

Figure 1. Diagram of the model construction used for downstream analysis. The linear genome was binned according to the resolution of the respective HiC experiment (either 5kb, 25kb, 50kb, or 100kb intervals). The response vector Y used for classification was determined by whether or not a genomic bin contained a TAD boundary. The positional coordinates of each functional genomic element, obtained from ENCODE, was used to define the feature space of the models.

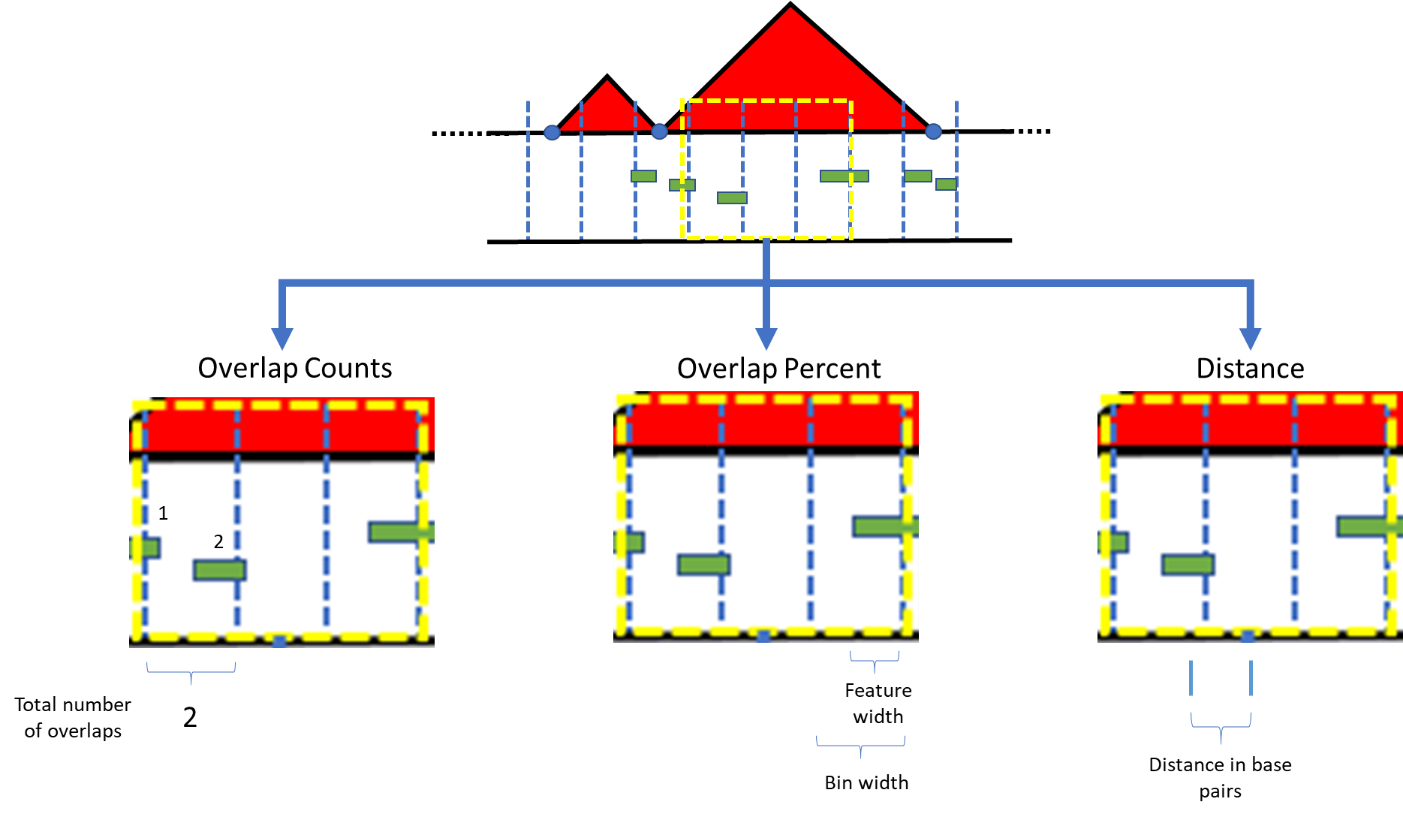


Figure 2. Diagram of the 3 predictor types considered when assessing the relationship between TAD boundaries and functional genomic elements. Each predictor type was used as the feature space in downstream analyses for predicting which functional genomic elements were associated with the formation of TAD boundaries. Featured above are bin-specific examples of the construction of each predictor type. (Left) The overlap count (OC) predictors were calculated by considering the total number of elemental regions that overlapped with each genomic bin. (Middle) The overlap percent (OP) predictors were calculated by dividing the sum of all feature widths within each specific bin and dividing by the total bin width (either 5, 25, 50, or 100 kilobases given resolution of boundary data). (Right) The distance predictors were calculated by measuring the distance (in base pairs) from the center of each genomic bin to the center of the nearest elemental region of interest.

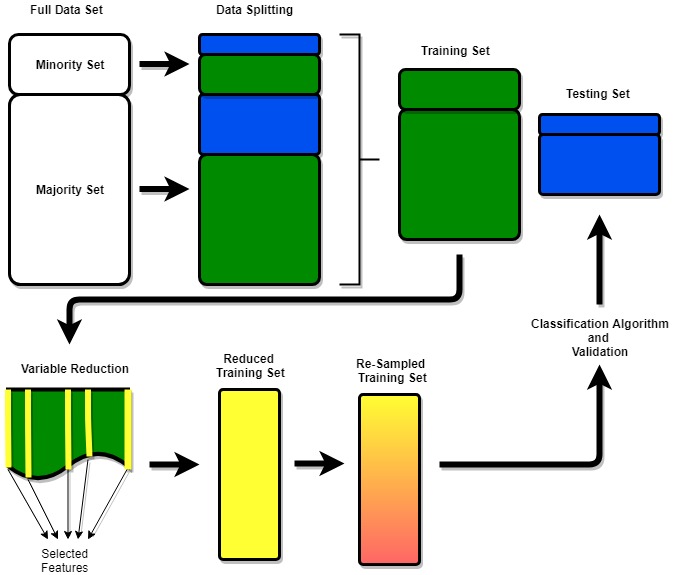


Figure 3. A diagram of the model building pipeline. The process was the same for each combination of predictor type, variable reduction technique, and resampling method considered. The data was split into a 7:3 training set to testing set ratio. The variable reduction technique of choice was performed (either LASSO or Elastic-Net regularization). The training set was then reduced and a random forest classification algorithm was performed. Each model was validated on the same testing set.

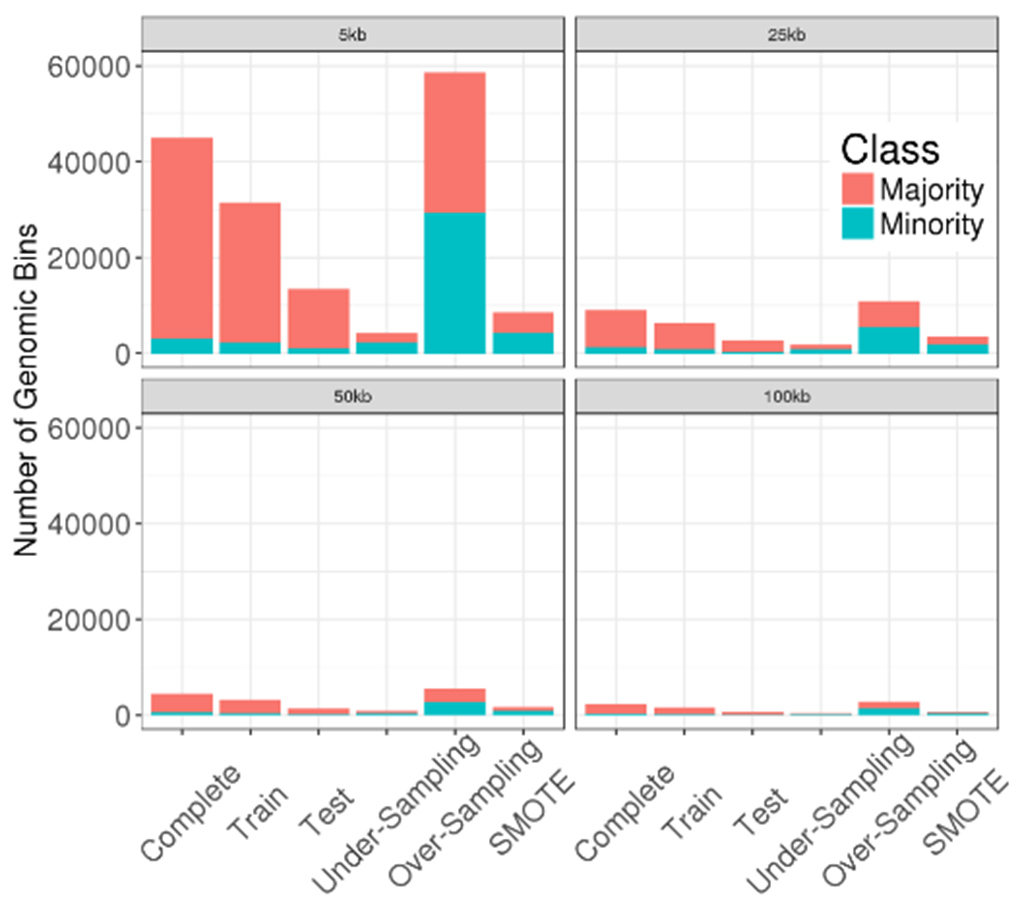


Figure 4. Barplots illustrating the class imbalance problem featured across each of the four different resolutions that were analyzed. Minority classes represent the number of genomic bins that contain a TAD boundary, while the majority classes represent the number of genomic bins that do not contain a TAD boundary. As the resolution increases, the subsequent sizes of the genomic bins also increase. This contributes to less genomic bins throughout the genome of the GM12878 cell line. This is evident by the decrease in sizes of the bars as the resolution increases.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Genomic Bin** | **Complete** | **Training** | **Testing** | **No Balancing** | **Random Under-Sampling** | **Random Over-Sampling** | **SMOTE** |
| **5 kb** |  |  |  |  |  |  |  |
| With TAD boundary | 3083 | 2140 | 943 | 2140 | 2140 | 29323 | 4280 |
| Without TAD boundary | 41865 | 29323 | 12542 | 29323 | 2140 | 29323 | 4280 |
| Total | 44948 | 31463 | 13485 | 31463 | 4280 | 58646 | 8560 |
| **25 kb** |  |  |  |  |  |  |  |
| With TAD boundary | 1266 | 895 | 371 | 895 | 895 | 5434 | 1790 |
| Without TAD boundary | 7776 | 5434 | 2342 | 5434 | 895 | 5434 | 1790 |
| Total | 9042 | 6329 | 2713 | 6329 | 1790 | 10868 | 3580 |
| **50 kb** |  |  |  |  |  |  |  |
| With TAD boundary | 621 | 431 | 190 | 431 | 431 | 2747 | 862 |
| Without TAD boundary | 3920 | 2747 | 1173 | 2747 | 431 | 2747 | 862 |
| Total | 4541 | 3178 | 1363 | 3178 | 862 | 5494 | 1724 |
| **100 kb** |  |  |  |  |  |  |  |
| With TAD boundary | 277 | 185 | 92 | 185 | 185 | 1410 | 370 |
| Without TAD boundary | 2002 | 1410 | 592 | 1410 | 185 | 1410 | 370 |
| Total | 2279 | 1595 | 684 | 1595 | 370 | 2820 | 740 |

Table 1. Summary of the data used across each resolution for each re-sampling technique. Each complete data set was split into a 7:3 training to testing ratio. All re-sampling techniques were applied to the training set, while the testing set was used to validate the model. Each re-sampling technique produced perfectly balanced classes, as evidenced by the 50/50 ratio of majority/minority classes across each resolution.

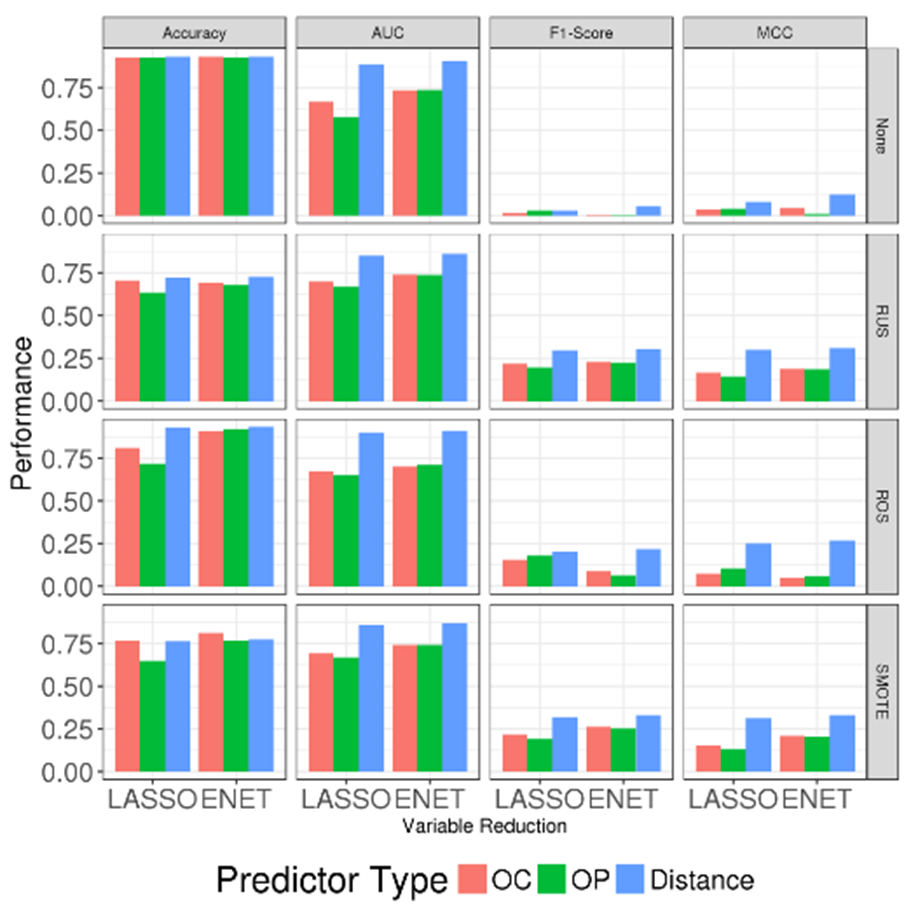


Figure 4. Model performances for TAD boundary data at 5kb resolution across each re-sampling technique. A total of four metrics were used to assess model performance: accuracy, AUC, F1-score, and MCC. Within each plot, models are evaluated across specific predictor types and variable reduction techniques.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | LASSO | | | Elastic-Net | | |
|  | Balancing Technique | OC | OP | Distance | OC | OP | Distance |
| Accuracy | None | 0.929 | 0.927 | 0.930 | 0.930 | 0.93 | 0.931 |
| RUS | 0.706 | 0.633 | 0.723 | 0.692 | 0.679 | 0.729 |
| ROS | 0.81 | 0.718 | 0.933 | 0.907 | 0.922 | 0.934 |
| SMOTE | 0.768 | 0.648 | 0.764 | 0.812 | 0.768 | 0.775 |
| AUC | None | 0.669 | 0.577 | 0.888 | 0.734 | 0.737 | 0.902 |
| RUS | 0.700 | 0.669 | 0.853 | 0.74 | 0.738 | 0.862 |
| ROS | 0.673 | 0.651 | 0.901 | 0.701 | 0.711 | 0.912 |
| SMOTE | 0.694 | 0.668 | 0.858 | 0.741 | 0.741 | 0.867 |
| F1-Score | None | 0.018 | 0.03 | 0.031 | 0.004 | 0.002 | 0.055 |
| RUS | 0.218 | 0.197 | 0.297 | 0.229 | 0.225 | 0.303 |
| ROS | 0.156 | 0.181 | 0.203 | 0.089 | 0.063 | 0.219 |
| SMOTE | 0.217 | 0.192 | 0.318 | 0.263 | 0.252 | 0.332 |
| MCC | None | 0.038 | 0.04 | 0.081 | 0.044 | 0.01 | 0.124 |
| RUS | 0.166 | 0.143 | 0.300 | 0.189 | 0.186 | 0.309 |
| ROS | 0.073 | 0.105 | 0.252 | 0.050 | 0.059 | 0.268 |
| SMOTE | 0.153 | 0.132 | 0.313 | 0.209 | 0.205 | 0.331 |

Table 2.

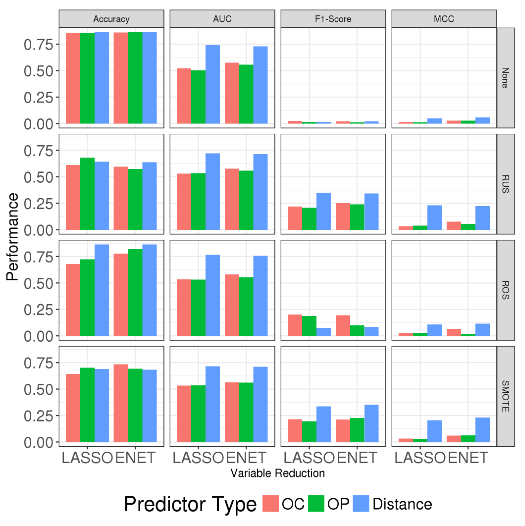


Figure 5. Model performances for TAD boundary data at 25kb resolution across each re-sampling technique, between different predictor types and variable reduction techniques.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | LASSO | | | Elastic-Net | | |
|  | Balancing Technique | OC | OP | Distance | OC | OP | Distance |
| Accuracy | None | 0.857 | 0.86 | 0.863 | 0.861 | 0.863 | 0.863 |
| RUS | 0.611 | 0.679 | 0.642 | 0.596 | 0.574 | 0.637 |
| ROS | 0.678 | 0.724 | 0.863 | 0.777 | 0.82 | 0.863 |
| SMOTE | 0.644 | 0.703 | 0.691 | 0.734 | 0.693 | 0.684 |
| AUC | None | 0.521 | 0.502 | 0.74 | 0.574 | 0.558 | 0.73 |
| RUS | 0.53 | 0.535 | 0.723 | 0.578 | 0.557 | 0.716 |
| ROS | 0.533 | 0.532 | 0.765 | 0.58 | 0.556 | 0.756 |
| SMOTE | 0.533 | 0.535 | 0.717 | 0.565 | 0.561 | 0.71 |
| F1-Score | None | 0.025 | 0.016 | 0.016 | 0.021 | 0.011 | 0.021 |
| RUS | 0.22 | 0.208 | 0.347 | 0.252 | 0.24 | 0.343 |
| ROS | 0.202 | 0.187 | 0.075 | 0.195 | 0.102 | 0.085 |
| SMOTE | 0.215 | 0.197 | 0.336 | 0.212 | 0.227 | 0.352 |
| MCC | None | 0.016 | 0.011 | 0.05 | 0.028 | 0.027 | 0.057 |
| RUS | 0.032 | 0.039 | 0.23 | 0.076 | 0.055 | 0.224 |
| ROS | 0.028 | 0.029 | 0.107 | 0.066 | 0.019 | 0.115 |
| SMOTE | 0.034 | 0.031 | 0.206 | 0.06 | 0.063 | 0.231 |

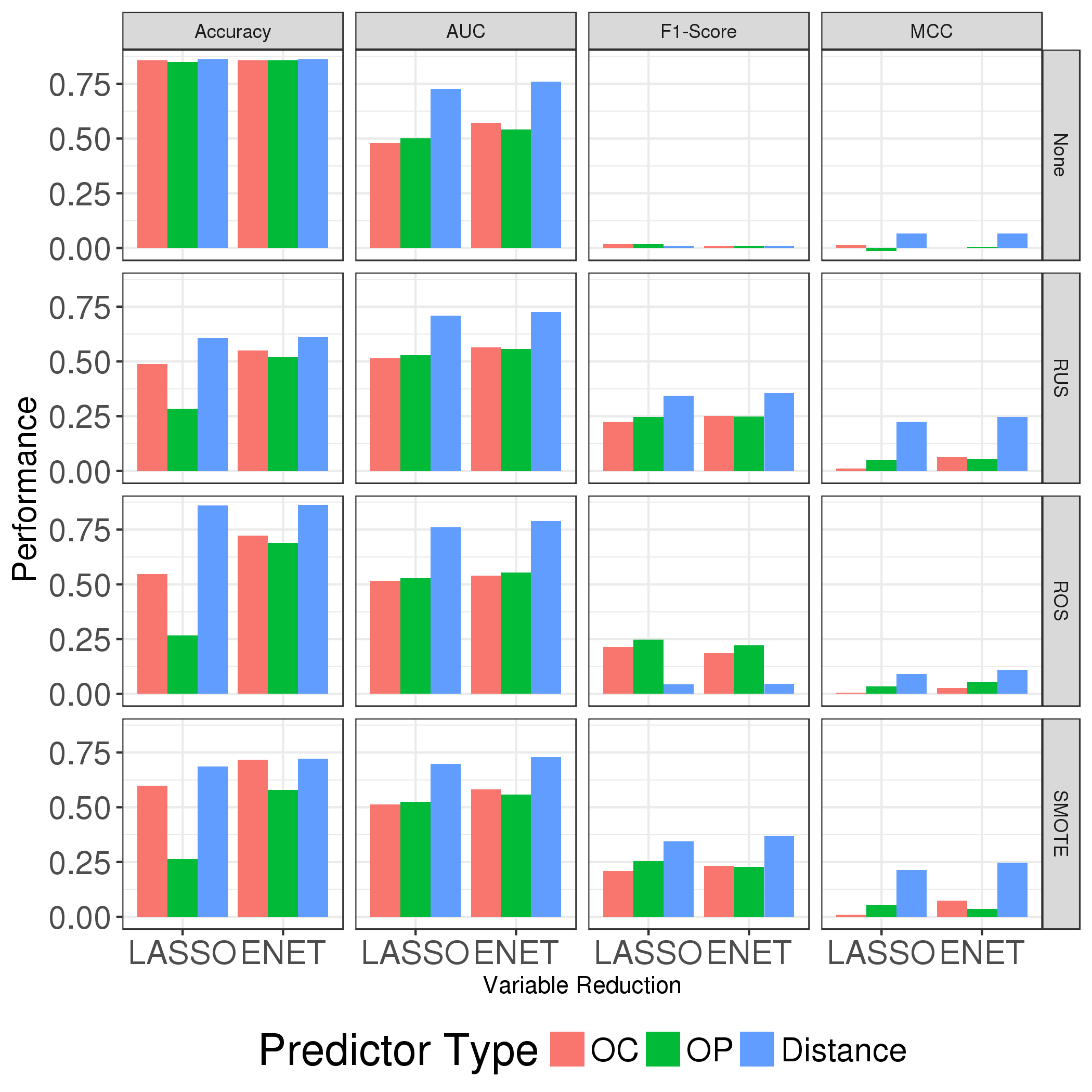


Figure 6. Model performances for TAD boundary data at 50kb resolution across each re-sampling technique, between different predictor types and variable reduction techniques.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | LASSO | | | Elastic-Net | | |
|  | Balancing Technique | OC | OP | Distance | OC | OP | Distance |
| Accuracy | None | 0.856 | 0.849 | 0.861 | 0.857 | 0.858 | 0.861 |
| RUS | 0.488 | 0.284 | 0.608 | 0.549 | 0.519 | 0.611 |
| ROS | 0.547 | 0.267 | 0.861 | 0.723 | 0.689 | 0.862 |
| SMOTE | 0.599 | 0.263 | 0.686 | 0.718 | 0.58 | 0.721 |
| AUC | None | 0.479 | 0.5 | 0.726 | 0.57 | 0.541 | 0.759 |
| RUS | 0.514 | 0.528 | 0.708 | 0.564 | 0.557 | 0.726 |
| ROS | 0.515 | 0.528 | 0.761 | 0.539 | 0.555 | 0.79 |
| SMOTE | 0.513 | 0.524 | 0.698 | 0.582 | 0.557 | 0.728 |
| F1-Score | None | 0.02 | 0.019 | 0.01 | 0.01 | 0.01 | 0.01 |
| RUS | 0.225 | 0.246 | 0.344 | 0.251 | 0.248 | 0.355 |
| ROS | 0.214 | 0.248 | 0.044 | 0.185 | 0.221 | 0.047 |
| SMOTE | 0.208 | 0.253 | 0.344 | 0.232 | 0.229 | 0.369 |
| MCC | None | 0.015 | -0.014 | 0.067 | 0.001 | 0.005 | 0.067 |
| RUS | 0.012 | 0.048 | 0.225 | 0.064 | 0.055 | 0.246 |
| ROS | 0.006 | 0.033 | 0.09 | 0.026 | 0.052 | 0.11 |
| SMOTE | 0.009 | 0.054 | 0.213 | 0.074 | 0.035 | 0.247 |

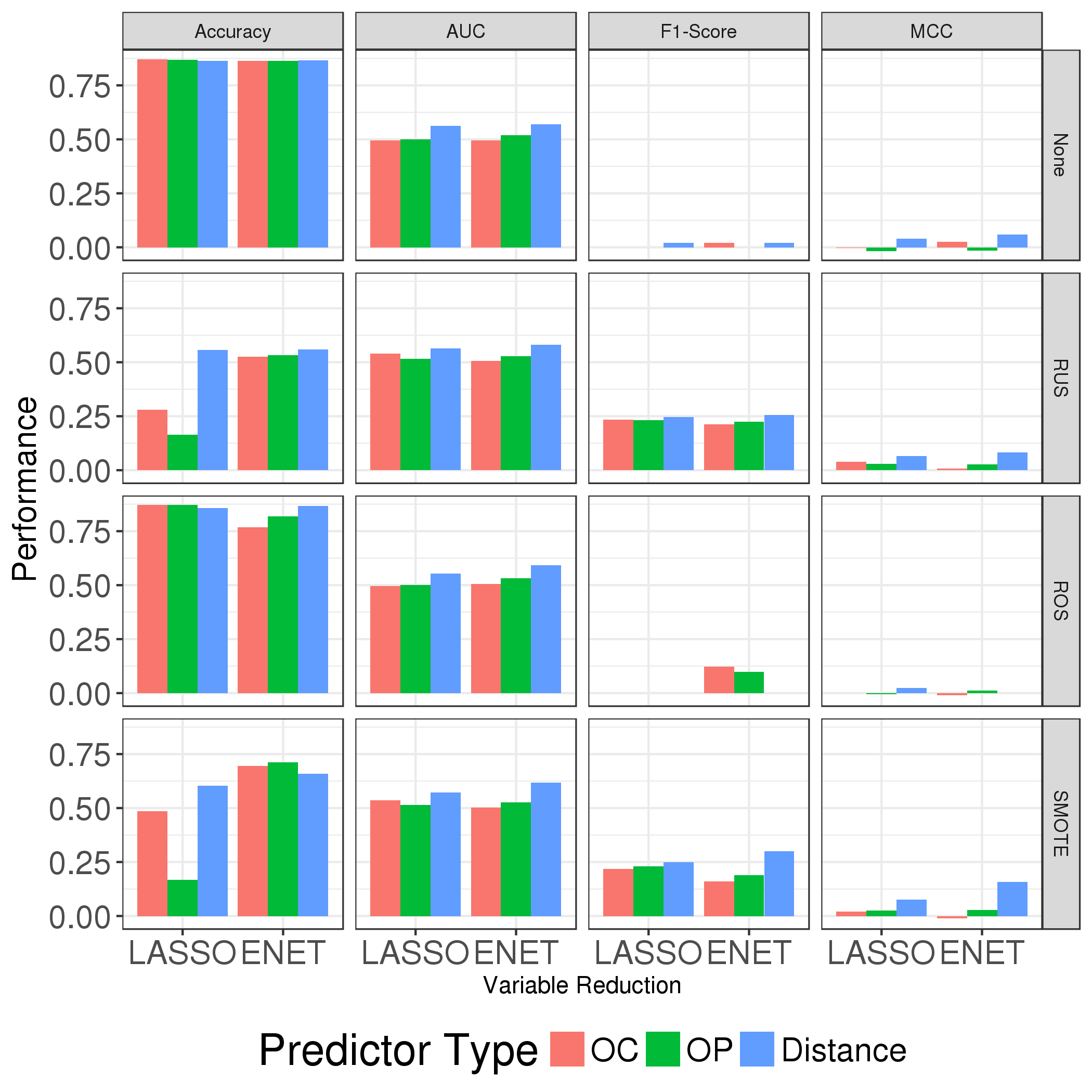


Figure 7. Model performances for TAD boundary data at 100kb resolution across each re-sampling technique, between different predictor types and variable reduction techniques.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | LASSO | | | Elastic-Net | | |
|  | Balancing Technique | OC | OP | Distance | OC | OP | Distance |
| Accuracy | None | 0.871 | 0.869 | 0.864 | 0.863 | 0.864 | 0.865 |
| RUS | 0.279 | 0.165 | 0.556 | 0.525 | 0.534 | 0.559 |
| ROS | 0.871 | 0.871 | 0.858 | 0.767 | 0.819 | 0.867 |
| SMOTE | 0.485 | 0.167 | 0.602 | 0.696 | 0.712 | 0.659 |
| AUC | None | 0.496 | 0.501 | 0.563 | 0.494 | 0.519 | 0.57 |
| RUS | 0.539 | 0.515 | 0.563 | 0.507 | 0.528 | 0.581 |
| ROS | 0.497 | 0.5 | 0.554 | 0.506 | 0.531 | 0.592 |
| SMOTE | 0.535 | 0.513 | 0.572 | 0.502 | 0.526 | 0.618 |
| F1-Score | None | NA | NA | 0.021 | 0.021 | NA | 0.021 |
| RUS | 0.233 | 0.231 | 0.245 | 0.213 | 0.224 | 0.255 |
| ROS | NA | NA | NA | 0.123 | 0.099 | NA |
| SMOTE | 0.218 | 0.231 | 0.249 | 0.161 | 0.189 | 0.3 |
| MCC | None | -0.004 | -0.018 | 0.039 | 0.026 | -0.015 | 0.058 |
| RUS | 0.038 | 0.03 | 0.066 | 0.007 | 0.027 | 0.083 |
| ROS | NA | -0.004 | 0.023 | -0.011 | 0.012 | NA |
| SMOTE | 0.02 | 0.025 | 0.076 | -0.01 | 0.028 | 0.158 |

A close up of a map

Description automatically generated

Figure 8. Comparing performances of random forest classification algorithms, across TAD boundary data at each resolution for models using distance-type predictors, with elastic-net regularization and random under-sampling. Performances were aggregated by taking the average of each metric across 50 iterations of random under-sampling.

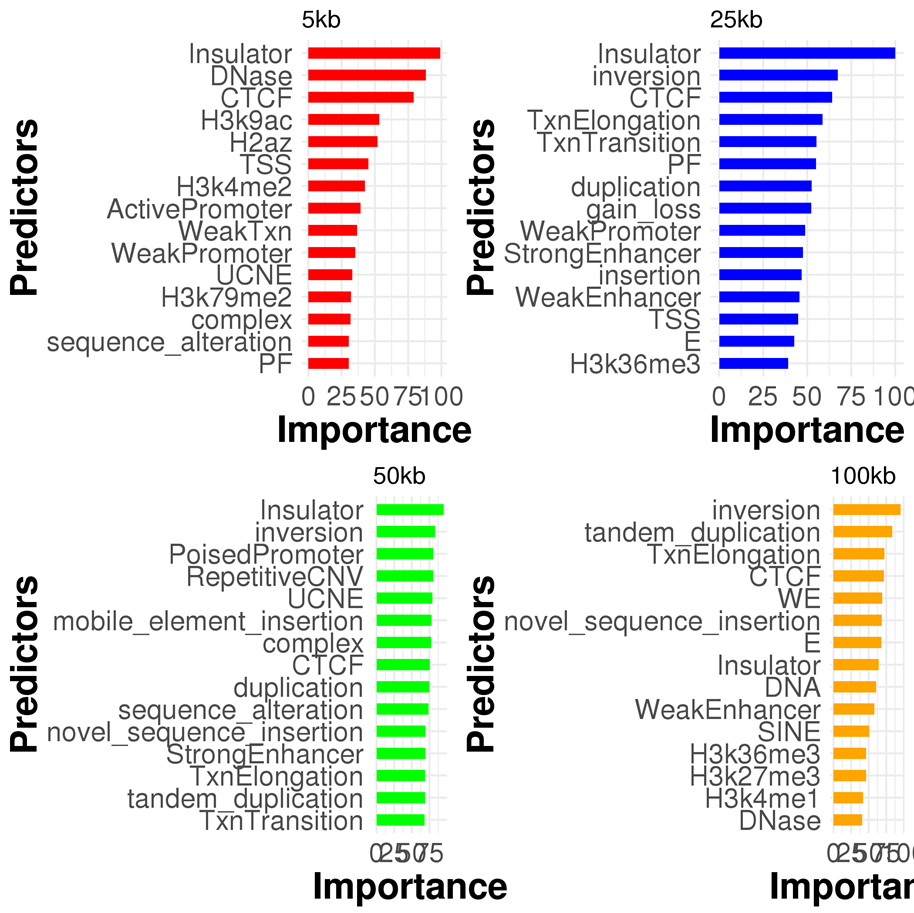


Figure 9. Variable importance plots for the top 15 most predictive functional genomic elements for each of the 4 different resolutions. The x-axis represents the standardized difference between the out-of-bag prediction accuracy after permuting each predictor variable, averaged across all trees. The greater the mean standardized difference, the more importance the predictor is to the model.

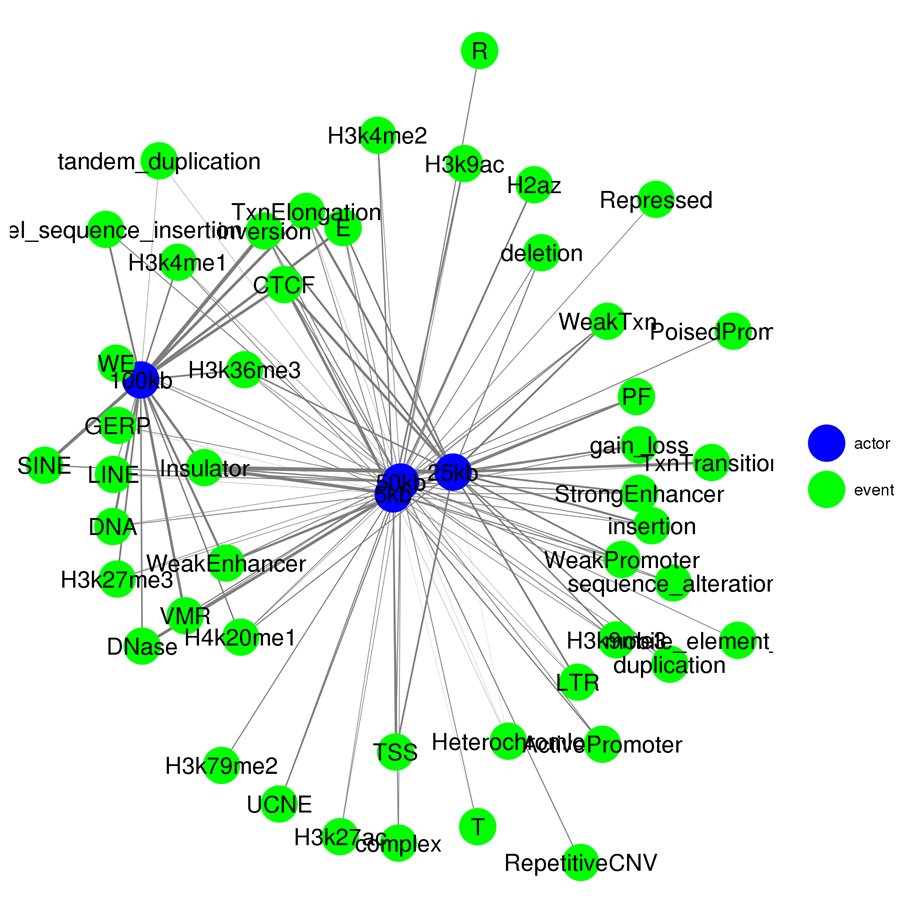


Figure 10. Network visualization of the variable importance for each functional genomic elements between each of the 4 data resolutions.