Integrative Deep Models For Alternative Splicing

Jha, Anupama et al.

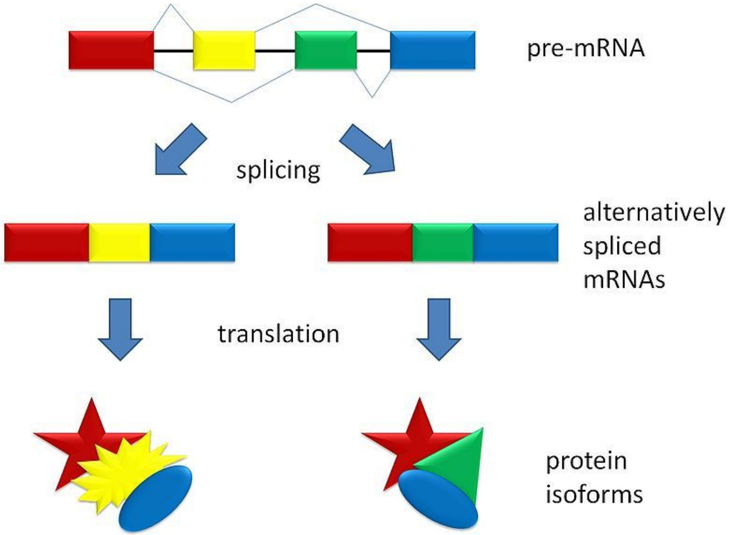
Group 11: Sebastian Astiz Le Bras, Ana-Maria Istrate, Michael Painter

Motivation

- High Conservation between Tissues in Alternative Splicing
- More than 90% of human multi-exon genes are alternatively spliced
- AS studied experimentally using RNA-Seq and CLIP-Seq
- Cannot directly measure from some parts of genome.
- Direct measurement is noisy and sparse.
- Growth in sequencing data makes computational approaches more feasible

Alternative Splicing (AS)

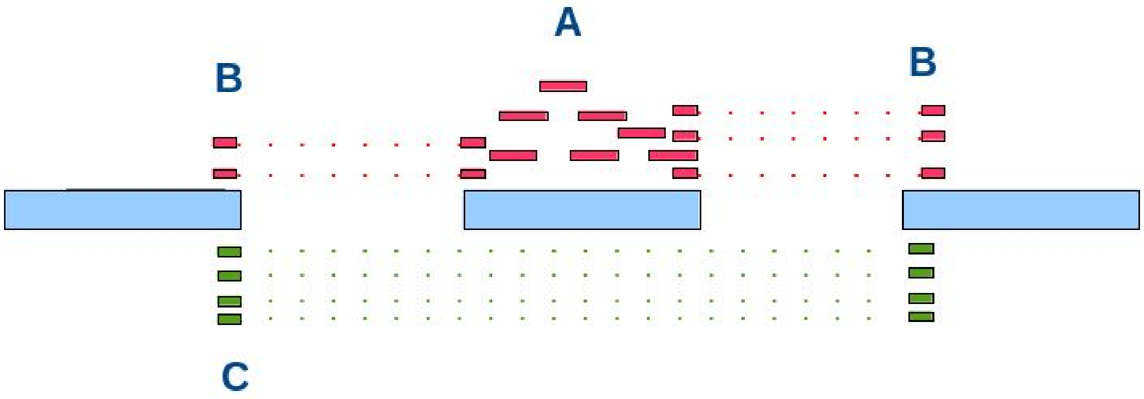
- One gene codes for multiple proteins.
- Cassette exon: may be excluded during splicing.
- Interested in the percent splicing index (PSI, or Ψ), the probability an exon is included during splicing.



Source: http://jonlieffmd.com/blog/alternative-rna-splicing-in-evolution

Percent Splicing Index

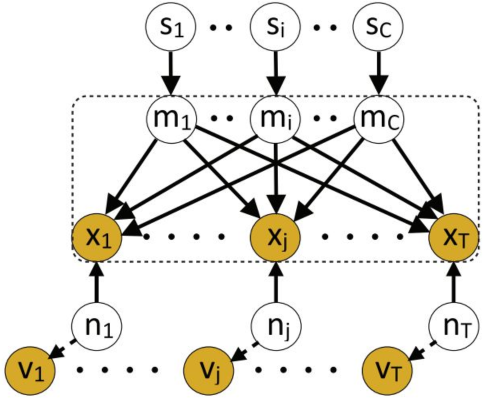
- Highly dependent on environment, or “condition”.
- Differential PSI (dPSI): $\Delta\Psi = \Psi_{C1} - \Psi_{C2}$, for conditions C1, C2.
- PSI = (A + B) / (A + B + C).



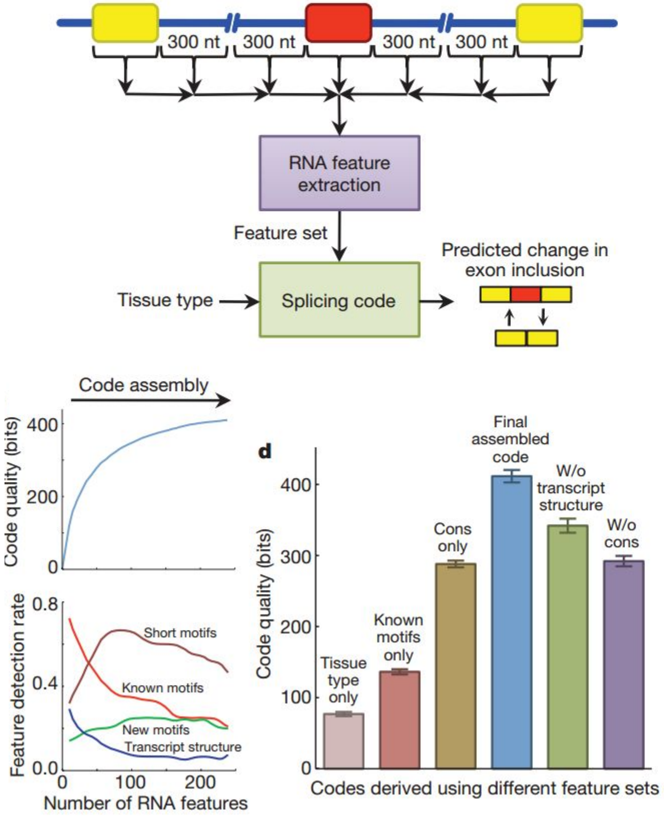
Source: http://geuvadiswiki.crg.es/index.php/Percentage_Splicing_Index

Previous Work - Pre Neural Networks

- Simple/manual feature extraction.
- Different ways to implement splicing code.
- Maximise “code quality”.
- Coarse output (low/medium/high PSI).



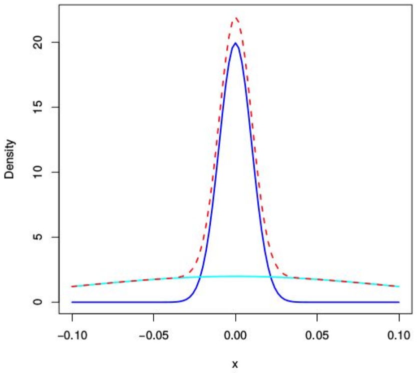
Source: Barash et al.



Source: Barash et al.

Previous Work - Neural Networks

- Use neural networks for “splicing code”.
- Two approaches: Bayesian (BNNs) and Deep (DNNs).
- Coarse output still.



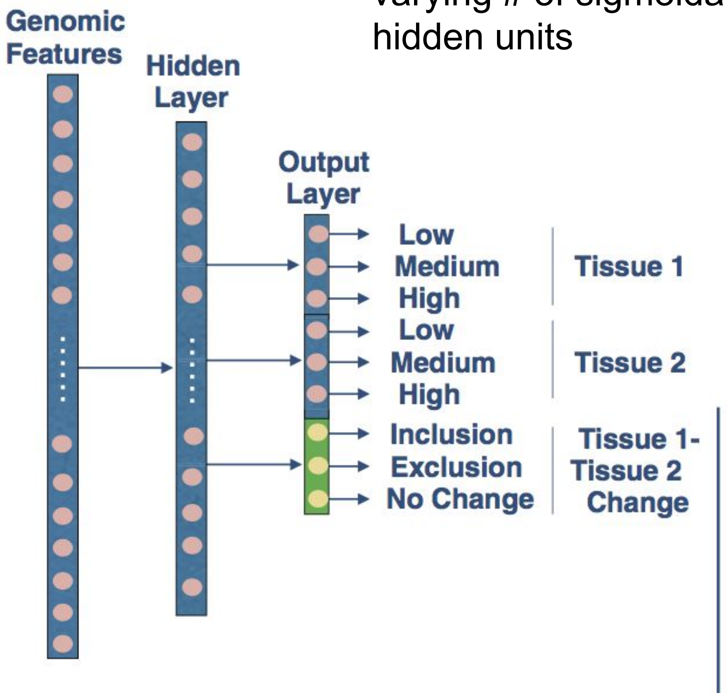
Source: https://stats.stackexchange.com/questions/180564/how-to-create-a-spike-and-slab-prior-plot-in-r

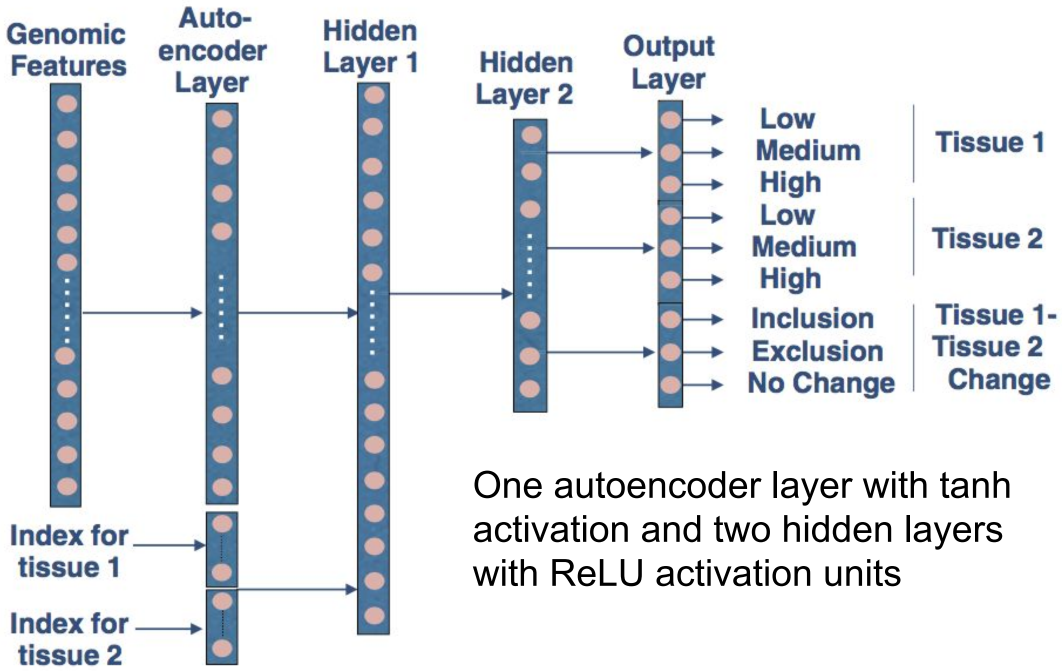
$$p(y|x) = \int p(\theta)p(y|x,\theta)\, d\theta$$

Bayesian Neural Network

Leung's Deep Neural Network

One hidden layer with varying # of sigmoidal hidden units





One autoencoder layer with tanh activation and two hidden layers with ReLU activation units

Initial research proved it was favorable to KNN, SVMs, Naive Bayes and DNN with dropouts

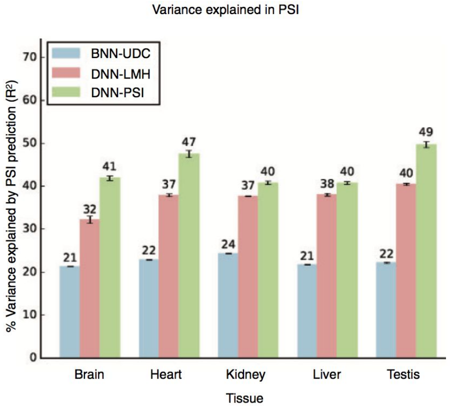
Subsequent research showed superior to the BNN model

Addition 1: new target functions Continuous output/prob regression rather than classification now

- Previous work unable to predict Ψ and $\Delta\Psi$ directly
- Initially, for any exon e in a given condition t, 3-way prediction task: $\{p^s_{t,e} | 0 \leq p^s_{t,e} \leq 1, \sum_s p^s_{t,e} = 1\}$

s = chances of inclusion, exclusion, no change

- Then, Ψ was seen as having three levels: ‘Low’ (0 < Ψ < 33%), ‘Medium’ (33% < Ψ < 66%) and ‘High’ (66% < Ψ < 100%)
- Now: new target functions which model Ψ directly
- => improvement in the % in variance explained by Ψ



New likelihood function

$$T_{\Psi_{e,c}} = E[\Psi_{e,c}]$$

$$T_{\Delta\Psi_{inc,c,c'}} = |\max(\epsilon, E[\Delta\Psi_{c,c'}])|$$

$$T_{\Delta\Psi_{exc,c,c'}} = |\min(\epsilon, E[\Delta\Psi_{c,c'}])|$$

- 2 => captures the dPSI for events with increased inclusion between condition c and c’

- 3=> captures the dPSI for events with increased exclusion between condition c and c’

$$\mathcal{L} = \sum_{c} \sum_{e}^{E} k_{c,e} w_{c,e} \sum_{t} \mathcal{L}_{t,c,e}$$

$$\mathcal{L}_{t,c,e} = t \log \widehat{t} + (1-t) \log (1-\widehat{t})$$

$$w_{c,e} = \sum_{\Psi=E[\Psi_{e,c}]-\Delta}^{E[\Psi_{e,c}]+\Delta} P(\Psi)$$

- t is one of the target functions
- k = 1 if exon is quantifiable in condition c

Addition 1: updating previous models to be able to compare them

BNN: supplemented the LMH (Low, Medium, High) Ψ variables with UDC variables (for inclusion levels going up, down or not changing)

=> made the BNN targets equivalent to those of the DNN and improved performance

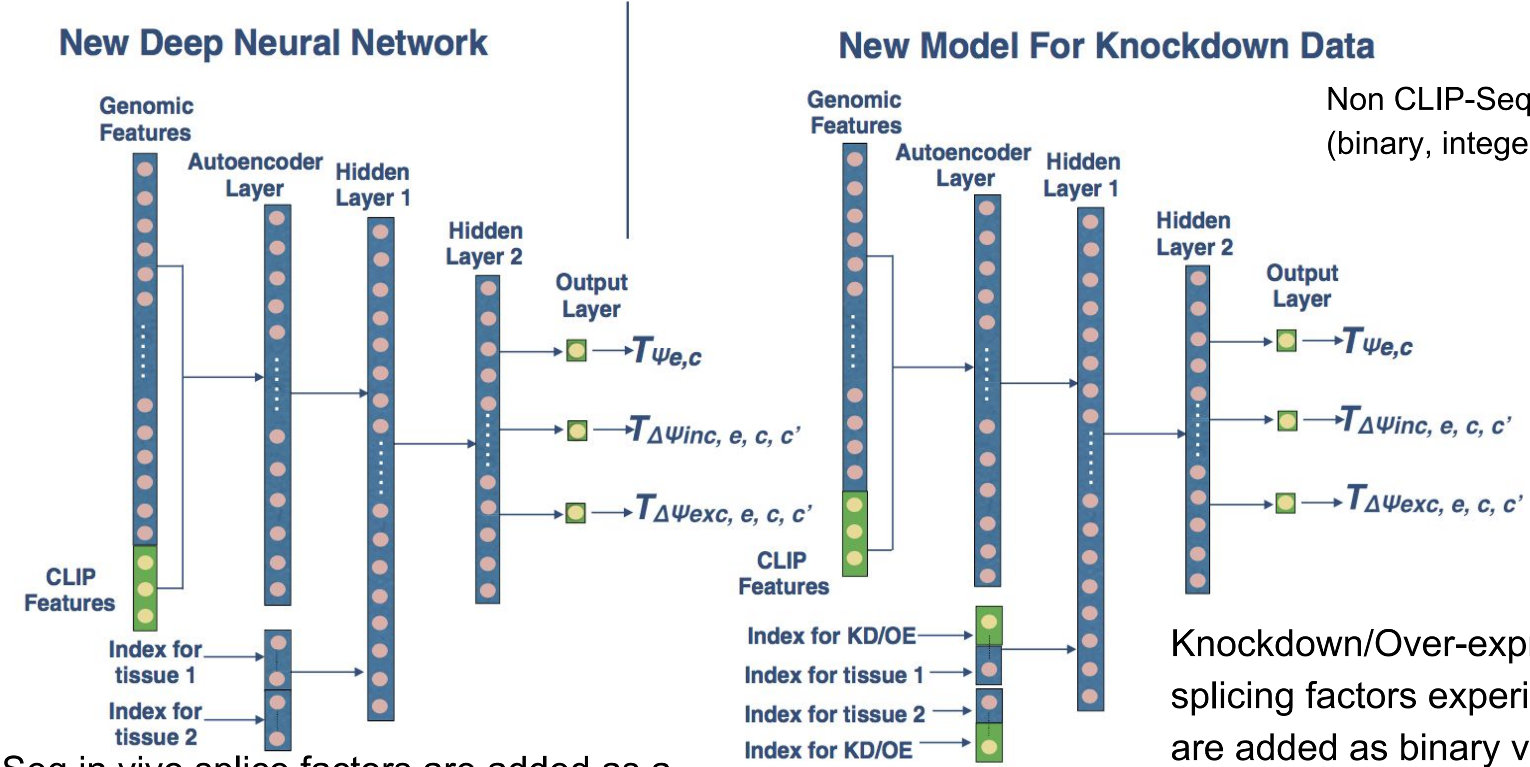
DNN: tissue type was input as two hot vectors when comparing two tissues

experimented with different types of network architectures/different types of hidden layers/units/activation units/batch normalizations => no significantly better results, so DNN architecture was not changed

added 874 CLIP features to the dataset

Addition 2: integrating experimental data into the model

Previous: Non-integrative models, require the same (≈12000) cassette exons trained on.

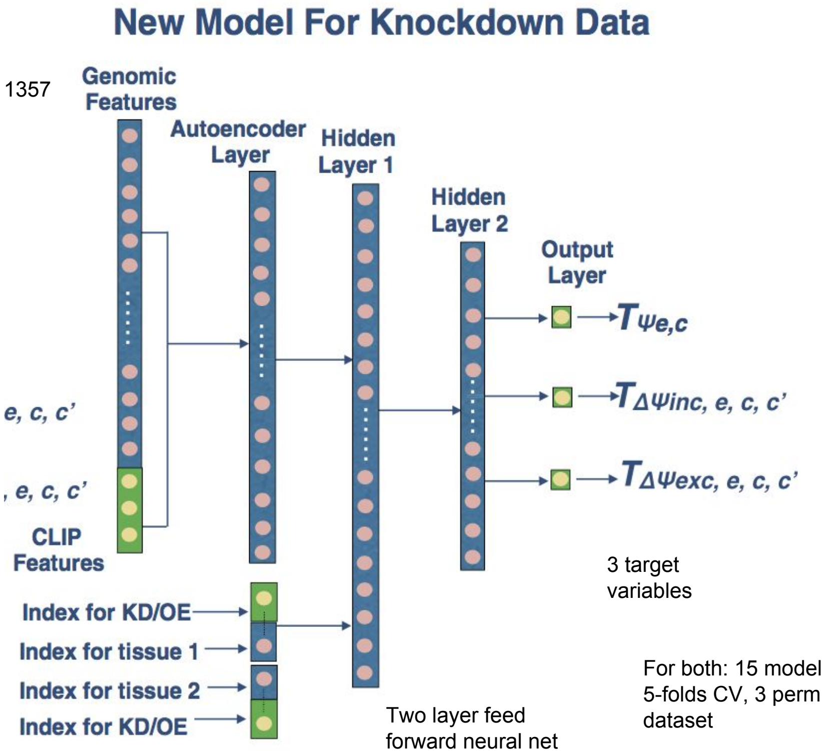


Non CLIP-Seq features (binary, integer, real)

CLIP-Seq in vivo splice factors are added as a set of input features

Knockdown/Over-expression splicing factors experiments are added as binary vectors coding the tissue and splice factor

Training the model



1357

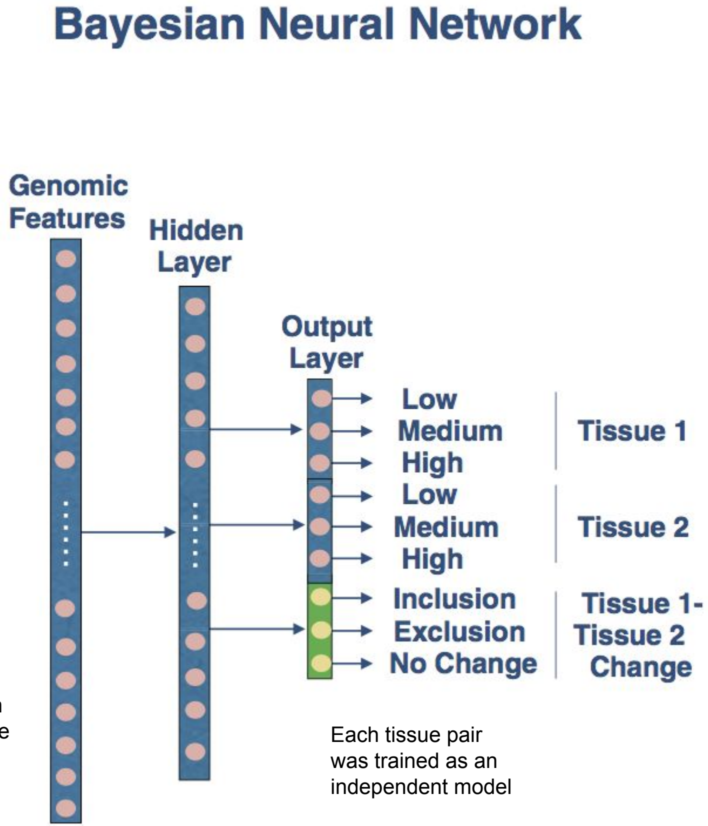
874

3 target variables

Fixed weights + tissue input was added

Two layer feed forward neural net was trained using SGD with momentum, weights tuned by backpropagation

For both: 15 models trained with 5-folds CV, 3 permutations of the dataset



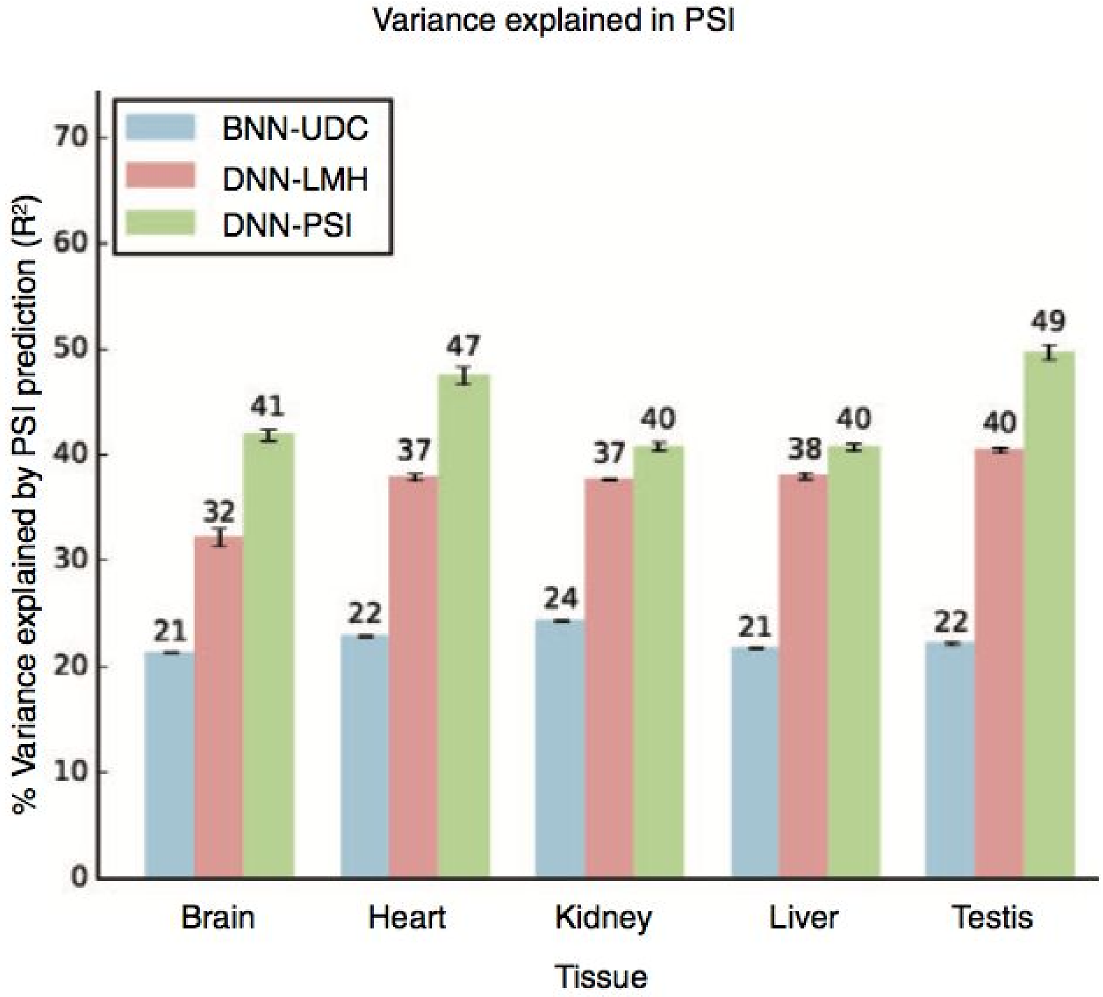
Each tissue pair was trained as an independent model

Final predictions generated by averaging over predictions from sampled weights

Data

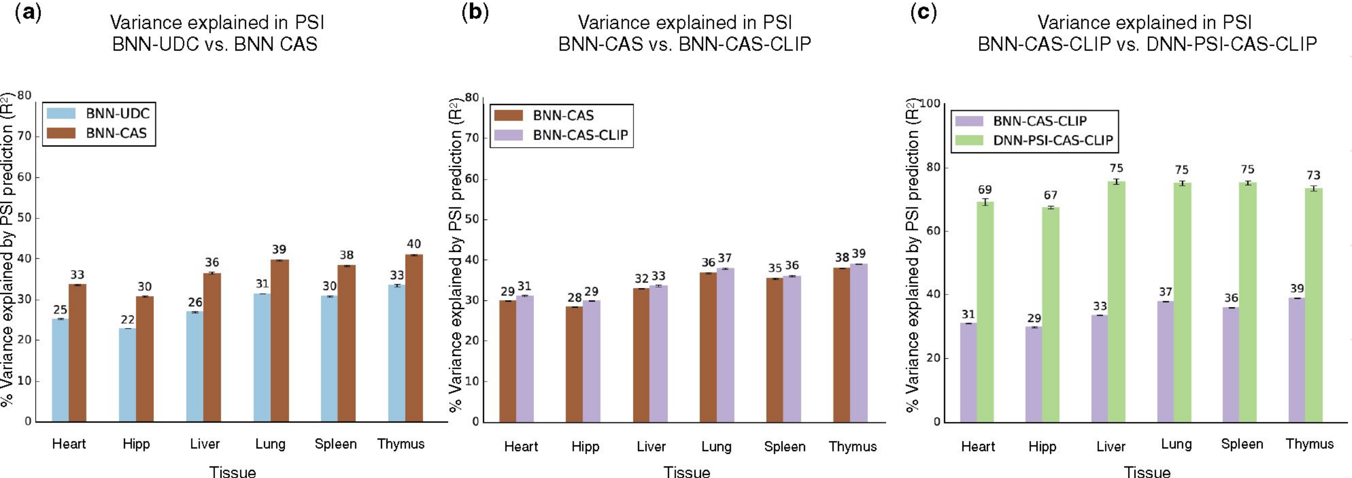
- RNA-seq data:
  - 11,019 mouse alternative exons from brain, heart, kidney, liver and testis
  - 1393 features in 55 groupings describing the exon, ints neighboring introns and adjacent exons.
- CLIP-seq data
  - 15 CLIP-seq experiments
  -

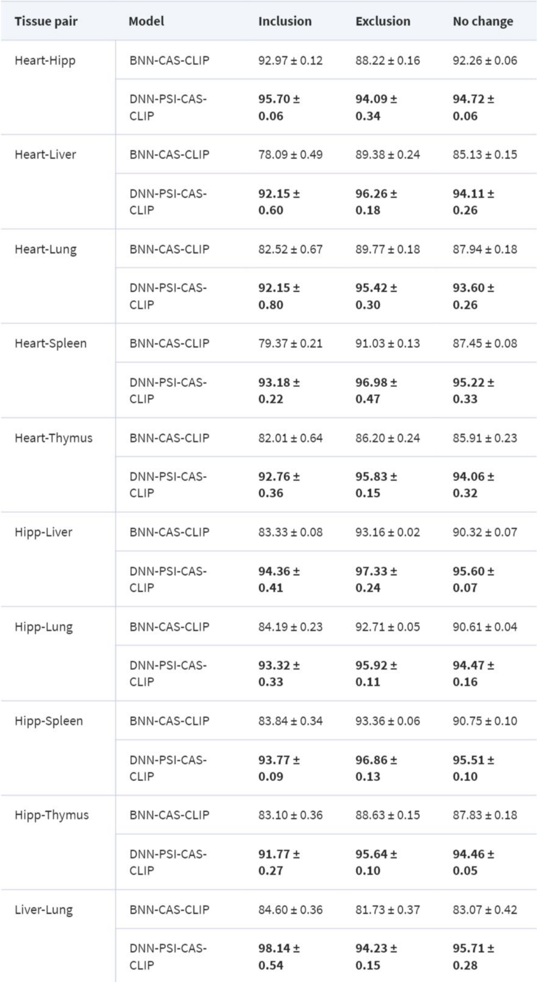
Results


Variance explained in PSI
BNN-UDC
DNN-LMH
DNN-PSI
% Variance explained by PSI prediction (R²)
0
10
20
30
40
50
60
70
21
32
41
22
37
47
24
37
40
21
38
40
22
40
49
Brain
Heart
Kidney
Liver
Testis
Tissue

Results

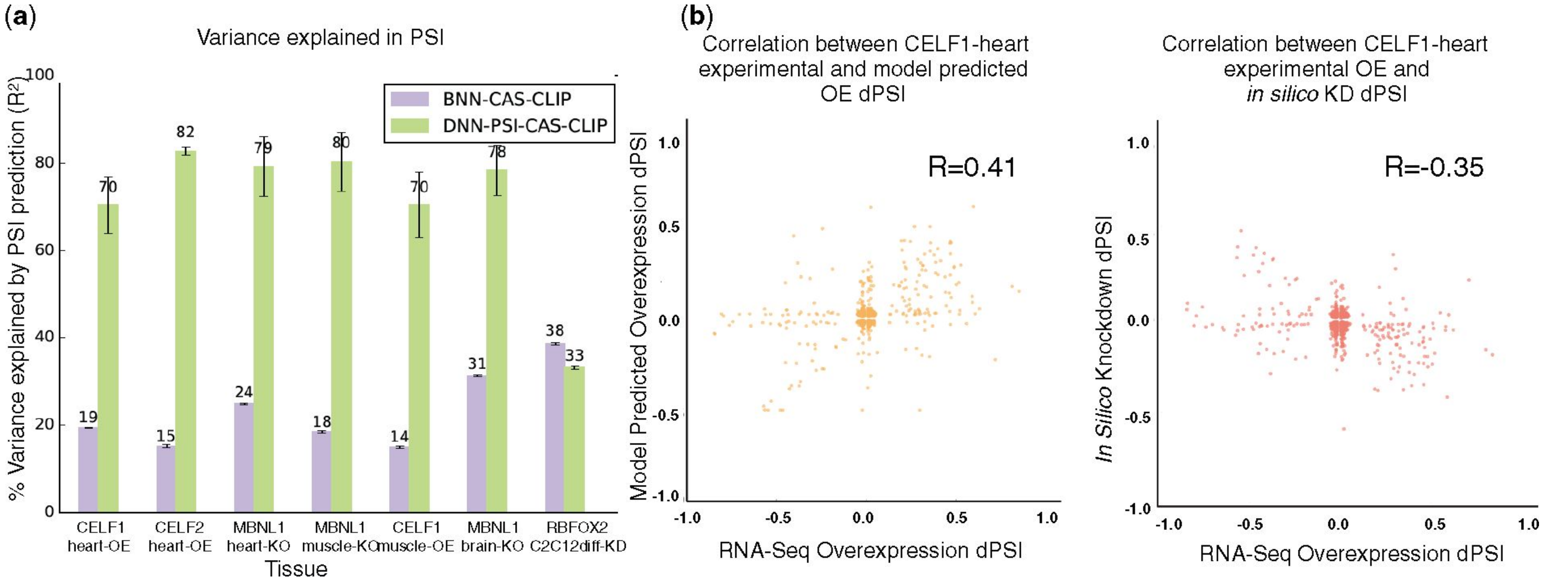
BNN-UDC	BNN model w/ added Up, Down, not Changing Prediction
BNN-CAS	BNN model w/ Cassetization
BNN-CAS-CLIP	BNN model w/ Cassetization and CLIP-Seq data
DNN-PSI-CAS-CLIP	DNN model w/ new target function, CLIP data, and Cassetization





Tissue pair	Model	Inclusion	Exclusion	No change
Heart-Hipp	BNN-CAS-CLIP	92.97 ± 0.12	88.22 ± 0.16	92.26 ± 0.06
	DNN-PSI-CAS-CLIP	**95.70 ± 0.06**	**94.09 ± 0.34**	**94.72 ± 0.06**
Heart-Liver	BNN-CAS-CLIP	78.09 ± 0.49	89.38 ± 0.24	85.13 ± 0.15
	DNN-PSI-CAS-CLIP	**92.15 ± 0.60**	**96.26 ± 0.18**	**94.11 ± 0.26**
Heart-Lung	BNN-CAS-CLIP	82.52 ± 0.67	89.77 ± 0.18	87.94 ± 0.18
	DNN-PSI-CAS-CLIP	**92.15 ± 0.80**	**95.42 ± 0.30**	**93.60 ± 0.26**
Heart-Spleen	BNN-CAS-CLIP	79.37 ± 0.21	91.03 ± 0.13	87.45 ± 0.08
	DNN-PSI-CAS-CLIP	**93.18 ± 0.22**	**96.98 ± 0.47**	**95.22 ± 0.33**
Heart-Thymus	BNN-CAS-CLIP	82.01 ± 0.64	86.20 ± 0.24	85.91 ± 0.23
	DNN-PSI-CAS-CLIP	**92.76 ± 0.36**	**95.83 ± 0.15**	**94.06 ± 0.32**
Hipp-Liver	BNN-CAS-CLIP	83.33 ± 0.08	93.16 ± 0.02	90.32 ± 0.07
	DNN-PSI-CAS-CLIP	**94.36 ± 0.41**	**97.33 ± 0.24**	**95.60 ± 0.07**
Hipp-Lung	BNN-CAS-CLIP	84.19 ± 0.23	92.71 ± 0.05	90.61 ± 0.04
	DNN-PSI-CAS-CLIP	**93.32 ± 0.33**	**95.92 ± 0.11**	**94.47 ± 0.16**
Hipp-Spleen	BNN-CAS-CLIP	83.84 ± 0.34	93.36 ± 0.06	90.75 ± 0.10
	DNN-PSI-CAS-CLIP	**93.77 ± 0.09**	**96.86 ± 0.13**	**95.51 ± 0.10**
Hipp-Thymus	BNN-CAS-CLIP	83.10 ± 0.36	88.63 ± 0.15	87.83 ± 0.18
	DNN-PSI-CAS-CLIP	**91.77 ± 0.27**	**95.64 ± 0.10**	**94.46 ± 0.05**
Liver-Lung	BNN-CAS-CLIP	84.60 ± 0.36	81.73 ± 0.37	83.07 ± 0.42
	DNN-PSI-CAS-CLIP	**98.14 ± 0.54**	**94.23 ± 0.15**	**95.71 ± 0.28**

Results



(a) Variance explained in PSI

(b) Correlation between CELF1-heart experimental and model predicted OE dPSI

Correlation between CELF1-heart experimental OE and *in silico* KD dPSI

Takeaways

- Adding new data sources can improve model Accuracy
    - Most likely by handling over fitting
- New Target function
    - Increases Accuracy of model
    - Provides continuous rather than discrete outputs for PSI
- Comparisons between DNN and BNN architectures
    - Re-evaluated existing work

Criticisms

- Deeper networks + L1/L2 regularization tried, but no data published
- “Cassettization” is a method introduced for feature extraction, but not explained.
- Vague on their hyperparameter selection and cross-validation methods.
- Problem with their cross-validation algorithm.
- CLIP-Seq, KD/OE experiments are noisy, have missing measurements - abstracting as binary indicators of binding may give false positives?

Impact and Future Work

Significant contribution to an area not very well researched, which opens up opportunities for future work in the field:

- Considering deep networks with multiple types of layers: CNN’s
- Extracting RNA features from core models (maybe RNN’s)
- Predicting Non-cassette splicing variations
- Generalizing current findings to other types of conditions and datasets

Questions?

References

[1] Jha, Anupama, Matthew R. Gazzara, and Yoseph Barash. "Integrative Deep Models for Alternative Splicing." bioRxiv (2017): 104869.

[2] Keane, Thomas M., et al. "Mouse genomic variation and its effect on phenotypes and gene regulation." Nature 477.7364 (2011): 289-294.

[3] Barash, Yoseph, et al. "Deciphering the splicing code." Nature 465.7294 (2010): 53-59.

[4] Barash, Yoseph, Benjamin J. Blencowe, and Brendan J. Frey. "Model-based detection of alternative splicing signals." Bioinformatics 26.12 (2010): i325-i333.

[5] Xiong, Hui Yuan, Yoseph Barash, and Brendan J. Frey. "Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context." Bioinformatics 27.18 (2011): 2554-2562.

[6] Leung, Michael KK, et al. "Deep learning of the tissue-regulated splicing code." Bioinformatics 30.12 (2014): i121-i129.