Supplementary Material

Nested deep transfer learning for modeling of multilayer thin films

Rohit Unni,[1,2] Kan Yao,[1,2] and Yuebing Zheng[1,2,*]

[1]Walker Department of Mechanical Engineering, The University of Texas at Austin, Austin, Texas 78712, USA

[2]Texas Materials Institute, The University of Texas at Austin, Austin, Texas 78712, USA

[*]Corresponding Author: zheng@austin.utexas.edu

This Supplementary Material includes:

1. **Model architectures and hyperparameters**
2. **Comparative studies**
3. **Post-processing method**

1. Model architectures and hyperparameters

The final forward model architecture consists of three bidirectional long short-term memory (LSTM) layers followed by three fully connected layers before the final output. The input takes an Nx5 array, with N in the first dimension representing the number of structure layers, and the second dimension encoding the dummy variables for the four materials and the scaled thickness. The LSTM layers have 128 units each, and the fully connected layers start with 500 neurons decreasing by 100 each layer until the final output dimension of 300.

The inverse model uses a series of convolutional and pooling layers, fully connected layers, and a final MDN layer for prediction. There are five pairs of convolutional and max pooling layers. All convolutional and pooling layers have 32 filters. The convolutional layers start with a dimension of 300 at the input and shrink by a factor of two after each pooling layer, rounding up if needed, finally reaching a dimension of 32x19 before being flattened and connected to the fully connected layers. The kernel for the first three convolutional layers has a dimension of 1x5 and the subsequent two have a dimension of 1x3. The fully connected layers start with 1000 neurons then decrease by 100 for the next two layers. The final 800 neuron layer is connected to 11 outputs, 10 of which have a dimension of four, representing the likelihood of a given layer being a certain material. The last output is the MDN layer, which uses 32 mixtures. These 32 mixtures are encoded by 672 neurons representing the mean and variance for each mixture for each of the 10 thickness variables, plus 32 mixture parameters. All models use the Adam optimizer with an initial learning rate of 0.01. A learning rate scheduler is used, which manually reduces the learning rate by 30% if the validation loss does not decrease for 7 consecutive epochs. The full architecture diagrams of both models are shown in Figure S1.

2. Comparative studies

One comparative study varied the number of weight layers in the network that were transferred between subsequent models. This transfer procedure used a step size of 2 between different structure layer numbers. For this, a dataset of 10,000 samples was generated for each model, split into 70% training and 30% validation, and the model is trained with the same hyperparameters as specified in the main text for 300 epochs. The loss results for the final 30-layer case are shown for each of the different weight transfer configurations in Figure S2. From this we found a general trend that transferring more weight layers in the network up to 4 layers reduced the RMSE, but further than this, the loss increased. This was tested under different conditions with material choices and wavelength ranges, as well as with deeper and shorter networks, and the consistent trend was that transferring up to the penultimate layer yielded the best results, but transferring all layers was not optimal.

A second study compared the step-size in model complexity, how many more layers the thin film structure for the next model had compared to the previous one. Step-sizes of one through five were tested going from a starting structure of 6 layers all the way up to 30 layers. In the step-size five case, since this didn't divide evenly to each 30 layers, the final transfer was only by four layers. For this, four weight layers were transferred at each step, and the same 10,000 samples split into training and validation were used. The results at the 30-layer training are shown in Figure S3. The conclusion was that a step size of four offered the best results at higher layer numbers. The final forward model used in the main text uses 4-layer weight transfer with a step size of 4 between successive models. A 30-layer structure forward model was also trained with the same dataset size without the use of the nested transfer method. The model without transfer converged to a much higher error than the model using nested transfer (Fig. S4). A further comparison between the two models using other regression metrics is shown in Table S1.

A comparison between the model setup used in this work and those used in other works featuring transfer learning for thin film structures is shown in Table S2. Of note, while the number of data used here is higher, the difficulty of the modeling task is significantly greater. The number of layers increases by fivefold from the base case, with the complexity of modeling raising non-linearly with the number of layers. The addition of free material choice at every layer and modeling both continuous and categorical variables in the same model is also a significant contributor to the higher data requirements. The higher complexity afforded with the greater number of layers and material choice allows the final model to be more flexible in modeling cases with real-world applicability.

3. Post-processing method

For the post-processing procedure, first the MDN takes the desired optical spectrum, denoted as R, as input and generates N probability distributions at the output, each corresponding to one design variable. The initial guess of the design is obtained by assigning each variable the value at the most prominent peak of its respective distribution, without requiring complex sampling strategies. Next, the initial design suggestion is forwarded to the forward network to predict the optical response, denoted as R_M, which represents the response of the active candidate design and is labeled as R_0. The performance of this design is evaluated by comparing R_0 with the ground truth optical spectrum R, using cosine similarity as a comparison metric. The optimization process commences, wherein one of the N variables of the active candidate design is randomly chosen and resampled a specified number of times, based on its probability distribution, to generate new guesses. The remaining N-1 variables are kept fixed during this resampling. Whenever the predicted response of a new guess R_i is closer to R than R_0, that guess becomes the new active candidate design, and R_0 is updated accordingly. The resampling and evaluation process repeats for all N variables in a random order, and this cycling through all variables can be repeated multiple times. If the forward model is accurate enough, the prediction of the design guesses' error relative to the ground truth will closely approximate the true error, leading to continuous improvement of the design over time. It is worth noting that throughout the design process, simulation is only employed once at the very end for verification, when the finalized design R' is simulated using an electromagnetic solver to compute its actual properties. Forward models were trained separately for both the standard dataset which included dielectrics and oxides on a glass substrate, and the thermal applications dataset, which included tungsten in the layer materials. These models were trained from scratch for 200 epochs with the larger datasets used for the inverse model. Since this is being used for a forward model, the target output data from the inverse model are used as the input, and vice versa. Both models converge to an RMSE below 0.02, allowing for near perfect reproduction of the target spectra. The training results for the forward models are shown in Figure S5.

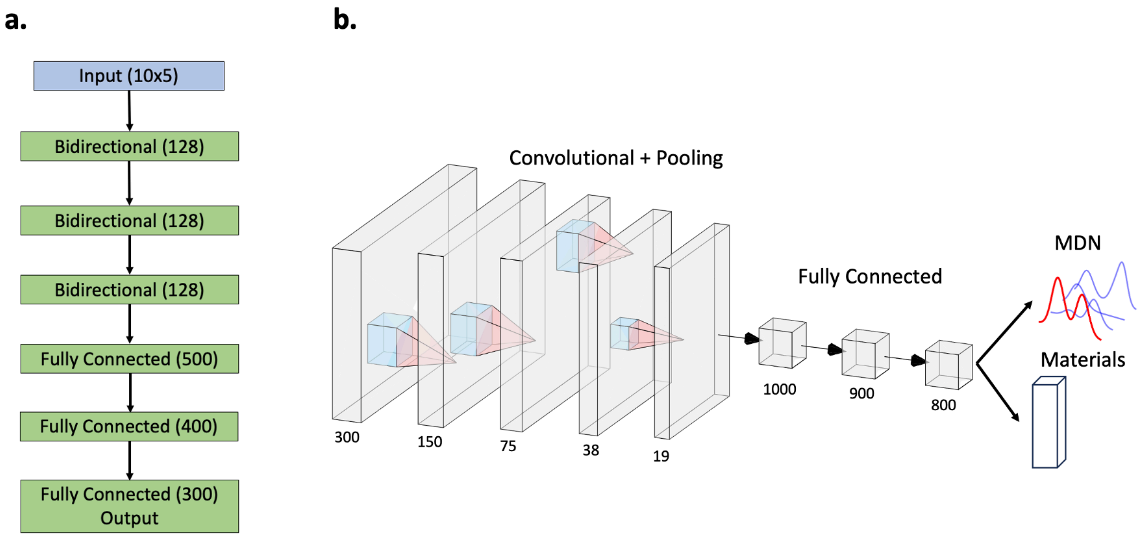


Figure S1: (a) Diagram of architecture used for the forward model. Three bidirectional LSTM layers are followed by three fully connected layers. The numbers in parentheses represent the number of units or neurons for each layer. (b) Diagram of architecture for the inverse model. Convolutional and pooling layers are followed by fully connected layers. The numbers on the bottom represent the dimension of the convolutional and pooling layers and the number of neurons in the fully connected layers.

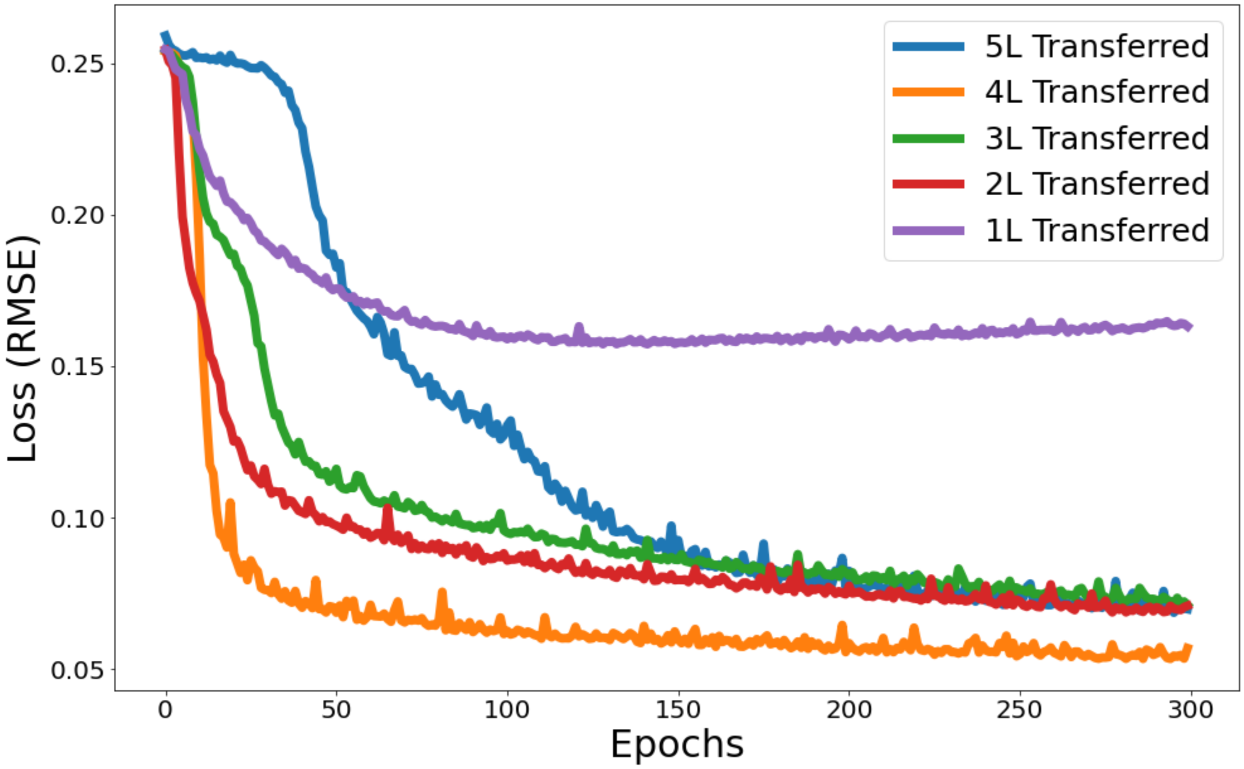


Figure S2: Training curve of the 30-layer over 300 epochs for different weight transfer configurations between subsequent models from an initial 6-layer model, using a step size of two between each model.

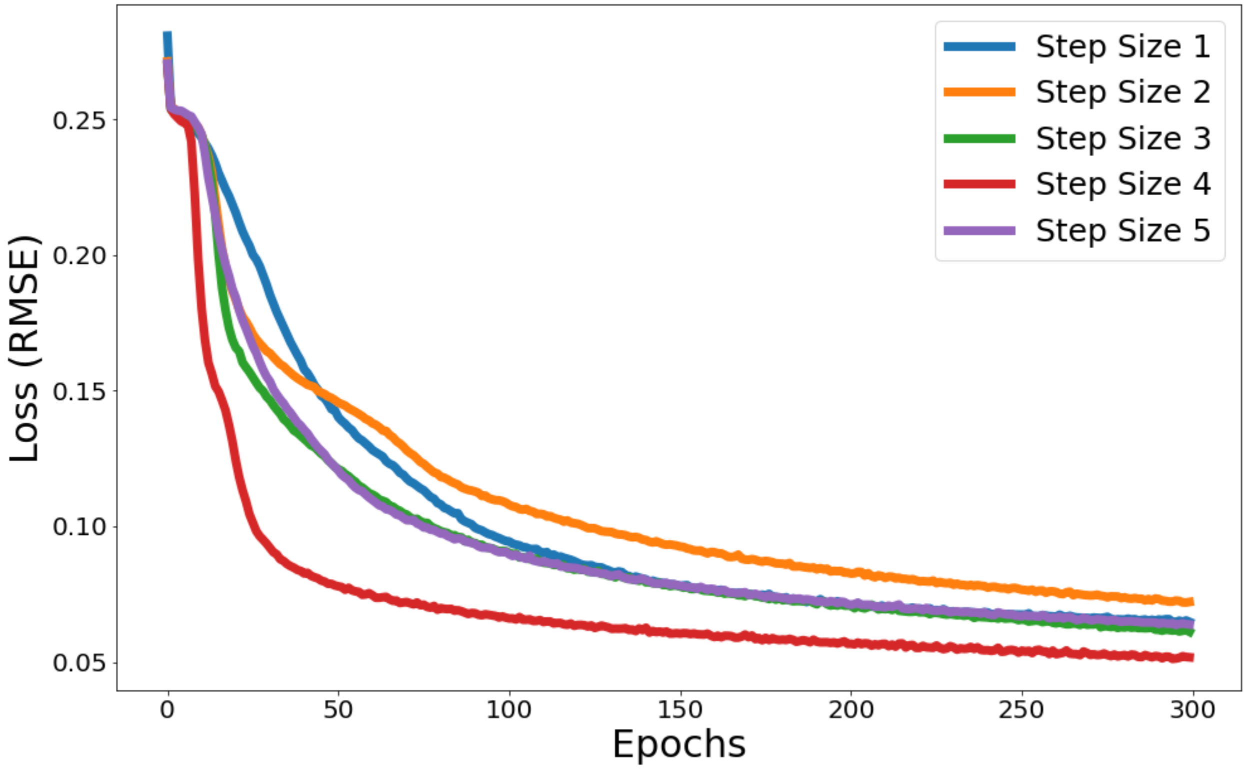


Figure S3: Training curve of the 30-layer over 300 epochs for different step sizes between subsequent models from an initial 6-layer model.

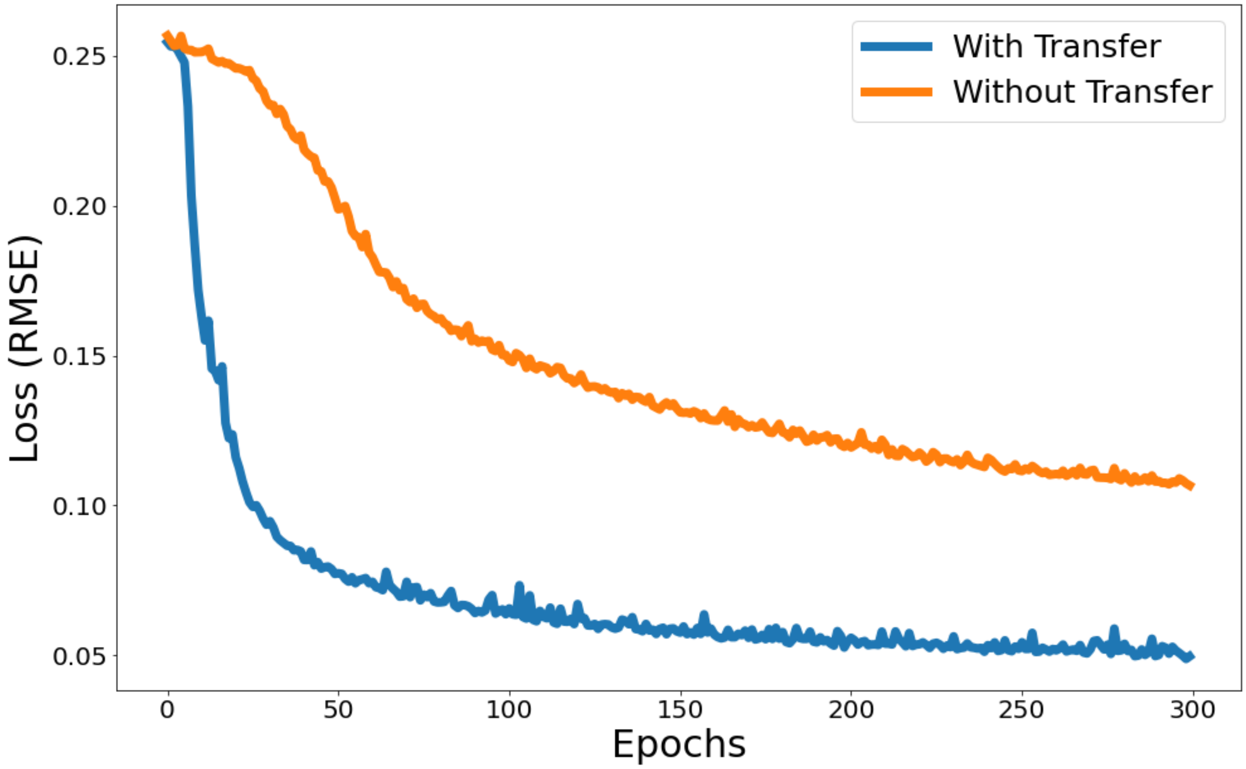
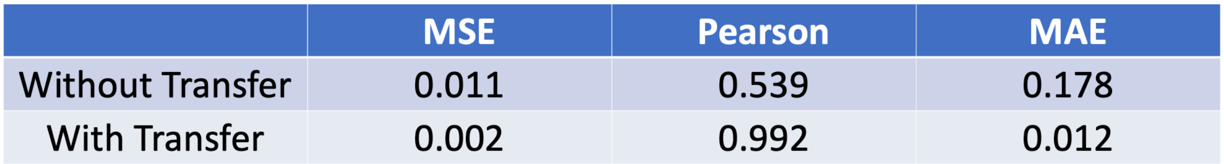
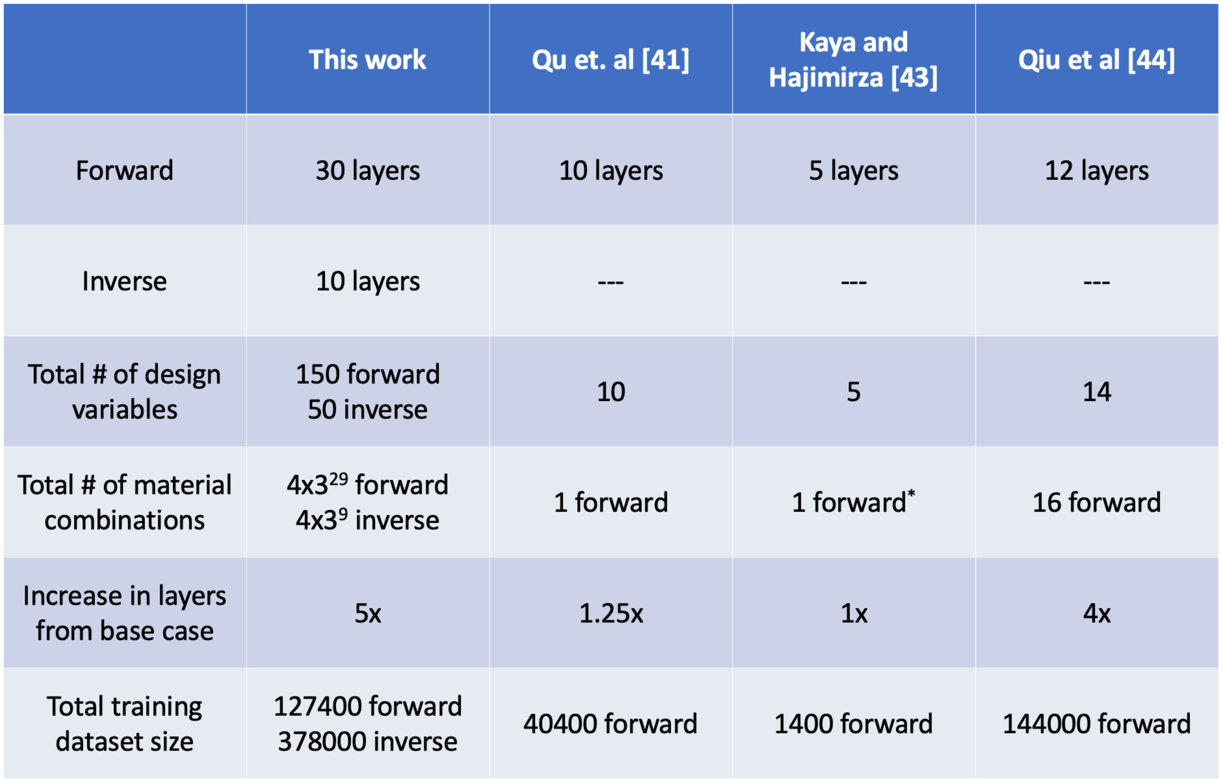


Figure S4: Training curve of two 30-layer forward models over 300 epochs, one using the nested transfer method, and one trained from scratch without transfer.



	MSE	Pearson	MAE
Without Transfer	0.011	0.539	0.178
With Transfer	0.002	0.992	0.012

Table S1: Comparison of 30-layer forward models using and without using nested transfer by different regression metrics. Mean squared error (MSE), Pearson correlation coefficient, and mean absolute error (MAE) were calculated from 5000 random samples from the test dataset.



	This work	Qu et. al [41]	Kaya and Hajimirza [43]	Qiu et al [44]
Forward	30 layers	10 layers	5 layers	12 layers
Inverse	10 layers	---	---	---
Total # of design variables	150 forward 50 inverse	10	5	14
Total # of material combinations	4x3^{29} forward 4x3^{9} inverse	1 forward	1 forward*	16 forward
Increase in layers from base case	5x	1.25x	1x	4x
Total training dataset size	127400 forward 378000 inverse	40400 forward	1400 forward	144000 forward

Table S2: Comparison of dataset and model parameters between this work and other works featuring transfer learning applied to thin film structures. Total training dataset size includes training data for base case and all unique transfer cases. * For Kaya and Hajimirza, transfer is conducted between models for different material combinations, but for any given model, the materials are fixed, as opposed to this work as well as Qiu et. al.

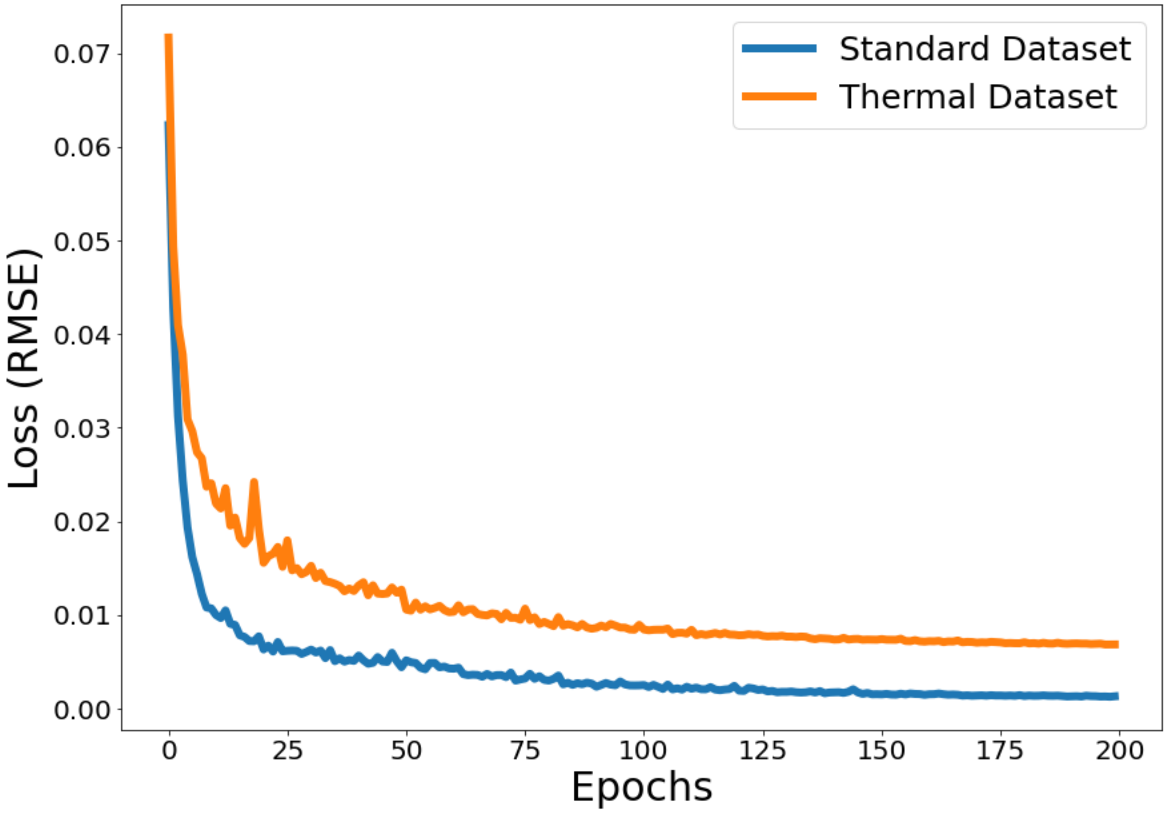


Figure S5: Loss curves for the test datasets for the 10-layer forward models used for the post-processing. The model used for the standard dataset (blue) converges to a lower value than the model for the thermal applications (orange).