# Supplementary Materials for

## Predicting PRDM9 binding sites by a convolutional neural network and verification using genetic recombination map

### Architecture and prediction accuracy without data augmentation

In addition to using data augmentation to train the EBCN, we used two other methods for hyperparameter search, training, and evaluation of prediction accuracy with test data. The first method is that 153,189 DNA fragments of 100-bp length with the ChIP-seq peak in its center were used as positive training data (training method 1). Another one is that the same number as training method 1 of DNA fragments of 100-bp length with ChIP-seq core at a random position were used as positive training data (training method 2). The negative data were prepared so that the ratio of positive data and negative data becomes 1:1. The hyperparameter search and calculation for metrics were done in the same way as the training method with data augmentation. The consequences for training methods 1 and 2 are shown in Table S1–S3 and Table S4–S6 respectively.

Table S1 Optimized network structure of final CNN for training method 1

| Layers | Parameters |
|---|---|
| Equivariant Conv1D Layer 1 | Filters = 34; kernel size = 22; activation function = LReLU |
| Equivariant MC Dropout Layer 1 | Dropout rate = 0.1731 |
| Spatial MaxPooling1D Layer | Pool size = 6 |
| Equivariant Conv1D Layer 2 | Filters = 34; kernel size = 4; activation function = LReLU |
| Equivariant MC Dropout Layer 2 | Dropout rate = 0.1731 |
| Equivariant Conv1D Layer 3 | Filters = 34; kernel size = 4; activation function = LReLU |
| Equivariant MC Dropout Layer 3 | Dropout rate = 0.1731 |
| Equivariant Conv1D Layer 4 | Filters = 34; kernel size = 4; activation function = LReLU |
| Equivariant MC Dropout Layer4 | Dropout rate = 0.1731 |
| Reverse Complement Sum Pooling Layer | None |
| Global Spatial MaxPooling1D Layer | None |
| Dense Layer | Activation function = Softmax |

Table S2 Hyperparameters of backpropagation for training method 1

| Hyperparameter | Optimal Value |
|---|---|
| Learning rate | $1.659 \times 10^{-3}$ |
| L2 coefficient | 0.0011 |
| Batch size | 64 |
| Optimizer | Adam |

Table S3 Metrics on prediction for training method 1

| Metrics | CNN |
|---|---|
| AUC | 0.9125 |
| Accuracy | 0.8422 |
| Sensitivity (True positive rate) | 0.8272 |
| Specificity | 0.8425 |

Table S4 Optimized network structure of final CNN for training method 2

| Layers | Parameters |
|---|---|
| Equivariant Conv1D Layer 1 | Filters = 54; kernel size = 40; activation function = LReLU |
| Equivariant MC Dropout Layer 1 | Dropout rate = 0.2360 |
| Spatial MaxPooling1D Layer | Pool size = 6 |
| Equivariant Conv1D Layer 2 | Filters = 54; kernel size = 2; activation function = LReLU |
| Equivariant MC Dropout Layer 2 | Dropout rate = 0.2360 |
| Equivariant Conv1D Layer 3 | Filters = 54; kernel size = 2; activation function = LReLU |
| Equivariant MC Dropout Layer 3 | Dropout rate = 0.2360 |
| Reverse Complement Sum Pooling Layer | None |
| Global Spatial MaxPooling1D Layer | None |
| Dense Layer | Activation function = Softmax |

Table S5 Hyperparameters of backpropagation for training method 2

| Hyperparameter | Optimal Value |
|---|---|
| Learning rate | $9.652 \times 10^{-2}$ |
| L2 coefficient | 0.0033 |
| Batch size | 32 |
| Optimizer | Momentum |

Table S6 Metrics on prediction for training method 2

| Metrics | CNN |
|---|---|
| AUC | 0.8951 |
| Accuracy | 0.8231 |
| Sensitivity (True positive rate) | 0.8083 |
| Specificity | 0.8234 |