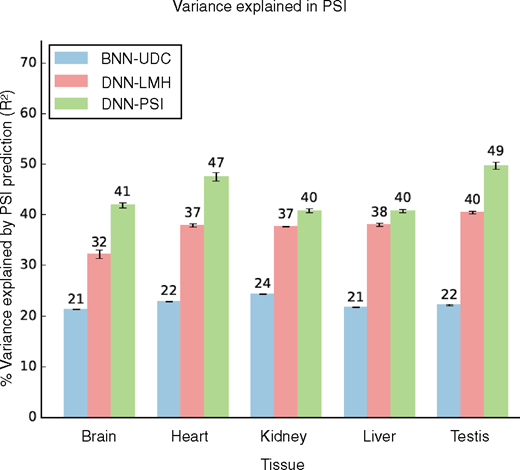
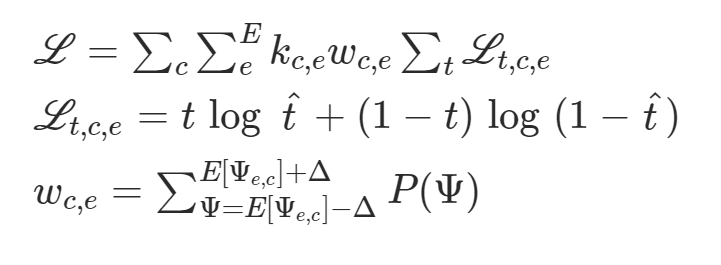
# Notes for CS 273b Paper Presentation

* Goal of paper is two-fold:
  + 1.) To increase the accuracy of AS predicting models using Deep Learning Methods
  + 2.) To incorporate additional data sources outside of the DNA sequence that may improve the predictions for the AS regulatory factors
* **Intro:**
  + 90% of human multi-exon genes are AS’d (Pan et al, 2008; Wang et al. 2008).
  + percent splicing inclusion (PSI, Ψ) -measured historically
    - “Ψ serves to capture the proportion of isoforms that include the alternative cassette exon versus those that skip it”
  + Previous work showed that BNN’s performed well
  + However, autoencoder DNN did better
  + “ Thus, in this work we reconstructed previous BNN and DNN models on the original dataset from (Leung *et al.*, 2014) to establish a baseline. Afterwards, we monitored the effect of a new target function, of increasing dataset size by exploiting improvements in RNA-Seq quantification algorithms (Vaquero-Garcia *et al.*, 2016), and adding new types of experimental data.”
  + Moved from an indirect formulation of PSI ({pst,e∣∣0≤pst,e≤1,∑spst,e=1}) to a direct one which improved the variance explained by PSI prediction:
    - 
  + Additional data: “CLIP-Seq based measurements of *in vivo* splice factors binding are turned into an additional set of input features while knockdown and over-expression experiments are added with binary vectors coding the tissue and splice factor (if any) measured.”
* **Dataset:**
  + TWO RNA-seq datasets
    - Brawand et al. 2011 data (also used in Leung et al. 2014)
      * 5 mouse tissues: brain, heart, kidney, liver, and testis
    - MGP data from Keane et al. 2011
      * 6 mouse tissues: heart, hippocampus, liver, lung, spleen, and thymus
      * Average read coverage = 60 million reads
  + CLIP-seq data
    - 15 CLIP-seq experiments were added for supplementary data
    - Gives noisey measurements of where splice factor binds.
    - Abstract the data into binary data for binding in areas of interest
* **Likelihood Learning function**
  + “Although useful, this target function suffers from several deficiencies when applied to RNA-Seq data. First, the binning results in a rudimentary estimation of ΨΨ and ΔΨΔΨ. Second, the optimization only aims to bring pec,spc,se and qec,sqc,se closer, without any relation to order or meaning. For example, if a cassette event has low inclusion (qc,s=L∼1qc,s=L∼1) then predicting pc,s=M∼1pc,s=M∼1 or pc,s=H∼1pc,s=H∼1 are just as bad. Moreover, in cases where an event suffers from insufficient or highly variable read coverage we may have qc,s=L∼qc,s=M∼qc,s=Hqc,s=L∼qc,s=M∼qc,s=H. In such cases, a model with prediction pc,s=H∼1pc,s=H∼1 based on sequence features will be penalized, even though there was no substantial evidence against it.”
  + 
* **Architecture**
  + BNN
    - 1 hidden layer
    - Varying number of sigmoid hidden untis
    - Weights are gaussian distribution w/ spike and slab prior to encourage sparsity
  + Original DNN
    - Autoencoder layer with tanh activation
    - 2 hidden layers with ReLU activation
    - Dropout with probability 0.5 was used in each layer except the autoencoder layer.
  + New DNN
    - Adding L1/L2 regularization had no impact on performance so it was excluded
    - Added 874 CLIP features as input
    - learning rates of the three target variables to vary to capture optimal model performance
* **Learning**
  + Trained first layer of model as Autoencoder to reduce the dimensionality of the data
    - Weights from the first layer were fixed and then rest of model is trained
    - Trained for 300-500 epochs
  + Trained with SGD w/ momentum
  + Feedworward network
    - Trained for 1000-1500 epochs
  + Weight tuning with backpropagation
  + Varying learning rates of PSI and inclusion and exclusion of dPSI improved performance
  + Once hyperparameters were fixed, the model was trained on both validation and training data
* **Results**
  + Measured R^2 and AUROC (for dPSI’s >= 0.15 or =< 0.05)
  + Compared to Leung et al. Paper the new model performs better
    - Leung et al. results/model were reconstruct to a comparable degree
  + Cassettization:
    - Detects and quantifies additional cassete and cassette like exons from the RNA-seq data.

**QUESTIONS:**

**What is MAJIC?**