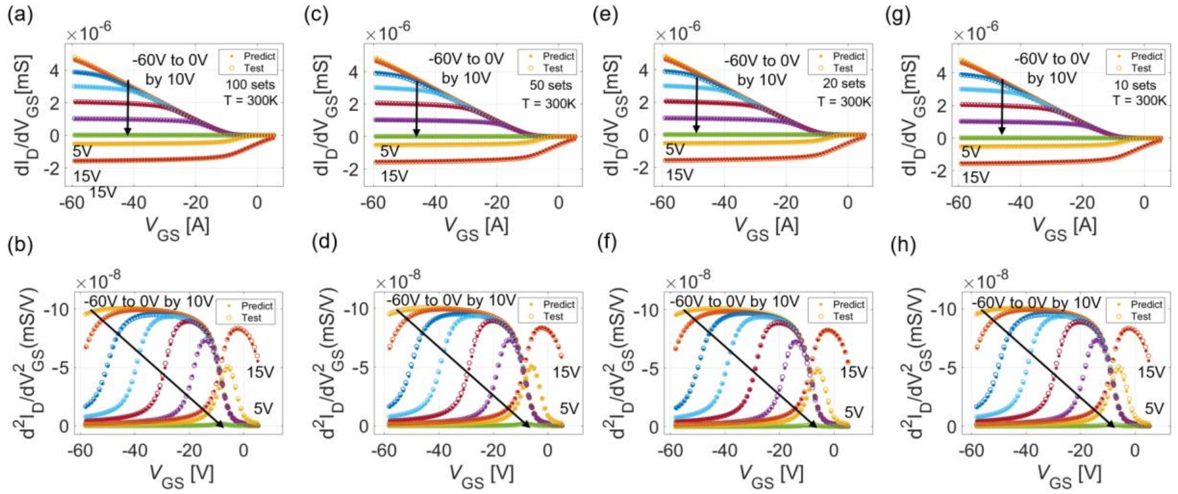

Figure S11. First and second derivative of I_D−V_{GS} characteristics under linear regime (Temperature = 300K) obtained by transfer learning of 100 sets (a) and (b), 50 sets (c) and (d), 20 sets (e) and (f), and 10 sets (g) and (h).

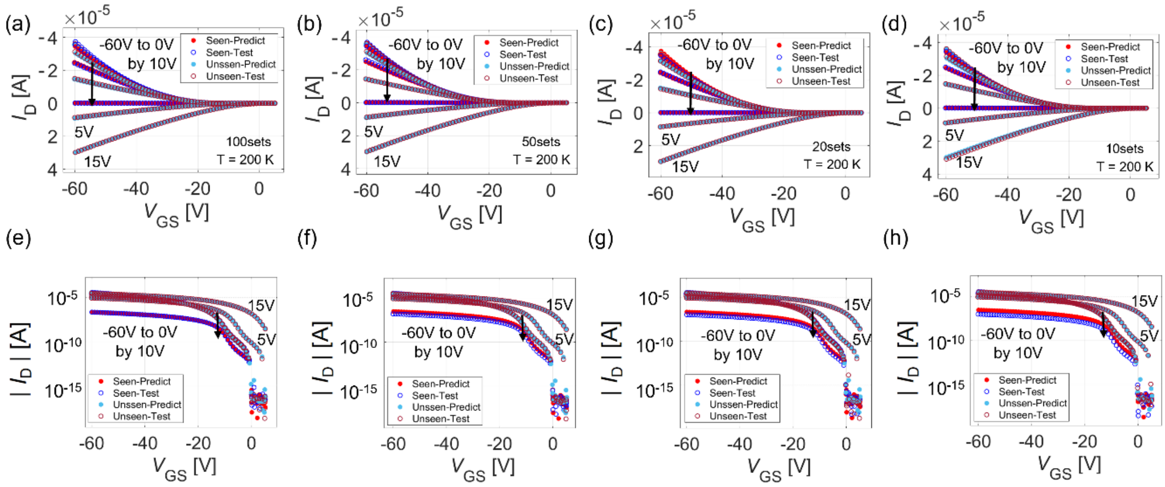

Figure S12. I_D − V_{GS} characteristics saturation (Temperature = 200K) in linear and semilog scale obtained by transfer learning of 100 sets (a) and (b), 50 sets (c) and (d), 20 sets (e) and (f), and 10 sets (g) and (h): Seen-Prediction data (red filled circle), Seen-Test data (blue open-circle), Unseen-Prediction data (blue filled circle), and Unseen-Test data (red open circle). The V_{DS} conditions of the data range from -60V to 15V in decrements of 10V. For transfer learning, the (V_{DS} conditions of the seen data are -60V, -40V, -20V and 0V. And the Unseen data are -50V, -30V, -10V, 5V and 15V.

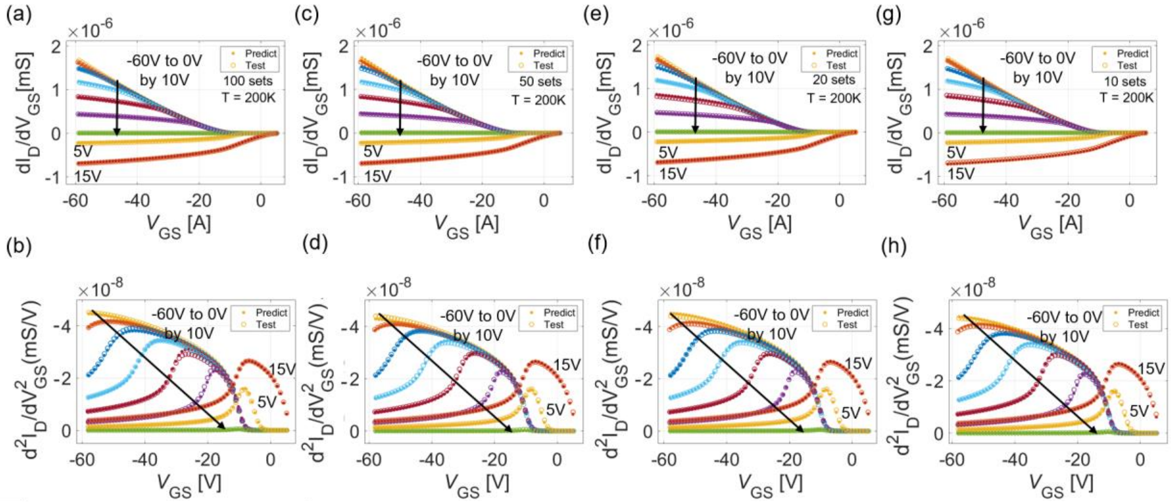

Figure S13. First and second derivative of I_D−V_{GS} characteristics under linear regime (Temp = 200K) obtained by transfer learning of 100 sets (a) and (b), 50 sets (c) and (d), 20 sets (e) and (f), and 10 sets (g) and (h).

Table S2. Time experiment result of each data samples and MSE loss, R squared and MAPE (percent, of the on region). Transfer learning in the drain voltage domain.

	100 SETS	50 SETS	20 SETS	10 SETS
PRE-TRAIN TIME	12h 11m 13s	12h 11m 13s	12h 11m 13s	12h 11m 13s
FINE-TUNE TIME	7h 24m 56s	5h 08m 07s	3h 18m 46s	2h 48m 14s
TOTAL TIME	7h 24m 56s	5h 08m 07s	3h 18m 46s	2h 48m 14s
MSE LOSS	0.0318	0.0713	0.112	0.132
R-SQUARED	0.998	0.971	0.915	0.901
MAPE (%)	7.28	11.16	15.9	28.7

Reference

[S1] Hinton, G.; Srivastava, N.; Swersky, K. Lecture 6e: RMSprop: Divide the gradient by a running average of its recent magnitude. https://www.cs.toronto.edu/~tijmen/ csc321/slides/lecture_slides_lec6.pdf.

[S2] Ruder, S. An overview of gradient descent optimization algorithms. 2017.

[S3] Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2017.

[S4] Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. 2012.