Development of Semi-Supervised Machine Learning Models to Predict Enhancer Regions in Polygenic Developmental Diseases

Savitha Srinivasan

Abstract: Polygenic developmental diseases affecting humans like autism spectrum disorder often have unknown causes and are difficult to cure. Gene expression and regulation is among the most important mechanistic elements to study. Enhancers, short regulatory stretches of non-coding DNA along chromosomes, significantly impact development through controlling how genes are spatially and temporally expressed. Thus, identifying enhancer regions precisely is key to better elucidating the mechanisms of complex developmental disorders but only a few enhancers have been found with in-vivo techniques. Hence computational approaches to find enhancer regions are appealing.

In this study, multiple genomic datasets from the VISTA Enhancer and UCSC Genome Browsers were integrated to yield 2200 mus musculus (mouse) enhancers, 10% of which are limb enhancers. Features were extracted by calculating RPKM values for gene expression signatures associated with the enhancers. Supervised machine learning models were developed to baseline the performance of multiple classifiers on limb enhancer prediction. The efficacy of five approaches to address dataset imbalance was systematically investigated. The neural network ensemble developed in this study surpasses prevailing precision/recall rates and was further improved with a newly proposed technique to architect ensemble models based on input zones. New candidate limb enhancers were identified using an algorithm developed with the model's predictions. Finally, semi-supervised learning techniques were investigated to gauge their effectiveness in improving model performance with unlabeled data. The self-training approach enables models to be improved as unlabeled enhancer regions are discovered, thereby supplementing in-vivo techniques effectively.

Introduction: Polygenic developmental diseases like autism spectrum disorder that affect humans often have unknown causes and are difficult to cure[10,17,20]. Scientists have taken a multi-pronged approach to elucidate their mechanisms, combining traditional biological and chemistry techniques with computational ones. Gene expression and regulation is one of the most significant mechanistic elements to study. Importantly, enhancer elements, which are short stretches of non-coding DNA along chromosomes, significantly impact organisms' development through controlling how genes are spatially and temporally expressed in complex developmental disorders[2,3,13,14]. Enhancer regions bind to protein activators and can increase the probability that a given gene will be transcribed. Thus, mutations in enhancers can cause diseases[7,18]. For instance, an enhancer (ZRS) 1 Mb upstream of the SHH gene in an intron within another gene has been directly implicated in preaxial polydactyly (PPD), the most frequent limb malformation and most common of all developmental malformations[6]. Therefore, to better elucidate these kinds of mechanisms, it is important to be able to identify enhancer regions with precision and accuracy.

Identifying enhancers in mammalian genomes in vivo is a challenging task. Many potential enhancer regions are unknown and only a few have been identified as well as validated in vivo[15]. Wetlab biological and chemical assays are limited in throughput[12]. This is why researchers have turned to computational approaches. Machine learning based computational approaches are especially effective. In this project, two machine learning approaches—supervised and semi-supervised methods— were explored, each with multiple classifiers. In supervised machine learning, all training data is labeled and the algorithms learn to predict the output from the input training data. In semi-supervised learning, some of the training data is labeled but most of it is unlabeled. The model is initially trained with the labeled data and subsequently refined with the predictions it makes on the unlabeled training data. Semi-supervised learning is useful in applications with large amounts of unlabeled data and thus, holds promise for the limb enhancer prediction task because most enhancers have not been classified.

Consequently, the aim of this investigation was to develop computational machine learning methods that can accurately, sensitively, and precisely predict limb enhancer elements. Both supervised and semi-supervised machine learning models were examined, and a variety of predictive classifiers was systematically explored. Since the training data set was moderately imbalanced (8.4:1), different methods to counter data imbalance were examined to improve model generalization. A second goal of the investigation was to determine whether it was effective to use unlabeled enhancer data to strengthen the computational model for predictions because of the real-world relevance of the scenario. Hence, semi-supervised learning was investigated. The ultimate goal of this investigation was to develop techniques to architect precise computational models and an algorithm that leverages the models to find new candidate limb enhancer regions in the genome that are unknown to date.

Data Acquisition: The challenge of building a model to predict enhancer regions was modeled by using mouse (mm10) genomic data because the mouse genome is very similar to the human genome. Specifically, the classification task at hand was determining whether a given enhancer region was involved in regulating limb activity in a developing mouse embryo or not.

Datasets: Mouse (mm10) genomic data was gathered from the UCSC Genome Browser[21] and a reference validated list of enhancers was found in the VISTA Enhancer Browser[23] to aggregate 2201 mouse enhancer element regions. 234 of these enhancer regions were known to have involvement in regulating limb activity in the developing mouse embryo and were labeled positive. The remaining 1967 regions had no involvement in regulating mouse embryonic limb activity and were labeled negative. Consequently, the dataset was imbalanced with a ratio of 8.4: 1.

Feature Extraction: Fifty features representing chromatin signatures (e.g. binding of the co-activator p300 and the acetylation of lysine residue 27 of histone H3, H3K27ac) in developing mouse embryonic limb tissue at day

11.5 were selected based on literature review[9,15,19,22]. The bedGraph files for these 50 features were obtained and then overlapped with the VISTA enhancer regions to compute the reads per kilobase per million (RPKM) over each enhancer region for all the features (see Figure 1). Finally, as shown in Figure 2, the RPKM values were overlapped with limb activity data to obtain the class labels.

Supervised Learning

Method: Since the enhancer region data set was imbalanced, five different data balancing techniques were investigated: oversampling, undersampling, imbalanced, weighted cost, and undersampling ensemble (Figure 3). Then, seven different classifiers, both linear and non-linear, were tested for each of the data balancing techniques. Thus, 35 different combinations were investigated. The classifiers tested were the following: L1 logistic regression, L2 Logistic Regression (LASSO), Support Vector Machine (SVM) with Linear Kernel, Random Forest, SVM with Radial Basis Function Kernel, and two different neural networks with differing numbers of features and layers. The dataset was split into 80% training and 20% testing and 10-fold stratified cross validation was used to train the supervised models (Figure 4).

Results: Figure 5 shows the AUROC and AUPRC trends for the various classifiers and balancing techniques. The linear classifiers (L1, LASSO and SVM with linear kernel) perform significantly better with the weighted cost balancing technique, while the non-linear classifiers (SVM with radial kernel, Random Forest, and neural networks) perform optimally with the US Ensemble balancing technique. Importantly, as seen in Figures 6 and 7, the neural network with two layers consistently outperforms all the other classifiers for all the data balancing methods, and thus serves as a promising candidate model to refine further. Adaptive boosting and stacked ensembles were also investigated but offered no additional improvement to the accuracy or precision.

Semi-Supervised Learning

Method: The training portion of the data set was split into 40% labeled and 40% unlabeled. After training a model with the labeled data the unlabeled data was classified. Only unlabeled data that met a chosen confidence threshold was added to the training data set (Figure 8). This was done repeatedly until none of the unlabeled data could be classified confidently. Moreover, another data balancing technique (weighted cost with minority samples) was tested in which only labeled samples that were classified as belonging to the minority class were added to the training data set. Thus, 40 models were explored with semi-supervised learning.

Results: As seen in Figure 9, the weighted cost with only minority samples balancing technique improved the Self-Training AUPRC over Supervised Learning for all the classifiers. Most significantly, the neural network with weighted cost performed equally well with 40% training data (semi-supervised) as with 80% training data (supervised) demonstrating that the Supervised Learning model trained with 80% of the dataset can be expected to improve with self training as new unclassified enhancers are discovered. This model would thus significantly reduce wet lab biologists' workload.

Enhancing the Model: Algorithmic Approach

Feature Ranking: The neural network model using the undersampling ensemble (NNE) was chosen for further refinement as it performed the best. To determine which chromatin signatures were most influential in classifying limb enhancers, the Pearson correlation coefficient was calculated between the Z-scores of each input feature and the neural net ensemble model's prediction scores. DNase1, p300, and H3K27ac were found to be the most influential features (Figure 10). This is also corroborated by wetlab research and provides confidence in the model's correctness.

Zoned Model: To further improve the model, the variation of the three principal chromatins across all enhancers was visualized (Figure 11). Figure 11 shows that the DNase Z-Score of most limb enhancers was positive while it

was negative for most non-limb enhancers. Moreover, the misclassifications of the NNE (Figure 12) illustrated that the specificity for the enhancers with positive DNase Z-Scores was very low. Since one-third of the enhancers were in the positive DNase zone and there was only a small data imbalance ratio (2.8 : 1) in the zone, a novel approach for refining the model was investigated. A zoned model ensemble was architected (Figure 13) by training a model on only the enhancers with positive DNase Z-Scores, and combining the results with a model that been trained on all the enhancers. The final zoned NNE developed dramatically increased the PPV and AUPRC, especially in the positive DNase zone which is very useful as most limb enhancers are in that zone (Figure 14) and reduced the misclassifications significantly (compare Figure 15 with Figure 12).

Discovering Enhancers— Algorithmic Approach: A novel algorithm was developed to predict new candidate limb enhancer regions. The algorithm first segmented the genome into overlapping 1 kb segments. Segments that overlapped known enhancers were discarded. The remaining segments were filtered to ensure that they had at least one of the three principal chromatin signatures to validate their biological relevance. The RPKM values for candidate segments were then fed into the final zoned model and prediction scores were obtained. Prediction scores that had a confidence less than 01.6 were discarded. Adjacent segments were merged, and the process above was repeated using a confidence threshold of >0.8 to identify the final limb enhancer predictions. The model identified over 100 new candidate limb enhancer regions (some of which are shown in Figure 16) in Chromosome 9 alone. Figure 17 shows one of the predicted enhancer regions visualized in the UCSC Genome Browser.

Conclusions:

1. The neural network, a deep-learning classifier, performed the best of all classifiers in both supervised and semi-supervised learning models. This is the first known application of neural networks to the enhancer discovery task. The performance of the supervised neural network itself exceeds current published results.
2. The application of semi-supervised learning in computational biology is also novel. Self-training with unlabeled data significantly improves the neural network model. Thus, the model developed here using the Supervised Learning method can continuously improve as new genomic datasets of unlabeled enhancer regions are generated through in vivo techniques. This minimizes the overhead of labeling new training data needed to improve a supervised learning model.
3. The novel algorithmic technique introduced to architect zoned ensembles (ensembles of models trained on different feature-dependent input zones potentially using ***different*** classifiers and methods to counter data imbalance) can generally be more effective in increasing accuracy and precision than traditional stacked ensembles or boosting techniques. Additionally, these novel computational models provide high sensitivity, specificity, and accuracy. The AUPRC, AUROC, and accuracy of the models developed with both the Supervised Learning and the Self Training methods surpass those found in published literature[4,5,9,16].
4. The algorithm to identify new limb enhancer regions leveraging the superior computation models developed in this study will enable the development of significantly improved genome-wide prediction maps of the enhancers active during limb development. Thus, these results will better guide both developmental biologists and human geneticists' research.
5. This is the first study to comprehensively explore the effect of a variety of methods to counter data imbalance across many different kinds of classifiers. The conclusions drawn in this study will guide researchers to thoughtfully consider the alternatives for their investigations.

References

[1] Anuj R. Shah, Christopher S. Oehmen, Bobbie-Jo Webb-Robertson; SVM-HUSTLE—an iterative semisupervised machine learning approach for pairwise protein remote homology detection, Bioinformatics, Volume 24, Issue 6, 15 March 2008, Pages 783–790, https://doi.org/10.1093/bioinformatics/btn028

[2] Banerji, J., Rusconi, S. Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299–308 (1981)

[3] Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;489: 57–74. pmid:22955616

[4] Busser BW, Taher L, Kim Y, Tansey T, Bloom MJ, et al. (2012) A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. PLoS genetics 8: e1002531.

[5] Burzynski GM, Reed X, Taher L, Stine ZE, Matsui T, et al. (2012) Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control. Genome research 22: 2278–2289.

[6] Debbie K. Goode, Philip Snell, Sarah F. Smith, Julie E. Cooke, Greg Elgar, Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3, In Genomics, Volume 86, Issue 2, 2005, Pages 172-181, ISSN 0888-7543, https://doi.org/10.1016/j.ygeno.2005.04.006.

[7] Davidson EH. Emerging properties of animal gene regulatory networks. Nature. Nature Research; 2010;468: 911–920. pmid:21164479.

[8] Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, et al. Integrating diverse datasets improvesdevelopmentalenhancerprediction. PLoSComputBiol. PublicLibraryofScience;2014;10: e1003677. pmid:24967590

[9] Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genet.39, 311–318 (2007)

[10] Guerra, Daniel J. “The Molecular Genetics of Autism Spectrum Disorders: Genomic Mechanisms, Neuroimmunopathology, and Clinical Implications.” Autism Research and Treatment 2011 (2011): 398636. PMC. Web. 15 Dec. 2017.

[11] Jason Weston, Christina Leslie, Eugene Ie, Dengyong Zhou, Andre Elisseeff, William Stafford Noble; Semi-supervised protein classification using cluster kernels, Bioinformatics, Volume 21, Issue 15, 1 August 2005, Pages 3241–3247, https://doi.org/10.1093/bioinformatics/bti497

[12] Maston, A. Glenn, Sara K. Evans, and Michael R. Green. Transcriptional Regulatory Elements in the Human Genome. Annual Review of Genomes and Human Genetics. Vol. 7: 29-59 (Volume publication date 22 September 2006).

[13] Michael Bulger, Mark Groudine, Enhancers: The abundance and function of regulatory sequences beyond promoters, In Developmental Biology, Volume 339, Issue 2, 2010, Pages 250-257, ISSN 0012-1606,

https://doi.org/10.1016/j.ydbio.2009.11.035.
(http://www.sciencedirect.com/science/article/pii/S0012160609014018)

[14] Mike Levine, Transcriptional Enhancers in Animal Development and Evolution, In Current Biology, Volume 20, Issue 17, 2010, Pages R754-R763, ISSN 0960-9822, https://doi.org/10.1016/j.cub.2010.06.070.
(http://www.sciencedirect.com/science/article/pii/S0960982210008560)

[15] Monti R, Barozzi I, Osterwalder M, Lee E, Kato M, Garvin TH, et al. (2017) Limb-Enhancer Genie: An accessible resource of accurate enhancer predictions in the developing limb. PLoS Comput Biol13(8): e1005720. https://doi.org/10.1371/journal.pcbi.1005720

[16] Narlikar L, Sakabe NJ, Blanski AA, Arimura FE, Westlund JM, et al. (2010). Genome-wide discovery of human heart enhancers. Genome research 20: 381-392.

[17] Nord A.S. et al. Cell 155, 1521-1531 (2013) PubMed

[18] Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13: 613–26. pmid:22868264

[19] Shaffer AL et al. A library of gene expression signatures to illuminate normal and pathological lymphoid biology.Immunol Rev. 2006 Apr;210:67-85.

[20] Tychele N. Turner, Bradley P. Coe, Diane E. Dickel, Kendra Hoekzema, Bradley J. Nelson, Michael C. Zody, Zev N. Kronenberg, Fereydoun Hormozdiari, Archana Raja, Len A. Pennacchio, Robert B. Darnell, and Evan E. Eichler; Genomic Patterns of De Novo Mutation in Simplex Autism, Cell 171, 710-722, October 19 2017. http://dx.doi.org/10.1016/j.cell.2017.08.047

[21] UCSC Genome Browser: https://genome.ucsc.edu/

[22] Visel A, Minovitsky S, DubchakI, Pennacchio LA. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35: D88–92. pmid:17130149.

[23] VISTA Enhancer Browser: https://enhancer.lbl.gov/

Bibliography

Davis, Jesse & Goadrich, Mark. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning, ACM. 06. 10.1145/1143844.1143874.

Habibi, Reza & Habibi, Hamed. (2012). Applications of Bootstrap method in Neural networks. Neural, Parallel & Scientific Computations. 20.

Introduction to Semi-Supervised Learning.
www.uio.no/studier/emner/matnat/ifi/INF5300/v16/undevisningsmateriell/chapelle_et_al_2006_bookchapter_introductiontosemi-supervisedlearning.pdf.

Refaeilzadeh, Payam & Tang, Lei & Liu, Huan. (2009). Cross-Validation. Encyclopedia of Database Systems. 532–538. 532-538. 10.1007/978-0-387-39940-9_565.

Yanjun Qi, Oznur Tastan, Jaime G. Carbonell, Judith Klein-Seetharaman, Jason Weston; Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins, Bioinformatics, Volume 26, Issue 18, 15 September 2010, Pages i645–i652, https://doi.org/10.1093/bioinformatics/btq394

FIGURES:

Figure 1: Example RPKM Calculation for the Chromatin Feature over the Enhancer Region

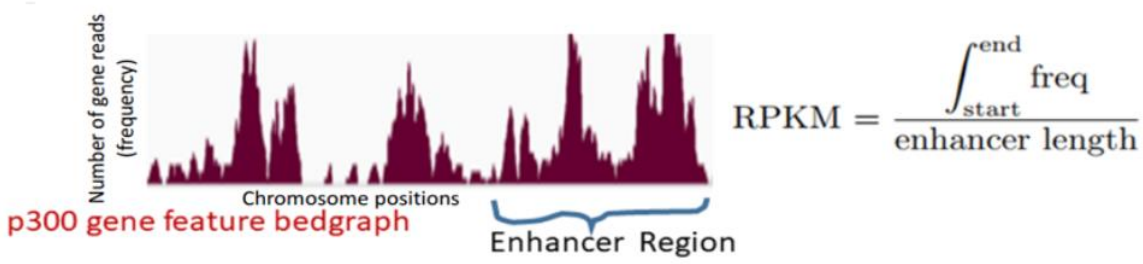

Figure 2: Expression Levels (RPKM) of Chromatin Signatures over Enhancer Regions

Enhancer Regions	RPKM Values for Chromatin Features				Label
	p300	H3K27ac	...	DNase 1	
chr9: (502160,503431)	0.342	0.496	...	...	Negative
chr12: (349028,349870)	0.489	0.238	...	...	Positive
...	1.345	12.123	...	...	Negative

Cleaned data illustrating expression levels (RPKM) of chromatin features for various enhancer regions

Figure 3: Methods to Counter Data Imbalance

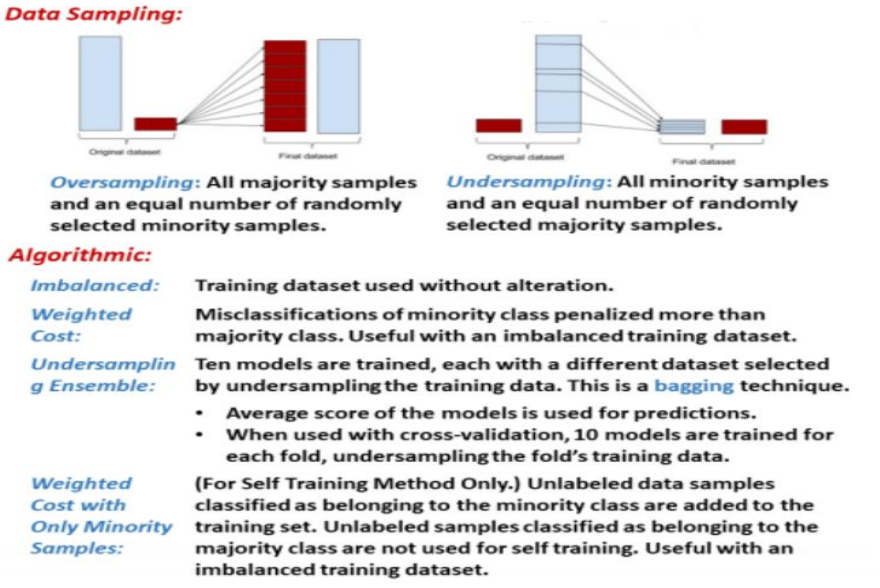

Figure 4: Supervised Learning Approach

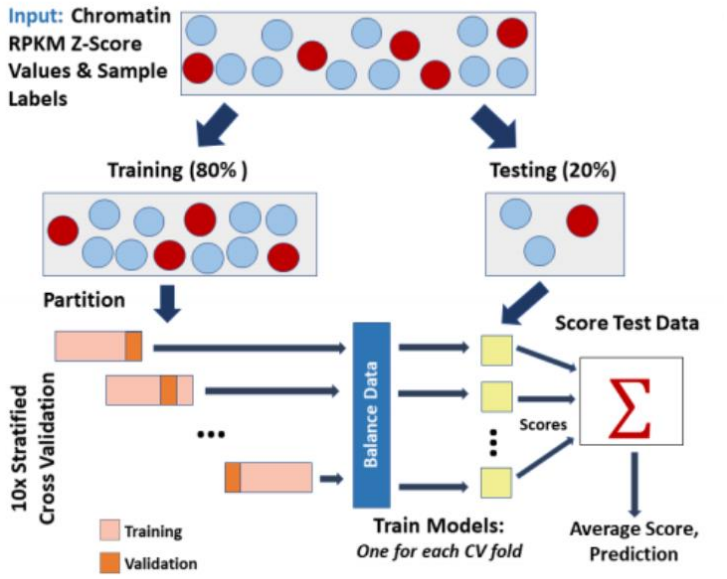

Figure 5: AUROC and AUPRC Trends for all the Classifiers and Balancing Techniques

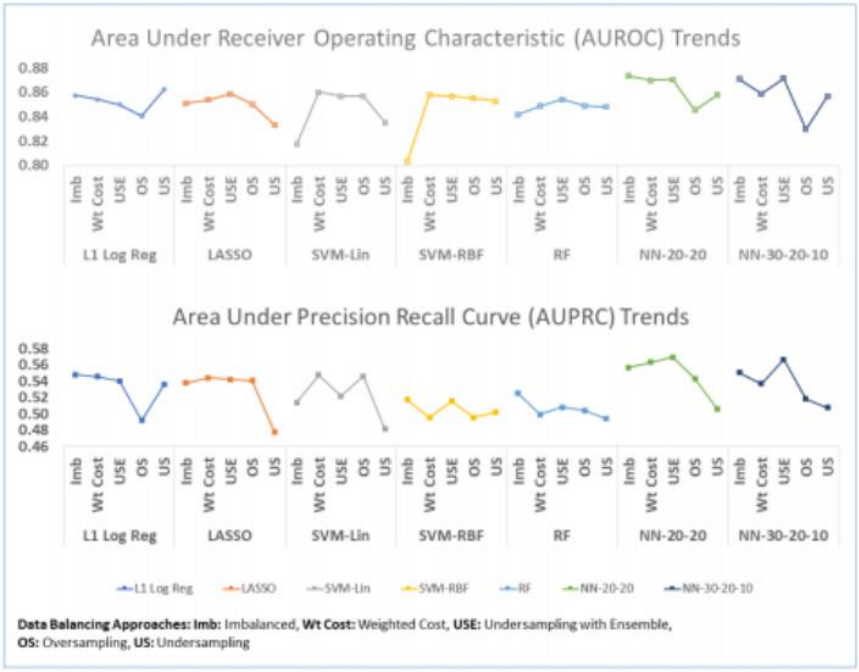

Figure 6: Summary Table for Most Performant Supervised Learning Models

Classifier	Data Method	Sensitivity	Accuracy	AUROC	AUPRC
L1 Logistic Regression	Imbalanced	0.3404	0.9166	0.8575	0.5483
	Weighted Cost	0.7277	0.8412	0.8512	0.5461
LASSO (L2 Logistic Regression)	Imbalanced	0.3362	0.9147	0.8509	0.5380
	Weighted Cost	0.7197	0.8372	0.8539	0.5442
SVM Linear Kernel	Imbalanced	0.2979	0.9152	0.8174	0.5141
	Weighted Cost	0.7319	0.8472	0.8603	0.5477
Neural Net, 2 layers: 20 features in each layer	Imbalanced	0.3489	0.9138	0.8735	0.5572
	Weighted Cost	0.7787	0.8390	0.8701	0.5634
	US Ensemble	**0.7957**	**0.8286**	**0.8707**	**0.5698**
Neural Net, 3 layers: 30, 20 10 features	Imbalanced	0.2851	0.9147	0.8709	0.5507
	US Ensemble	0.7872	0.8281	0.8716	0.5671
Random Forest (1000 trees)	Imbalanced	0.2766	0.9102	0.8419	0.5255
	US Ensemble	0.7787	0.8036	0.8541	0.5080
SVM Radial Kernel	Imbalanced	0.2426	0.9079	0.8028	0.5173
	US Ensemble	0.7702	0.8345	0.8570	0.5154

Average over 5 Test Runs

Figure 7: ROC and PRC for the Supervised Learning Neural Net Ensemble

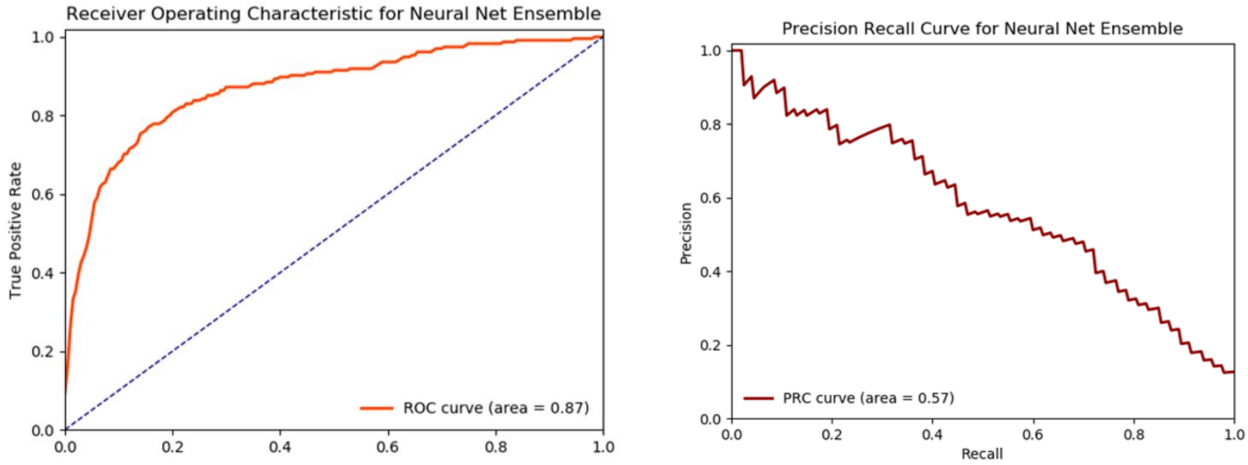

Figure 8: Semi-Supervised (Self-Training) Approach

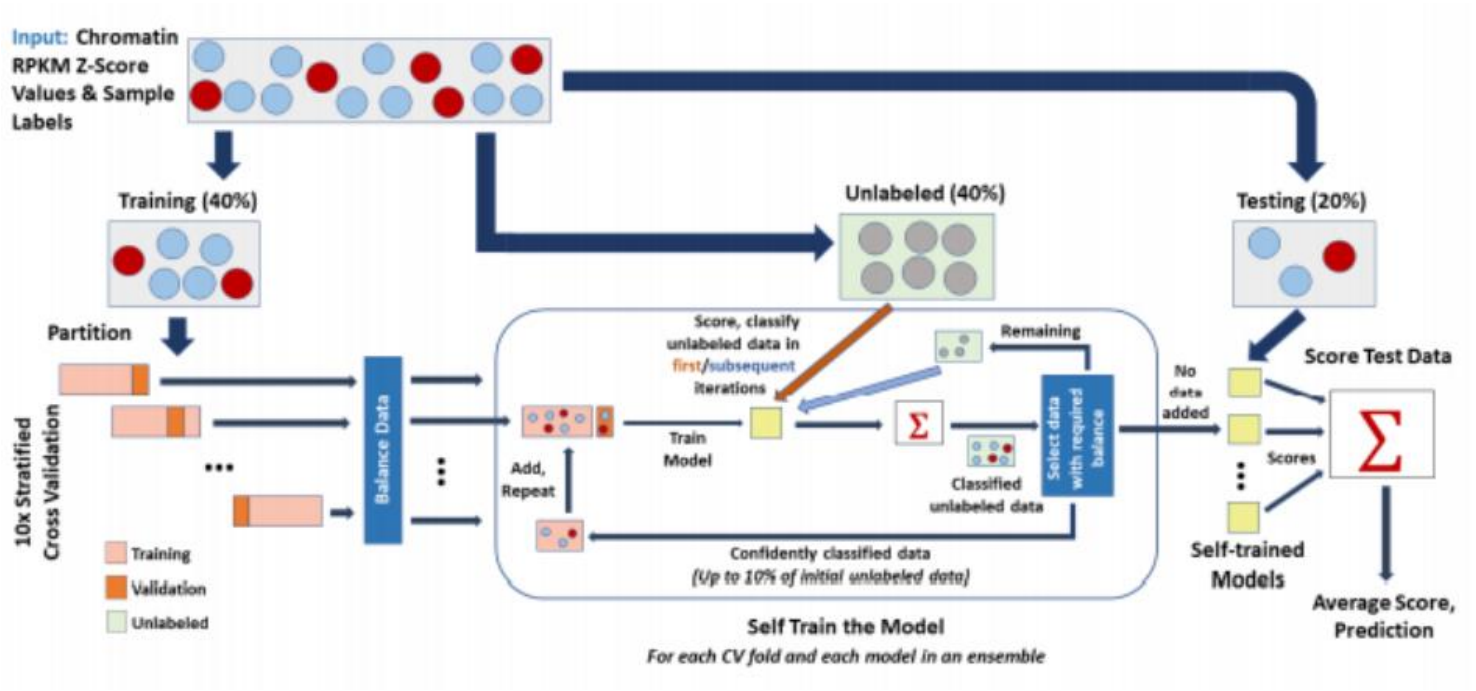

Figure 9: Summary Table for Most Performant Semi-Supervised Learning Models

Self Training Data Method	Samples Added	Sensitivity	Accuracy	AUROC	AUPRC	AUPRC Delta
Add positives only	79.56	0.6596	0.8635	0.8368	0.5216	0.0195
Add positives only	79.74	0.6468	0.8599	0.8288	0.5144	0.0179
Add positives only	80.1	0.6255	0.8717	0.8304	0.5092	0.0184
Add positives only	80.82	0.7021	0.8653	**0.8682**	**0.5640**	0.0057
US Ensemble	162	0.7532	0.8122	0.8518	0.5109	**0.0180**

Figure 10: Influence of Chromatin Features on the Neural Network Ensemble's Predictions

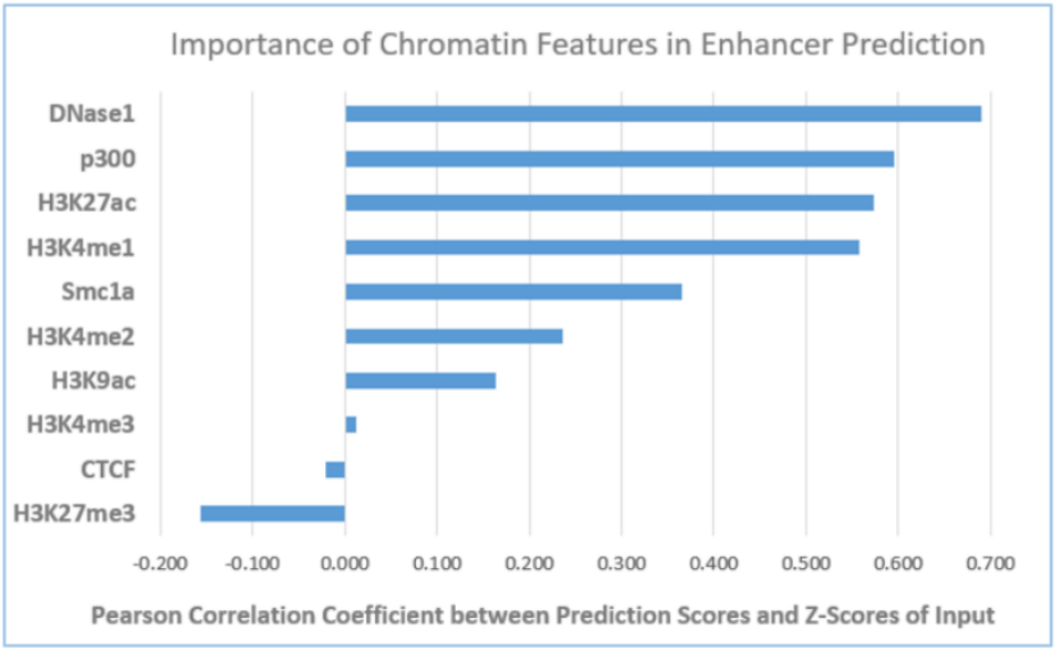

Figure 11: Variation of Principal Chromatins Across All Enhancers (Neural Network Ensemble)

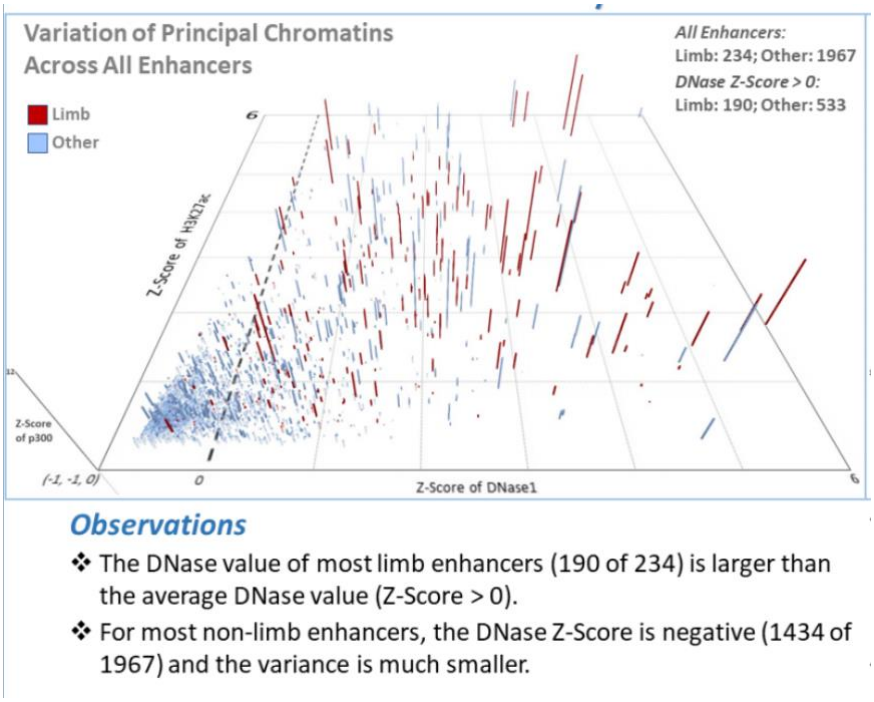

Figure 12: Misclassifications of the Neural Network Ensemble Model

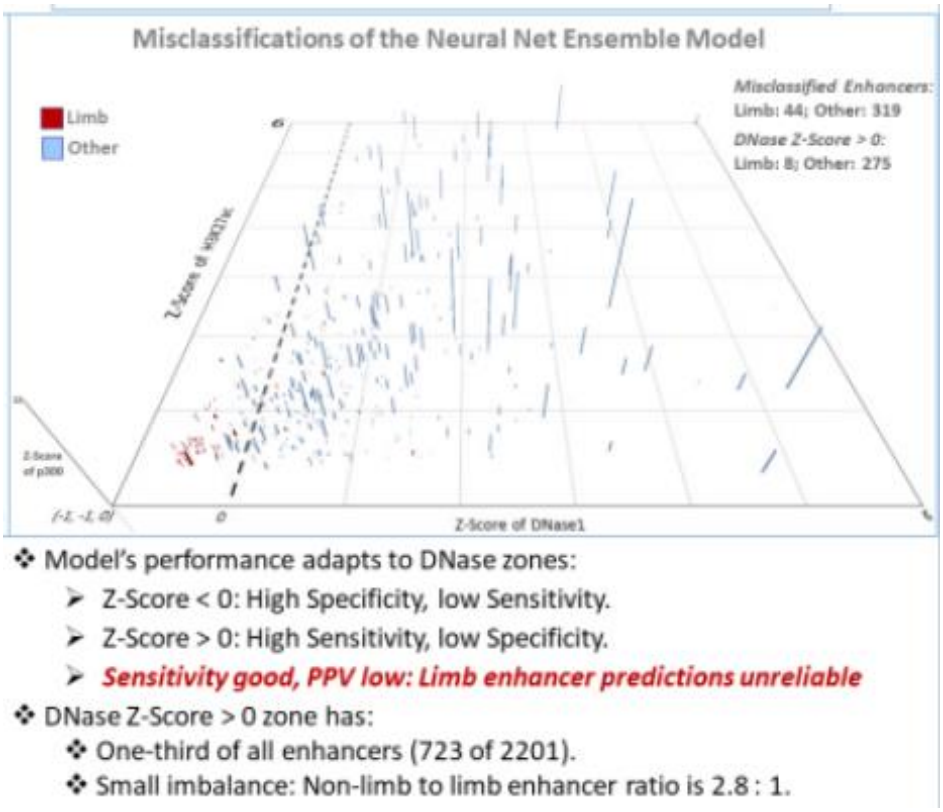

Figure 13: Method for Architecting Zoned Ensemble

Architecting the Zoned Model

Evaluate the performance of the base model M_0 to select K input zones based on features of the data samples:

- A data sample's features alone must determine its zone.
- Data samples may be in 1 or no zone. (Non-overlapping zones.)
- The number of data samples in each zone and the population size of classes in the zone should not be too small. (Mitigates overfitting.)

Divide dataset into training set R and test set S, preserving the data set's class imbalance ratio.

- Train model M_i for zone i with data samples from R in zone i with any classifier. Evaluate M_i with *all* data samples from S in zone i.
  - Select the best classifier for each zone.
  - ***Different zones can have different classifiers.***
- The zoned model is the collection $< M_0, M_1, \ldots, M_K >$:
  - Model M_i is used to label data samples in zone i.
  - M_0 is used to label data samples that are not in any input zone.
- Repeat with different training and test datasets. Average the performance results for these runs.

Figure 14: Summary of Zoned Model Ensemble Results

Summary Table of Results

Classifier	Zone	Sensitivity	PPV	Accuracy	AUROC	AUPRC
Neural Net Ensemble	All Zones	0.7957	0.3621	0.8286	0.8707	0.5698
	DNase > 0	0.9428	0.4002	0.6045	0.8338	0.6652
Zoned Neural Net Ensemble	All Zones	**0.6436**	**0.4595**	**0.8821**	**0.8681**	**0.5878**
	DNase > 0	**0.7559**	**0.5180**	**0.7434**	**0.8315**	**0.6886**

Average over 5 Test Runs

Figure 15: Misclassifications of the Zoned Neural Network Ensemble Model

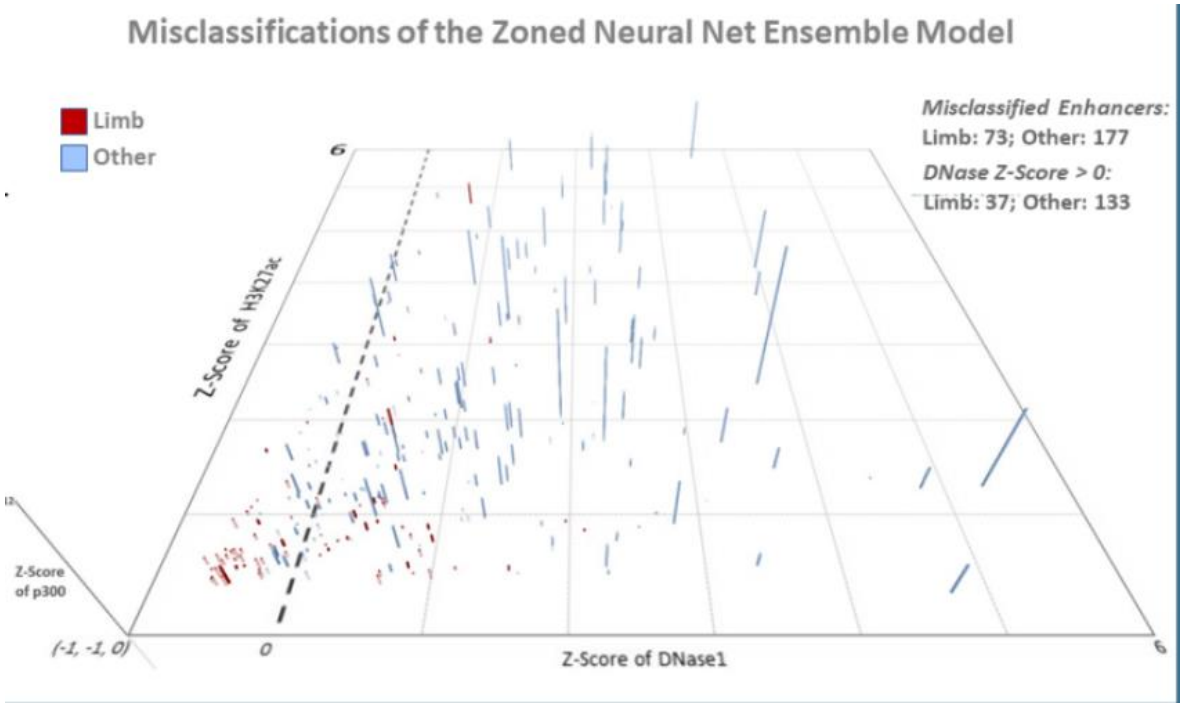

Figure 16: New Candidate Enhancer Predictions in Chromosome 9

Predictions of new Limb Enhancer Segments by my Discovery Algorithm in Chromosome 9

Examined chr9 genome from 13600000 bp to 121500000 bp; known enhancers are from 13697970 to 121357466.

1. Pick Segments that have DNase, p300, H3K27ac

Columns A - G: output of OneKSegmentsChr9.py to file Chr9CandidateSegments.tsv

Segments without DNaseScore were added to extend contiguous segment groups by 500 bp before and after the group

2. Calculate RPKM values for feature inputs; score segments

3. Analyze scores, merge segments

Scores were calculated by predict16dComboModel.py and output to file Chr9CandidateSegmentScores.tsv

119 of the discovered segments have score >= 0.8

10 of the merged segments are shown below

Chrc	Left	Right	Segment	Index	DNaseSco
chr9	25525500	25526500	Seg938	938	
chr9	25526000	25527000	Seg939	939	11118
chr9	25526500	25527500	Seg940	940	29620
chr9	25527000	25528000	Seg941	941	30611
chr9	25527500	25528500	Seg942	942	11728
chr9	29317500	29318500	Seg1138	1138	10299
chr9	29318000	29319000	Seg1139	1139	11275
chr9	29318500	29319500	Seg1140	1140	12977
chr9	29319000	29320000	Seg1141	1141	12397
chr9	29319500	29320500	Seg1142	1142	
chr9	30741500	30742500	Seg1308	1308	
chr9	30742000	30743000	Seg1309	1309	17745
chr9	30742500	30743500	Seg1310	1310	20468
chr9	30743000	30744000	Seg1311	1311	21191
chr9	30743500	30744500	Seg1312	1312	16630
chr9	30744000	30745000	Seg1313	1313	
chr9	32719000	32720000	Seg1607	1607	
chr9	32719500	32720500	Seg1608	1608	22628
chr9	32720000	32721000	Seg1609	1609	43265
chr9	32720500	32721500	Seg1610	1610	38061
chr9	32721000	32722000	Seg1611	1611	18474
chr9	32721500	32722500	Seg1612	1612	
chr9	33103000	33104000	Seg1687	1687	
chr9	33103500	33104500	Seg1688	1688	8709
chr9	33104000	33105000	Seg1689	1689	36963
chr9	33104500	33105500	Seg1690	1690	46150
chr9	33105000	33106000	Seg1691	1691	26913
chr9	33105500	33106500	Seg1692	1692	14343
chr9	33106000	33107000	Seg1693	1693	
chr9	35323500	35324500	Seg1871	1871	11576
chr9	35324000	35325000	Seg1872	1872	26787
chr9	35324500	35325500	Seg1873	1873	29365
chr9	35325000	35326000	Seg1874	1874	13191
chr9	35326000	35327000	Seg1875	1875	27459
chr9	35326500	35327500	Seg1876	1876	47713
chr9	35327000	35328000	Seg1877	1877	28319
chr9	35327500	35328500	Seg1878	1878	
chr9	37029000	37030000	Seg2032	2032	
chr9	37029500	37030500	Seg2033	2033	27657
chr9	37030000	37031000	Seg2034	2034	44233
chr9	37030500	37031500	Seg2035	2035	26538

Index	Segme	Scor	Grou		Order of seg same?
938	Seg938	0.64652	1		TRUE
939	Seg939	**0.836**	1	Merge	TRUE
940	Seg940	**0.938**	1		TRUE
941	Seg941	**0.882**	1		TRUE
942	Seg942	0.7421	1		TRUE
1138	Seg1138	0.29876	2?		TRUE
1139	Seg1139	**0.912**	2	Merge	TRUE
1140	Seg1140	**0.938**	2		TRUE
1141	Seg1141	0.62856	2		TRUE
1142	Seg1142	0.26771	2-after		TRUE
1308	Seg1308	0.48755	3-before		TRUE
1309	Seg1309	0.61513	3		TRUE
1310	Seg1310	0.73117	3		TRUE
1311	Seg1311	**0.815**	3	Merge	TRUE
1312	Seg1312	**0.845**	3		TRUE
1313	Seg1313	0.57951	3-after		TRUE
1607	Seg1607	0.47273	4-before		TRUE
1608	Seg1608	**0.817**	4	Merge	TRUE
1609	Seg1609	**0.935**	4		TRUE
1610	Seg1610	**0.883**	4		TRUE
1611	Seg1611	**0.809**	4		TRUE
1612	Seg1612	0.643	4-after		TRUE
1687	Seg1687	0.50158	5-before		TRUE
1688	Seg1688	0.62794	5		TRUE
1689	Seg1689	0.79675	5		TRUE
1690	Seg1690	**0.852**	5	Merge	TRUE
1691	Seg1691	**0.842**	5		TRUE
1692	Seg1692	0.75154	5		TRUE
1693	Seg1693	0.50537	5-after		TRUE
1871	Seg1871	0.24617	6-before		TRUE
1872	Seg1872	0.70247	6?		TRUE
1873	Seg1873	0.63373	6?		TRUE
1874	Seg1874	0.26858	6?		TRUE
1875	Seg1875	**0.862**	6	Merge	TRUE
1876	Seg1876	**0.949**	6		TRUE
1877	Seg1877	**0.913**	6		TRUE
1878	Seg1878	0.76749	6-after		TRUE
2032	Seg2032	0.38167	7-before		TRUE
2033	Seg2033	**0.838**	7	Merge	TRUE
2034	Seg2034	**0.908**	7		TRUE
2035	Seg2035	0.77204	7		TRUE

Figure 17: Example Limb Enhancer found on Chromosome 9 by the Model. *Note the high expression of p300, H3K27ac, DNase1 illustrating how the prediction is biologically relevant*