

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

Zhongxia Yan

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Regulation Saturation Prediction

Zhongxia Yan and Anat Kreimer

Nir Yosef's lab

UC Berkeley

Previous Work

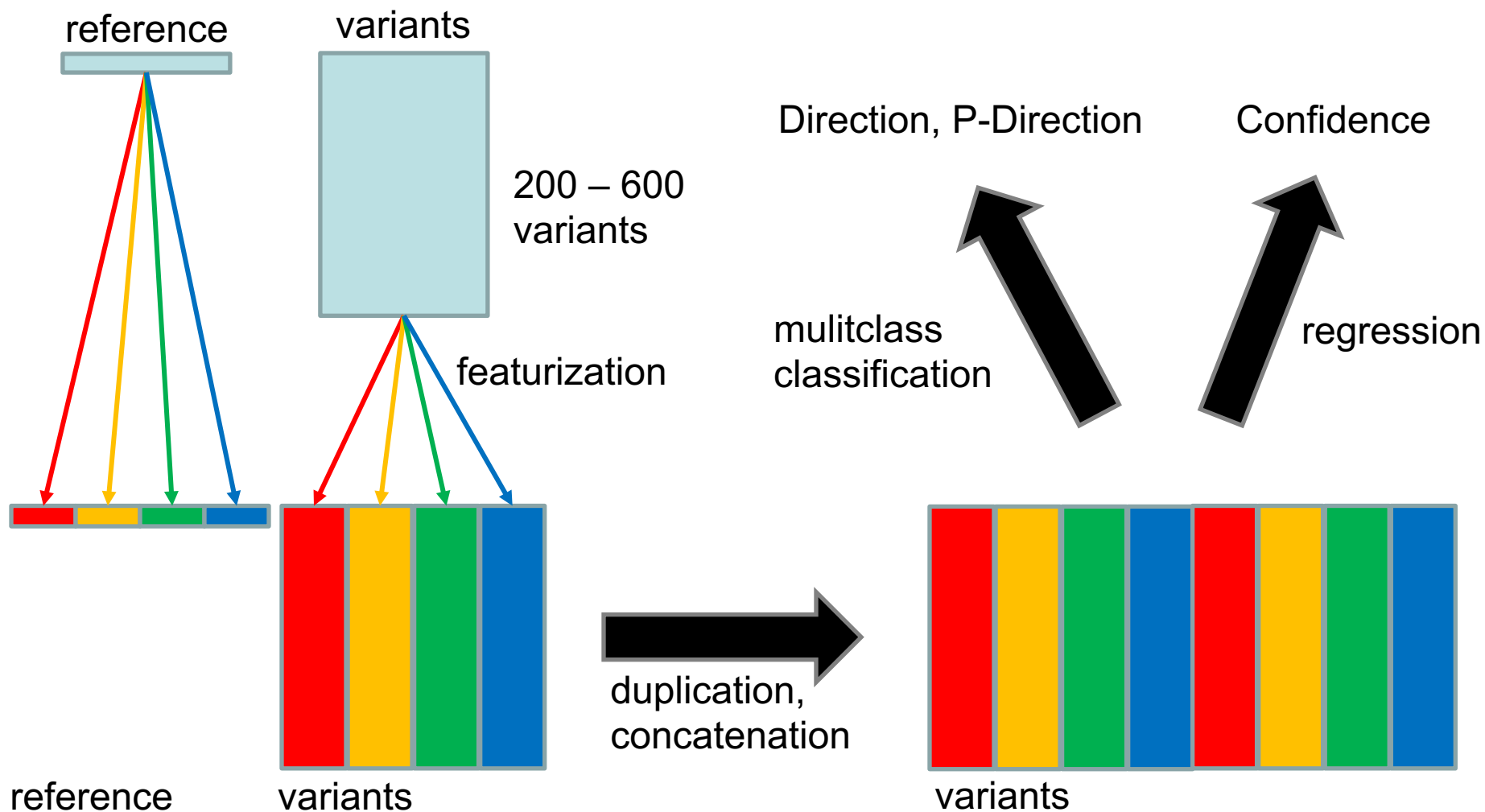
Prediction of functional regulatory elements

- Goal: explore features and models that may predict regulatory activity of candidate enhancer regions
- Datasets: mostly publicly available datasets of MPRA activity
- Features: aggregate of experimental features, outputs of other models, and sequence-based features
- Models: ensemble of nonlinear models (e.g. RandomForest)

Insights from Previous Work

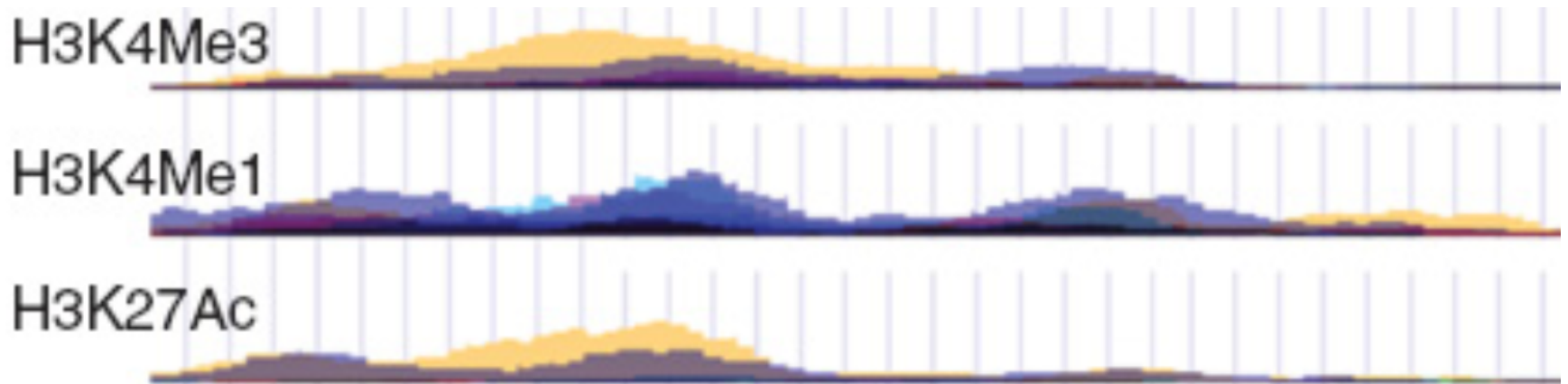
- Ensemble models produce better accuracy
- Concatenating all features produce better results than using individual features
- Supervised models learn well on MPRA datasets
- Models trained on one dataset can be applied to predict on another with moderate loss of accuracy

Overview



Featurization per Region

- Experimental
 - Number of base-pair overlap with measured peaks of ENCODE epigenetic factors
 - 1095 features

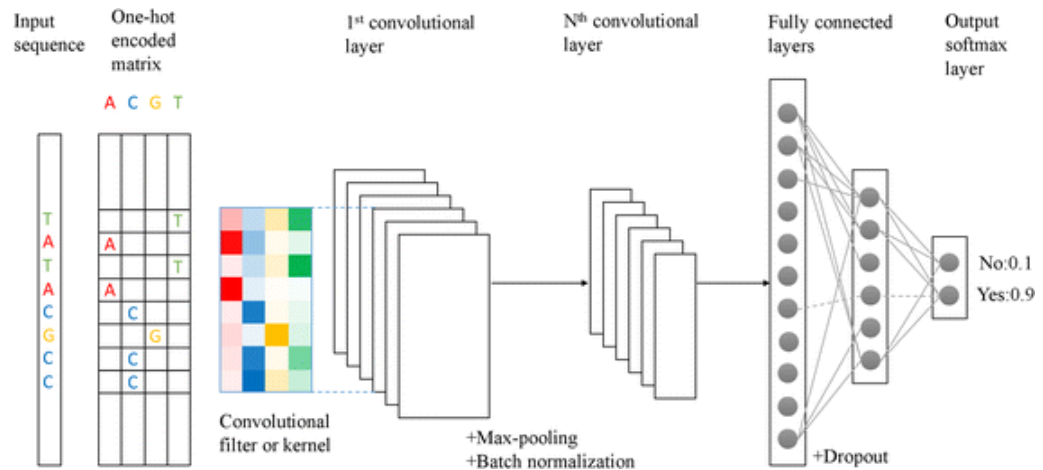


Featurization per Region Cont.

- DeepSea
 - Quantifications of transcription factor binding, DNase I sensitivities, and histone marks
 - 919 features

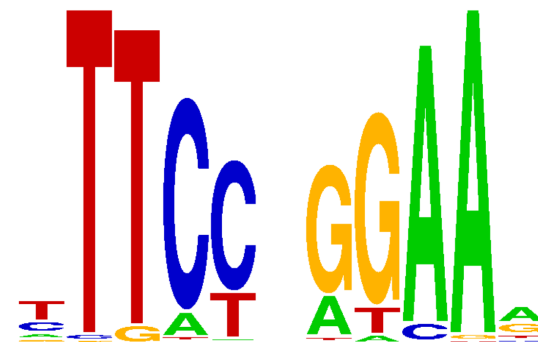
Both methods utilize convolutional neural networks

- DeepBind
 - Quantifications of transcription factor binding
 - 515 features



Featurization per Region Cont.

- Motifs
 - Scoring of ENCODE sequence motifs
 - 2065 features
- 5mer
 - Counts for each possible 5mer
 - 1024 features



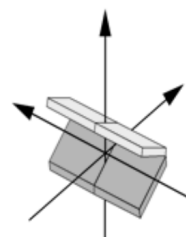
Featurization per Region Cont.

- DNA Shape

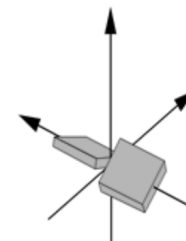
- Predicted DNA shape parameters: Minor groove width, Roll, Propeller twist, Helix Twist
- 4 features



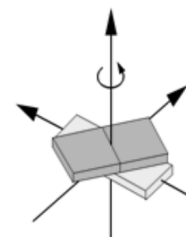
MGW



Roll



ProT



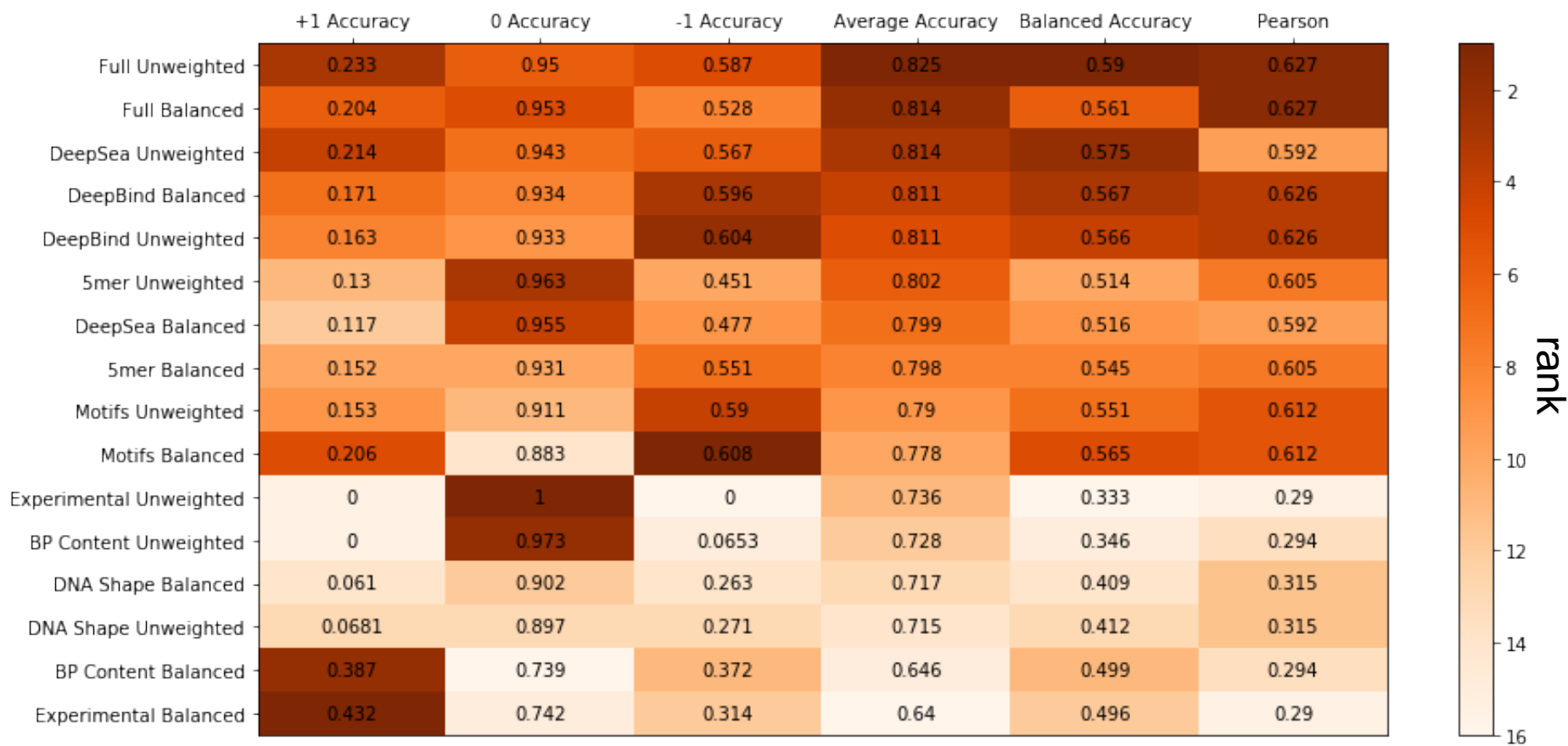
HelT

- BP Content

- Maximum length of polyA and polyT. GC content
- 3 features

Supervised Ensemble Model

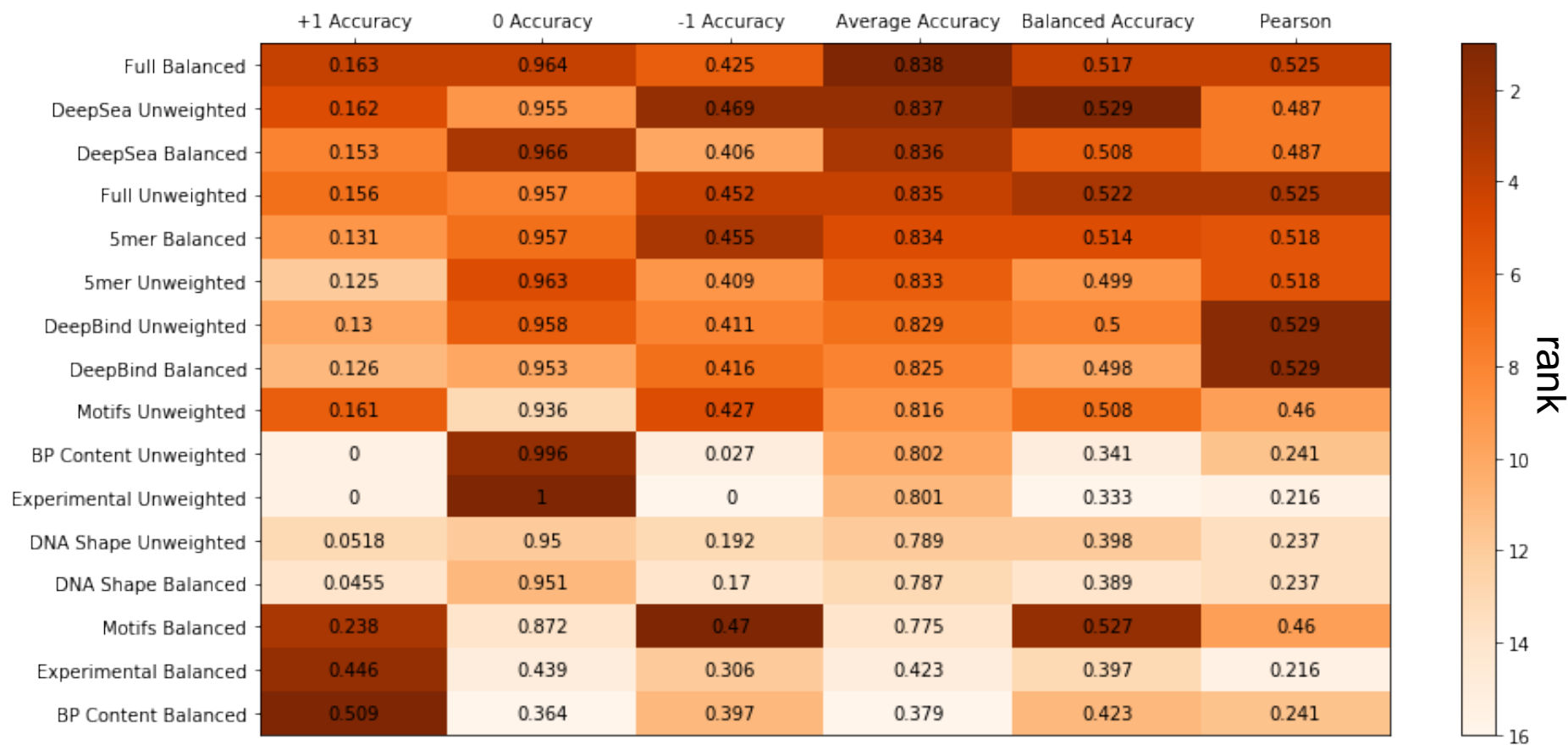
- Separate models for promoters and enhancers
- Ensemble of RandomForest and ExtraTrees classifiers for predicting direction
- RandomForest regressor for predicting the confidence
 - Rescale the predicted confidence to the distribution of the training confidences
- One attempt balances the weight of the classes by the inverse frequencies



Cross validation values by feature sets for **Enhancers**, sorted by Average accuracy, colored by rank.

Accuracies correspond to the Direction classification

Pearson corresponds to the Confidence regression



Cross validation values by feature sets for **Promoters**, sorted by Average accuracy, colored by rank

Accuracies correspond to the Direction classification

Pearson corresponds to the Confidence regression

Supervised Ensemble Model Analysis

- Predicts class 0 well, but does worse on +1 and -1
- Full set of features performs the best overall in both Enhancers and Promoters
- Balancing class weights by class frequencies does not help prediction

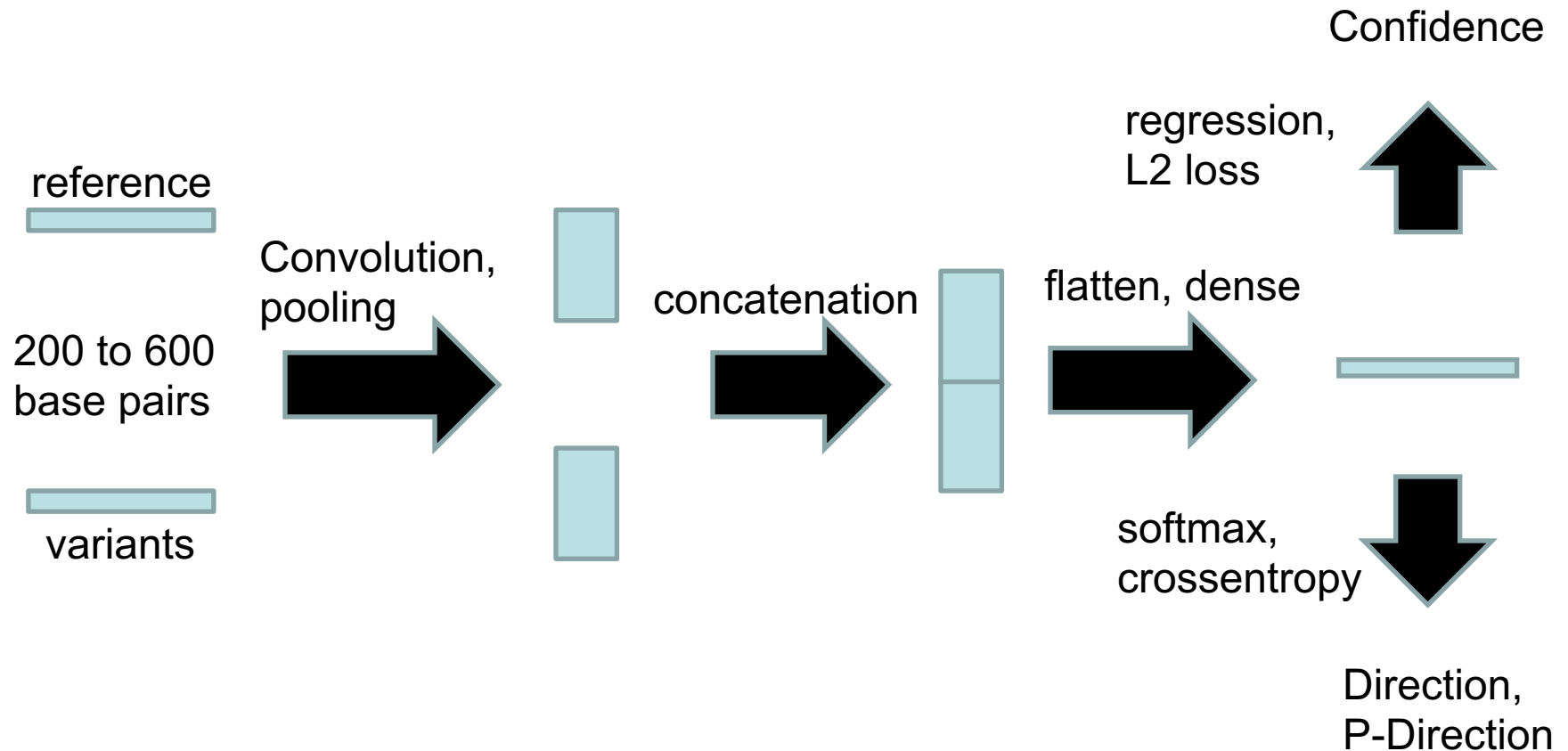
Transfer Ensemble Model

- Utilize other MPRA datasets
 - Putative regulatory regions (Kwasnieski et al. 2014)
 - Expression quantitative trait loci (eQTL) regions (Tewhey et al. 2016), also used in CAGI 4
 - Putative liver enhancers (Inoue et al. 2017)
 - Putative enhancers (Inoue et al. unpublished)
- Trained a regressor on featurization to predict regulatory activity
- Applied regressor on each pair of reference and variant data, then subtract to get log fold change – predict direction from log fold change
- Did not use training labels

Transfer Ensemble Model Analysis

- Performance is bad:
 - +1 Accuracy = 0.26
 - 0 Accuracy = 0.74
 - -1 Accuracy = 0.21
- Cannot predict “confidence”
- Maybe can use previous datasets in conjunction with training set

Siamese CNN Model



Siamese CNN Model Analysis

- Never desirably converged during training
 - Always predicts class 0
- Maybe due to not enough data
 - Only around 4000 total variants
 - Variants of the same region are very similar to one another

Submission Labels

1. Prediction from balanced Supervised model
2. Prediction from balanced Supervised model – duplicate of previous
3. Prediction from unbalanced Supervised model
4. Prediction from balanced Supervised model with enhancer directions from Transfer model

Conclusion

- Ensemble model is effective in distinguishing single nucleotide variants
- Full featurization is the most effective, and works on both promoters and enhancers
- Neural network may need more tuning or more data

Acknowledgments

Yosef Lab:

- Anat Kreimer
- Nir Yosef

Various labs for borrowed models and data

CAGI organizers!

Overview of Challenge

- Predict effect of single nucleotide variation in enhancers and promoters on MPRA activity
- Dataset consist of 9 promoter regions and 5 enhancer regions from 187 bp to 600 bp
- Each region has around $3 * \text{<num bp>}$ single nucleotide variants, and around a quarter of those are labeled

Data Format

- Training Data

#Chrom	Pos	Ref	Alt	Value	Confidence
X	138612669	T	A	-0.17	0.07

- Value is variant effect

- Test Output Fortmat

Chrom	Pos	Ref	Alt	Promoter_Enhancer	Direction	P_Direction	Confidence	SE	Comments
1	109817273	C	A	SORT1	*	*	*	*	*

- Direction:

- Derived as $\text{sign}(\text{value}) * (\text{confidence} > 0.1)$

- P-direction: probability that direction is correct

Data Preprocessing

- For each region, featurize the reference region and all of its variants
- Full feature for each variant is the concatenation of the reference and variant features