Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

Miriam Shiffman
2/28/17

Genome-wide prediction of cis*-regulatory regions using supervised deep learning methods*

(Li, Shi, Wasserman 2016)

Massachusetts
Institute of
Technology

http://mit6874.github.io

Overview

- Key Claim
  - ***DECRES (DEep learning for identifying Cis-Regulatory ElementS) can be used to identify active enhancers and promoters with better sensitivity and specificity than previous models***
- Importance
  - ***Early steps toward demonstrating how (with sufficient data) neural networks can identify putative regulatory elements, with some interpretability***
- Issues
  - ***Overstated claims***
  - ***Failure to compare to randomized background or comparable models***

Assumptions

- Enhancers & promoters = discrete classes
- Enhancers (E) & promoters (P) = always active or inactive in a given cell line
- CAGE (cap analysis of gene expression) to read 5’ transcript “tag” = good proxy for regulatory status
    - E: tags per million >0 or =0
        - Bidirectional eRNAs
    - P: tags per million >5 or =0

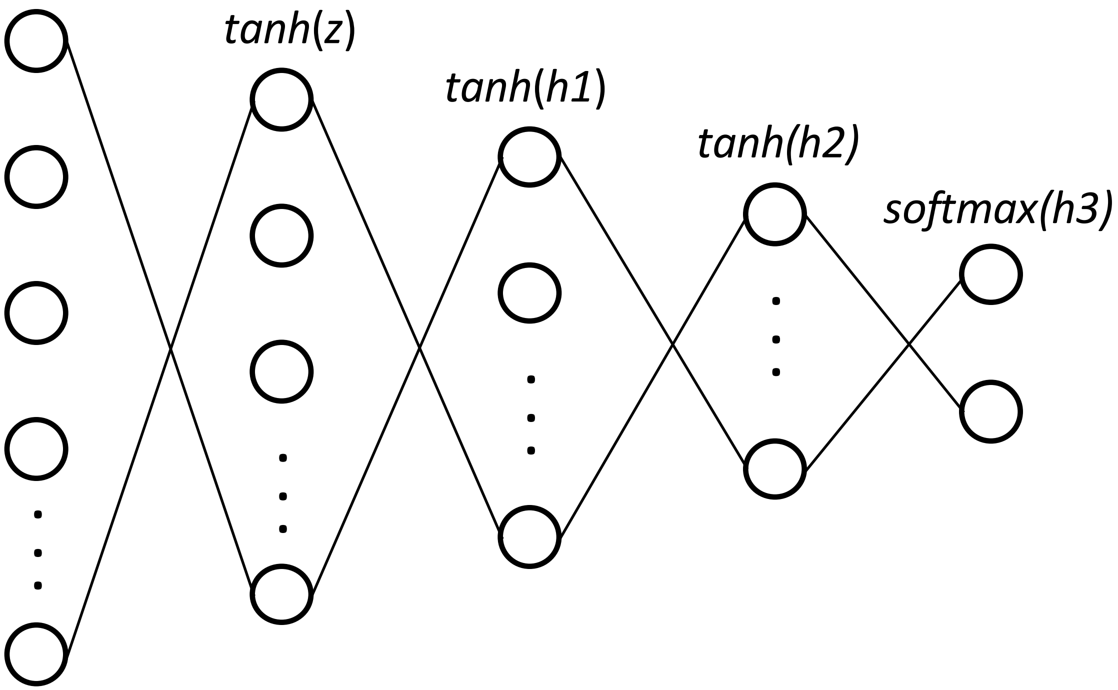
Data

- 8 (really 4) well-characterized cell lines
  - Segmented into 200 bp bins
- Encyclopedia of DNA Elements (**ENCODE**)
  - Cell-line specific
  - Histone modification ChIP-seq, transcription factor binding ChIP-seq, …
- Sequence data
  - Universal
  - CpG islands, …
- Functional Annotation of Mammalian Genomes (**FANTOM**)
  - Atlas of transcriptionally-active promoters & enhancers
  - Based on CAGE

Methods

- Fully-connected neural net classification
  - In: experimental data (ENCODE) or sequence features
  - Out: sequence feature class in cell line (FANTOM)
    - A-E (active enhancer), A-P (active promoter), …
- NLL == softmax cross-entropy loss
- Regularization
  - ℓ_2-regularization
  - Early-stopping (held-out validation, ≤1000 iters)
- Model assessment
  - 10-fold cross-validation
  - Comparison to other models
    - Unsupervised (ChromHMM, Segway)
    - Supervised (dReg SVM)
  - Independent experimental data
  - Functional & motif enrichment

Model #1: Deep learning for *cis*-regulatory region (CRR) classification



input *x*	***h1***	***h2***	***h3***	**class *y***
n=73-136 (ENCODE) *n*=351 (sequence)	*n*=256	*n*=128	*n*=64	*n*=2 : {bg, CRR} or {A-E, A-P} or … *n*=3: {bg, A-E, A-P}

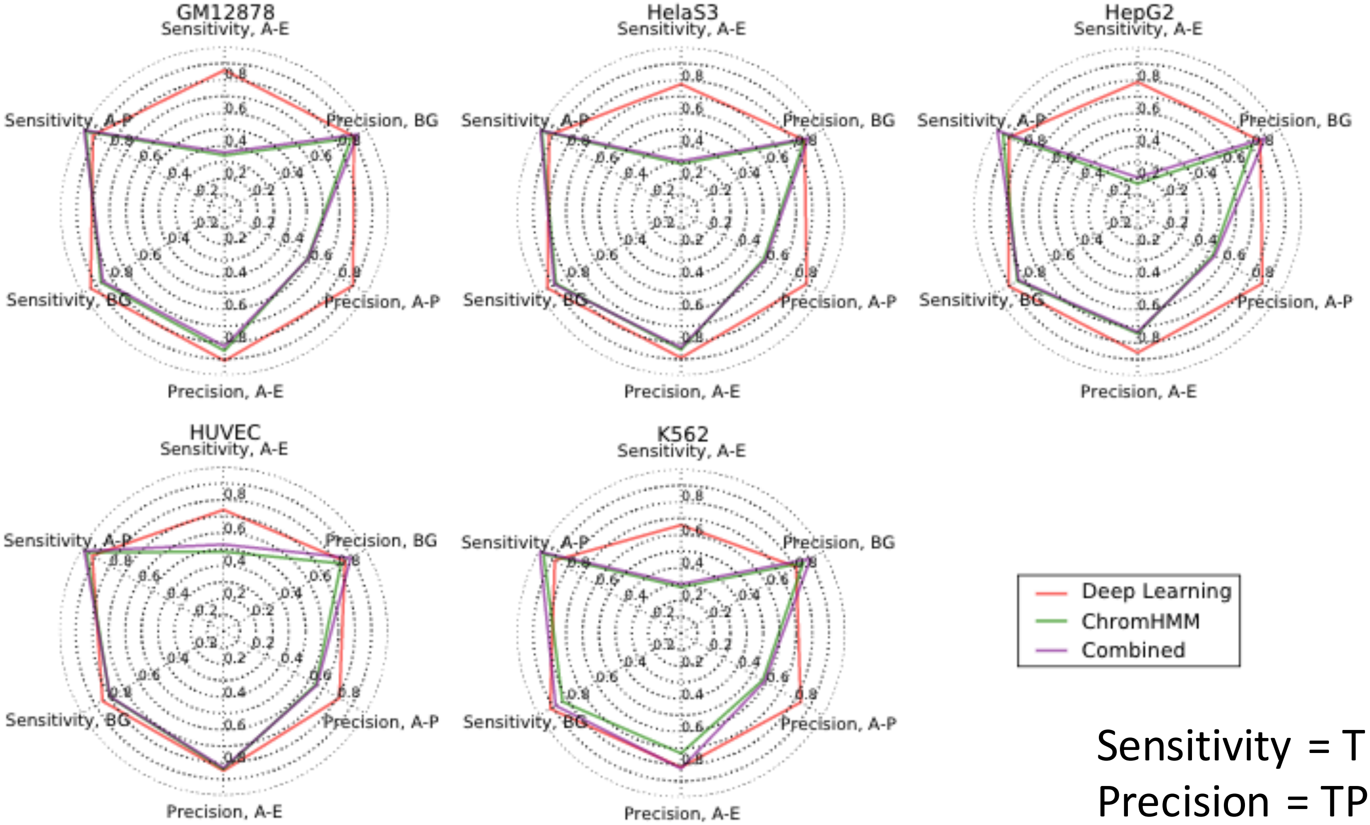
Results: Deep learning for *cis*-regulatory region (CRR) classification

- ***DECRES accurately distinguishes between classes of CRRs, <u>with sufficient data</u> (cell-line specific)***

Classes	**Test accuracy**
CRR type: {A-E, A-P}	87.78 – 93.59%
Activity status: {A-E, I-E} *or* {A-P, I-P}	≈ 90 – 95.87%
CRR across genome: {A-E, A-P, background}	≈ 84 – 90%

- ***Performance correlated with training data size***
- ***Not improved by adding universal sequence features (though predictive alone)***

DECRES ≥ unsupervised methods for detecting active enhancers & promoters.

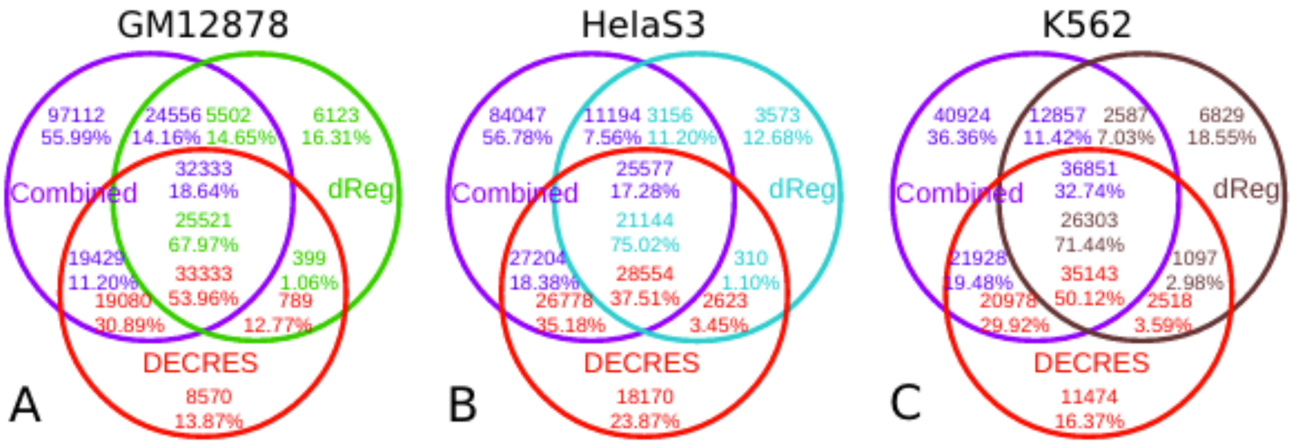


Sensitivity = TP / (TP+FN)
Precision = TP / (TP+FP)

Figure 2: Comparison of the supervised method (Deep Learning) and unsupervised methods (ChromHMM and Combined) on five FANTOM annotated test sets. The ENCODE segmentations were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation. We relabelled the annotations of ChromHMM and Combined. For ChromHMM segmentations, the Tss, TssF, and PromF classes were merged to A-P; the Enh, EnhF, EnhW, EnhWF classes were merged to A-E; and the rest were denoted by BG. When processing the Combined annotations, TSS and PF were relabelled to A-P; E and WE were relabelled to A-E; and the rest to BG.

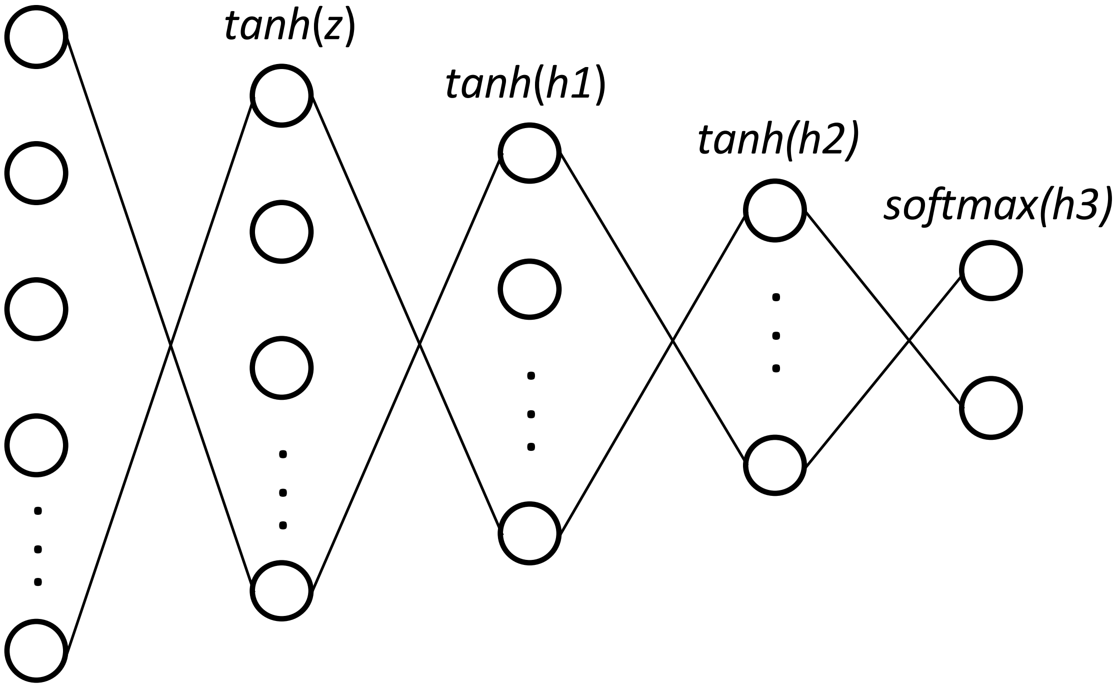
Results: Deep learning for CRR classification

- ***Most predictions agree with other models***



- ***Agreement with independent experiments***
  - CRE-seq: 65.5% sensitivity, 40.2% precision
  - Lower confidence associated with false positives (Mann-Whitney U / Wilcoxon rank sum)
- ***Sensible cell-line specific functional enrichment***
  - e.g. Immune response, B-cell signaling, and leukemia pathways in lymphoblastoid lineage

Model #1: Deep learning for CRR classification



tanh(*z*)

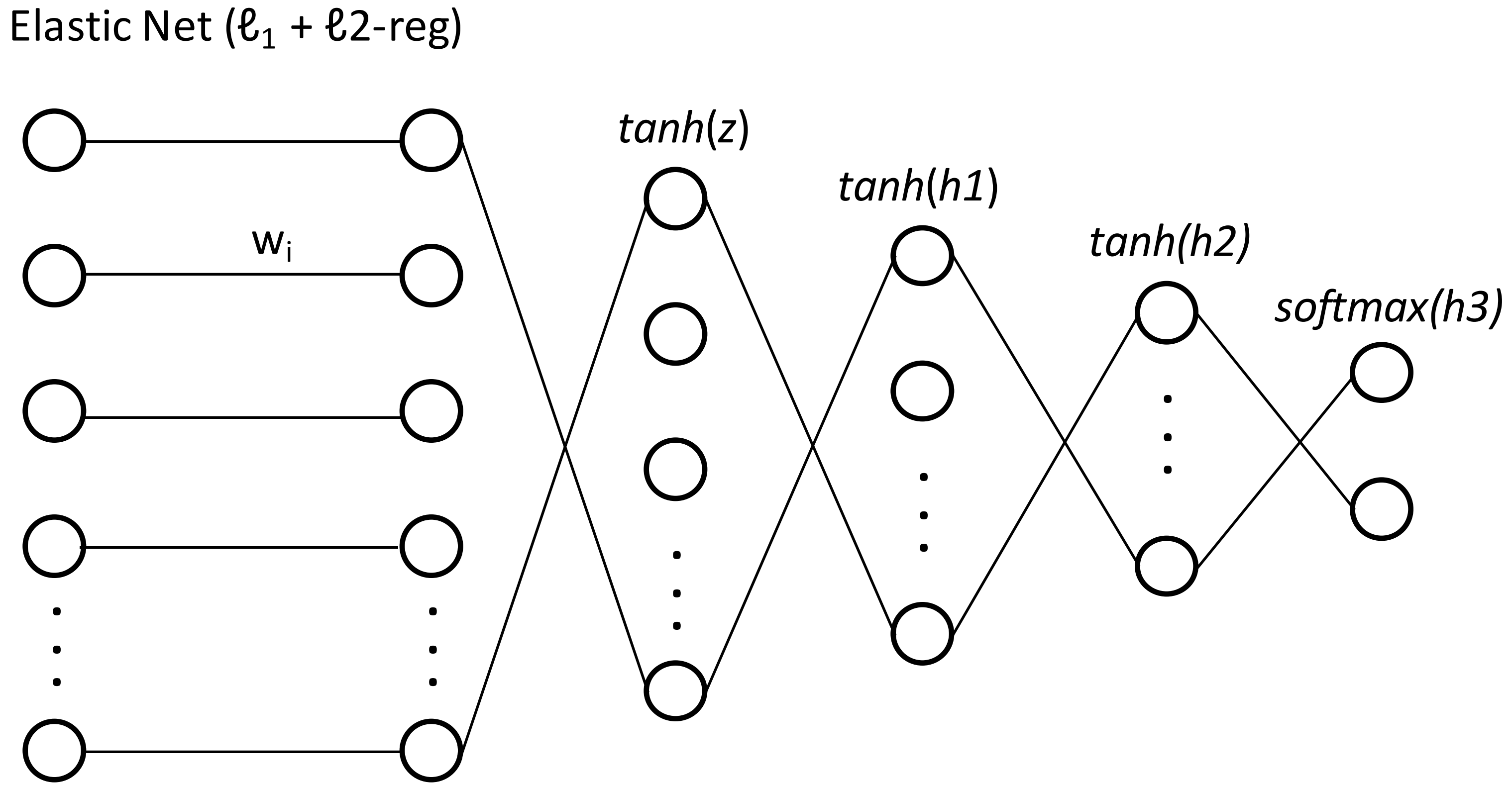
tanh(*h1*)

tanh(h2)

softmax(h3)

input *x*	***h1***	***h2***	***h3***	**class *y***
n=73-136 (ENCODE) *n*=351 (sequence)	*n*=256	*n*=128	*n*=64	*n*=2 : {bg, CRR} or {A-E, A-P} or … *n*=3: {bg, A-E, A-P}

Model #2: Deep feature selection



input *x*
n=73-136 (ENCODE)
n=351 (sequence)

feature selection

h1
n=256

h2
n=128

h3
n=64

class *y*
n=2 : {bg, CRR} or
{A-E, A-P} or …
n=3: {bg, A-E, A-P}

Accuracy increases but plateaus as features increase.

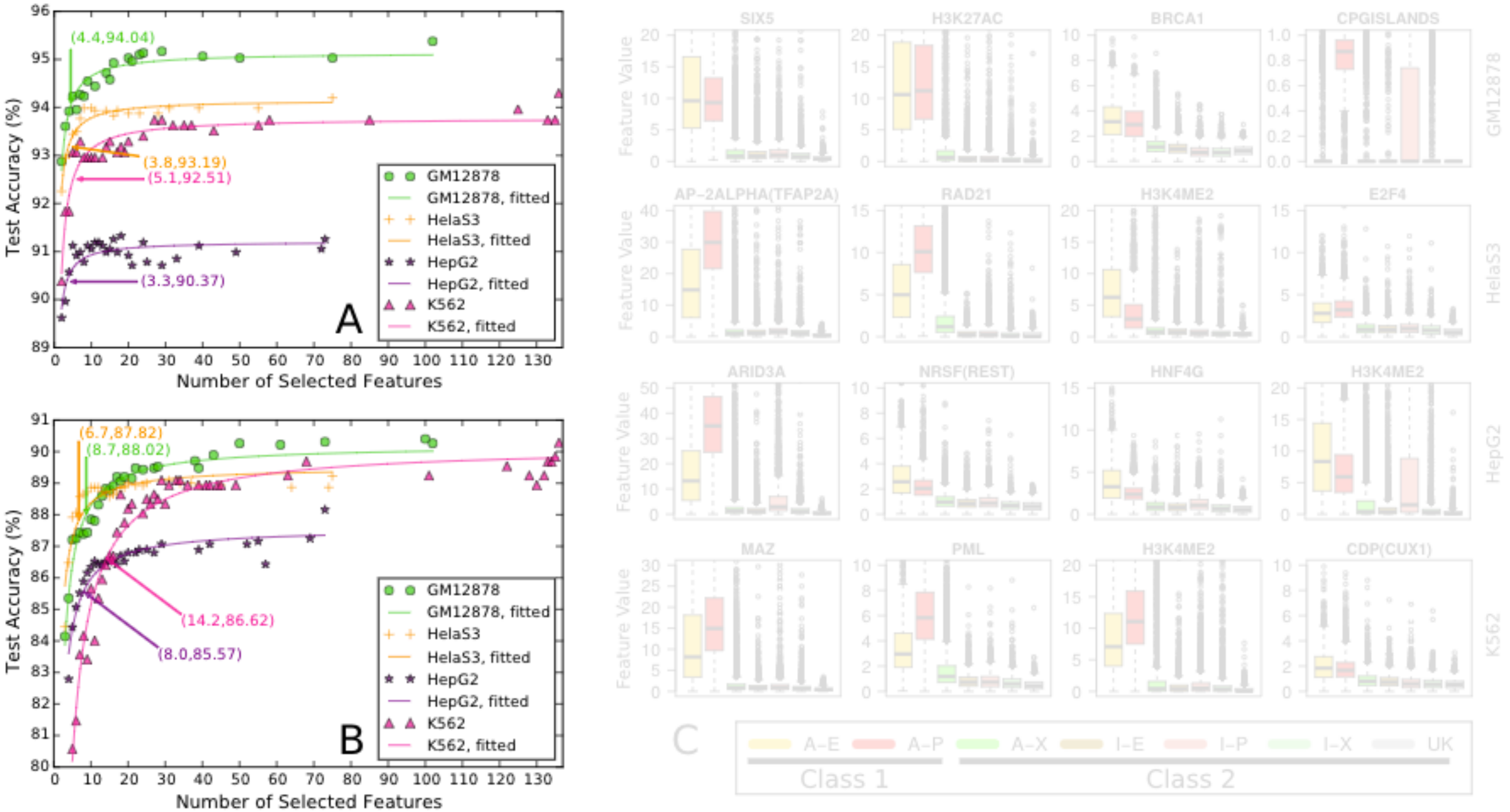


Figure 4: Feature analysis. (A) Accuracy versus the number of features incorporated into the model for 2-class prediction (distinguishing active CRRs (A-E + A-P) from BG (background: A-X, I-E, I-P, I-X and UK). The annotated points indicate where a line with slope 0.25 intersects a fitted curve). (B) Accuracy versus the number of features for a 3-class prediction (distinguishing A-E, A-P and BG). Points as described for (A). (C) For the top 4 features of the 2-class models generated for four well-characterized cell lines, box-plots depict the range of observed feature values (log2 scale) for 7 sequence classes.

Minimal feature set is informative.

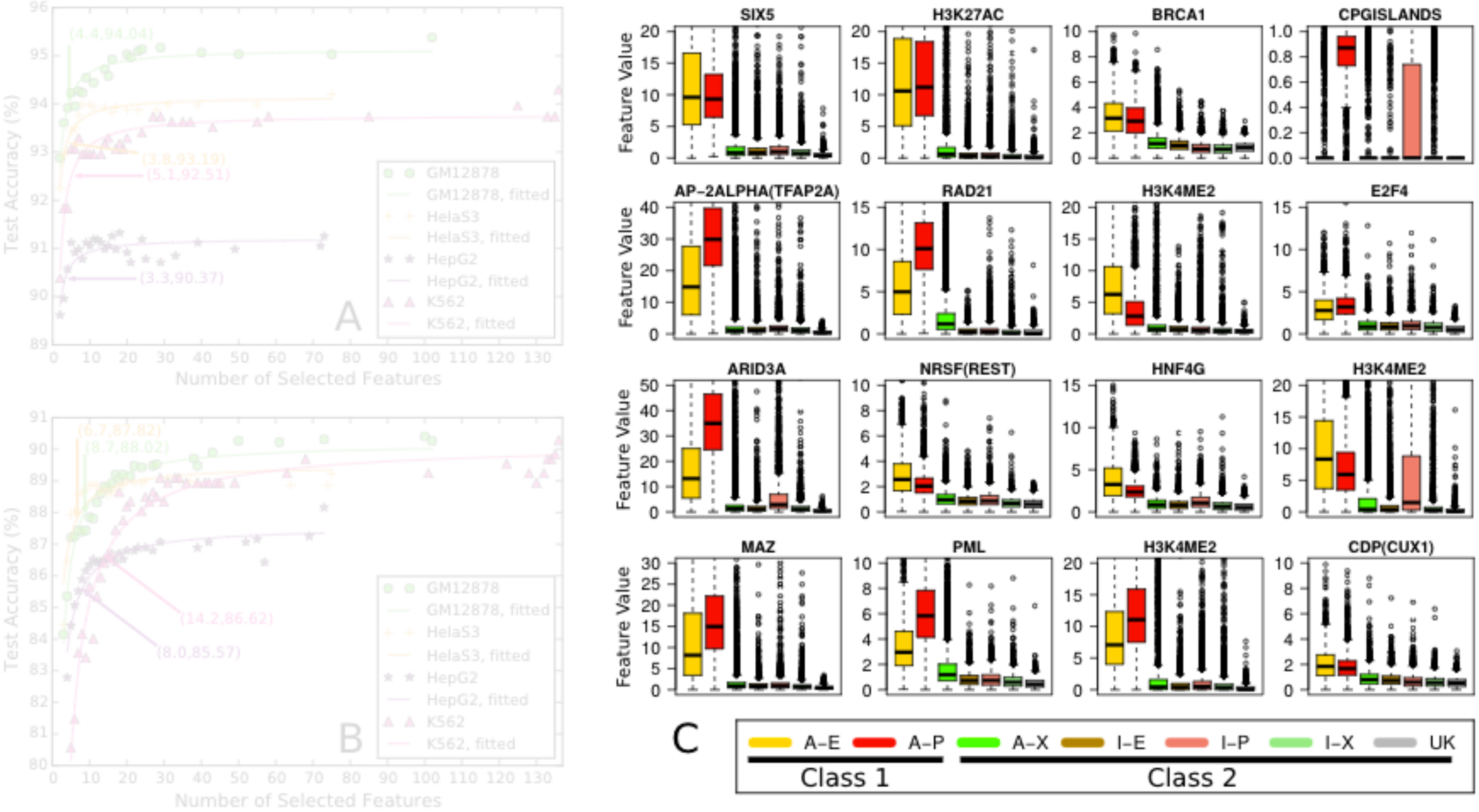


Figure 4: Feature analysis. (A) Accuracy versus the number of features incorporated into the model for 2-class prediction (distinguishing active CRRs (A-E + A-P) from BG (background: A-X, I-E, I-P, I-X and UK). The annotated points indicate where a line with slope 0.25 intersects a fitted curve). (B) Accuracy versus the number of features for a 3-class prediction (distinguishing A-E, A-P and BG). Points as described for (A). (C) For the top 4 features of the 2-class models generated for four well-characterized cell lines, box-plots depict the range of observed feature values (log2 scale) for 7 sequence classes.

Key Claims

- Identified “300,000 candidate enhancers genome wide (6.8% of the genome, of which 40,000 are supported by bidirectional transcription data) and 26,000 candidate promoters (0.6% of the genome)” in 1+ cell lines
- Predictions supported by other models and independent experimental data
- Deep feature selection enables some interpretability

Analysis

- Poor optimization of hyperparameters
  - "It is well-known that the model selection of neural networks is time consuming…"
- Not necessarily reproducible
  - Code but no trained models
  - How to combine ensemble of models ?
- No statistical comparison to randomized background
  - "Proximal genes…are consistent with...lineage"
  - "79% of predicted promoters are less than 5 kbps to the annotated gene TSSs, while 47% of predicted promoters are less than 5 kbps to the annotated gene TSSs"
- Unfair model comparisons (more/better data)
  - RF? Kernelized SVM? Shallow neural net?

Summary

- Key innovation
  - ***Better prediction of active cis-regulatory regions than previous models***
- Issues
  - ***Overstated claims***
  - ***Insufficient statistical comparison to background***
  - ***Lack of fair comparisons to other models***
- Impact / future directions
  - ***Emphasis on interpretability, based on minimal feature selection***
  - ***Extensibility to continuum of enhancer-promoter activity***

Thx.