Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data

Nikolaus Fortelny and Christoph Bock

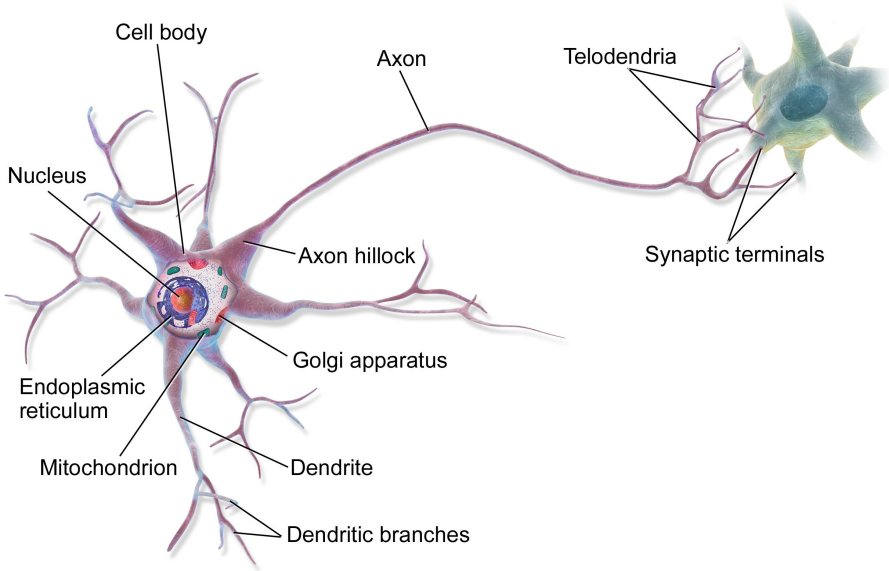
Genome Biology

Presented by Ethan Weinberger and Lee Organick

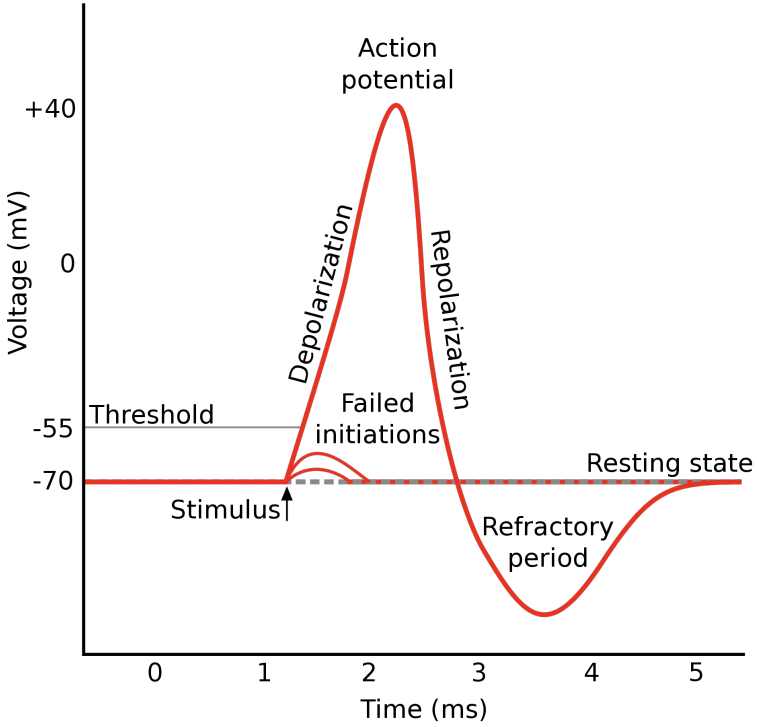
A (brief) primer on neural nets

Biological Neurons

- Neurons receive inputs on *dendrites*
- Enough stimulation “activates” the neuron
- Sends signal along its axon to other neurons

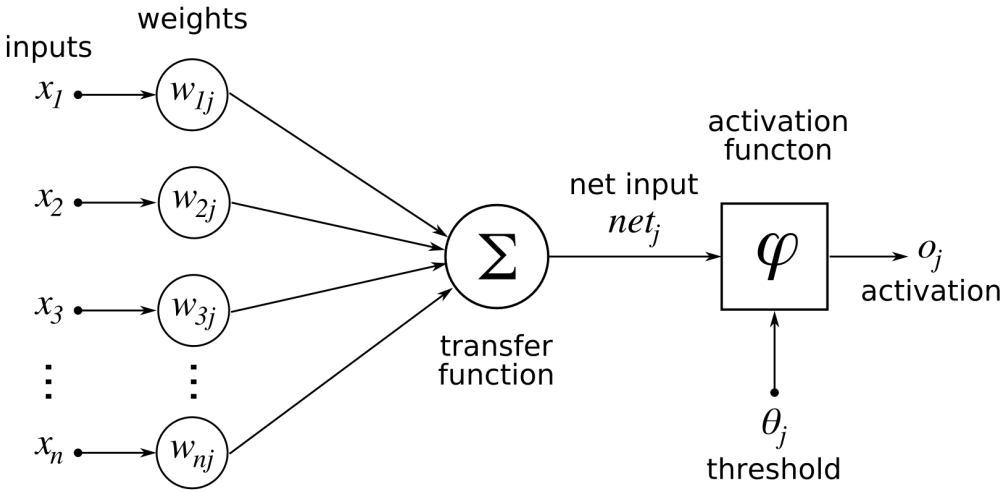


Neuron Activation

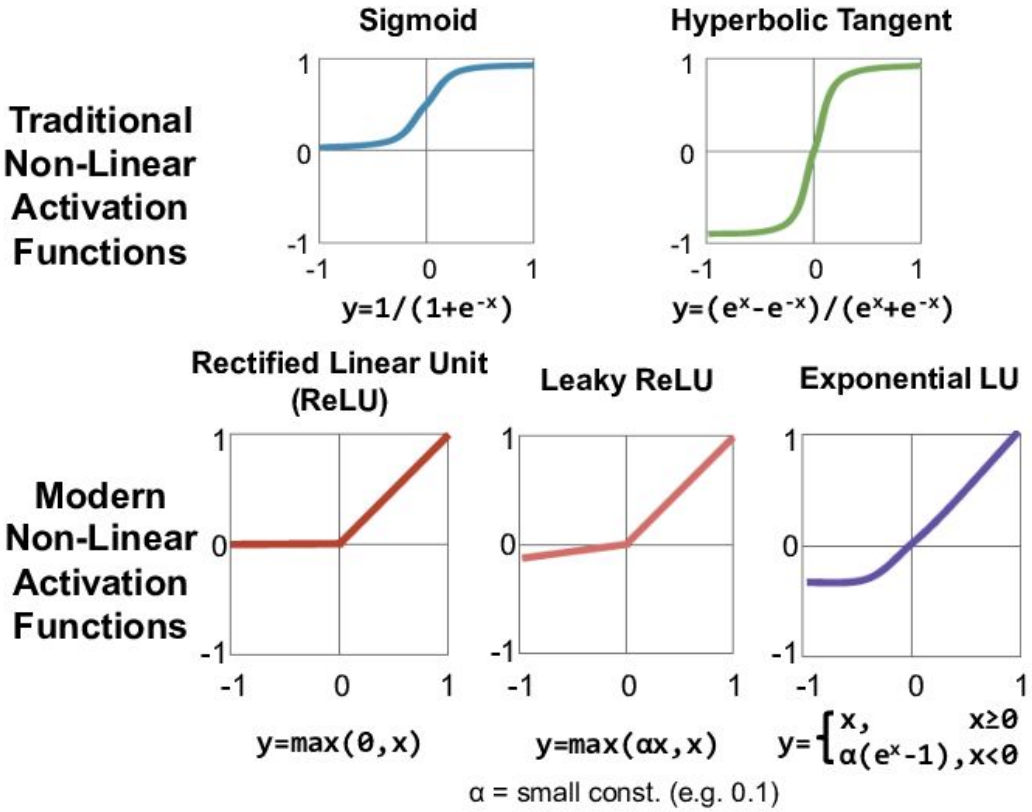


Artificial Neurons

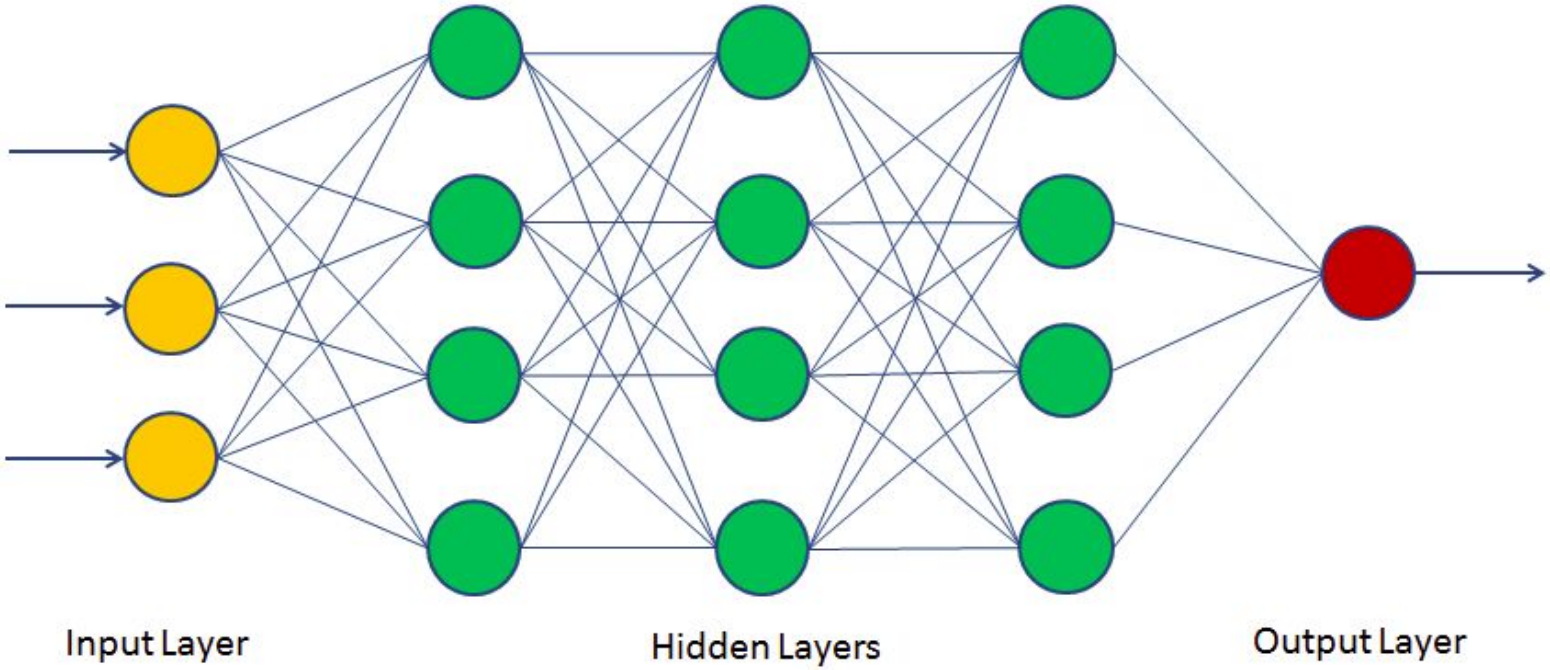
- Edges like dendrites/axons
- Inputs to edges multiplied by edge weights → summed up to “activate” neurons



Activation Functions



Neural Networks



Problem

- Models are hard to interpret
- Too many parameters for a human to comprehend
- Intermediate nodes don’t correspond to interpretable concepts

Interpretable ML- Bio edition

AKA- why Lee's excited for this revolution

1) No black box → fewer dumb errors
2) No black box → potentially less bias
3) No black box → faster results?
4) No black box → better results?

Interpretable ML- Bio edition

AKA- why Lee's excited for this revolution

1) No black box → fewer dumb errors
2) No black box → potentially less bias
3) No black box → faster results?
4) No black box → better results?

Learned that an address to the specialty clinic was more likely to be a specific kind of cancer*

*I could not find the paper on this, maybe I saw it in a casual presentation of someone's work?

Interpretable ML- Bio edition

AKA- why Lee's excited for this revolution

1) No black box → fewer dumb errors
2) No black box → potentially less bias
3) No black box → faster results?
4) No black box → better results?

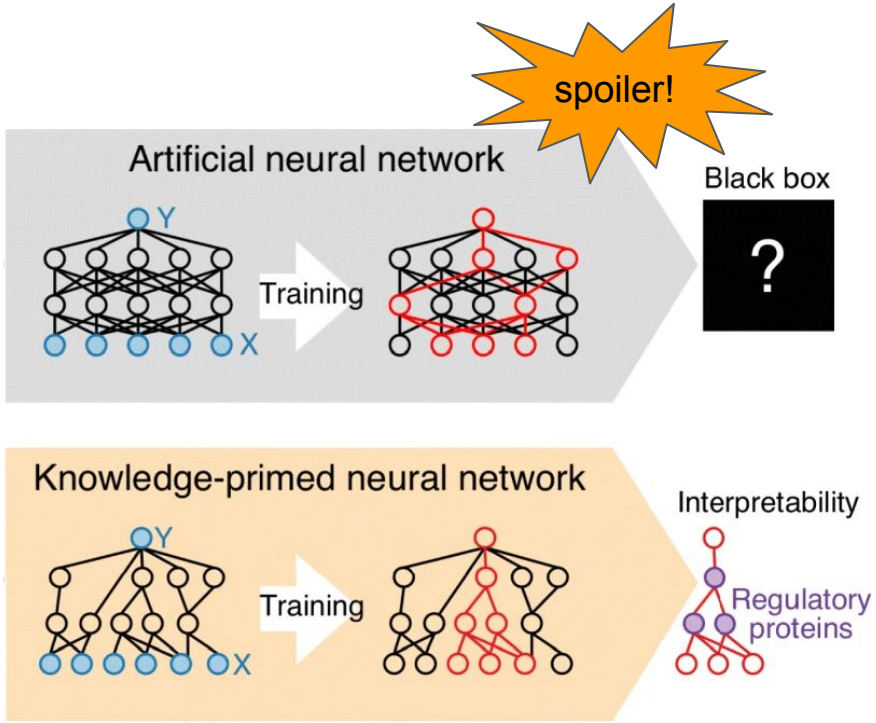
"Within the field of anaesthesiology, a preliminary multicentre analysis of data from 40 institutions by White and colleagues11 revealed that Black patients received inferior care (with respect to postoperative nausea and vomiting prophylaxis) both in aggregate and individually at nearly every single centre."

Bias and ethical consideration in machine learning and the automation of perioperative risk assessment.
British Journal of Anaesthesia. 2020. *O'Reilly-Shah et al.*
DOI: https://doi.org/10.1016/j.bja.2020.07.040

Interpretable ML- Bio edition

AKA- why Lee's excited for this revolution

1) No black box → fewer dumb errors
2) No black box → potentially less bias
3) No black box → faster results?
4) No black box → better results?



This paper's KPNN is much sparser and has few layers.

Interpretable ML- Bio edition

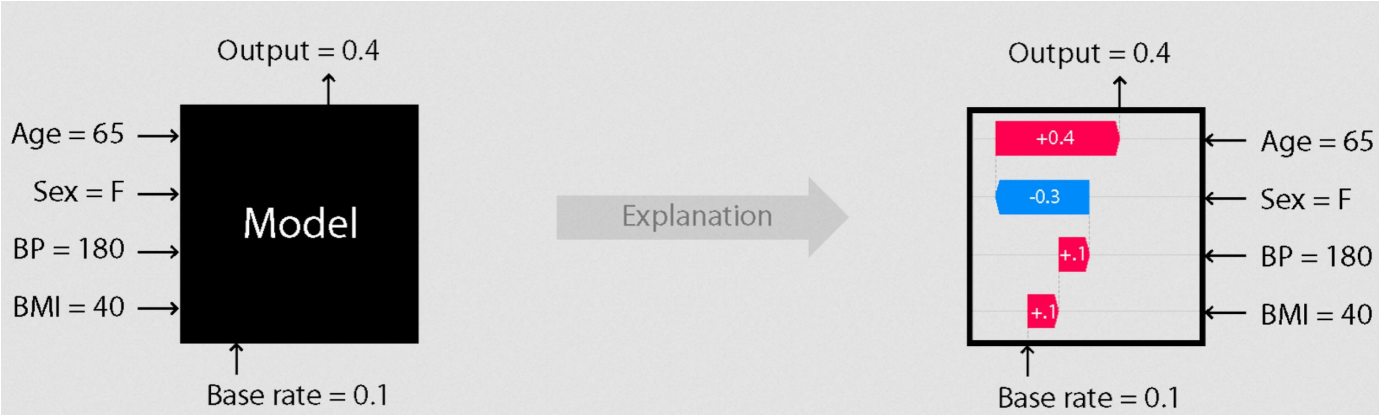
AKA- why Lee's excited for this revolution

1) No black box → fewer dumb errors
2) No black box → potentially less bias
3) No black box → faster results?
4) No black box → better results?

What does it mean to be "better"? More accurate? More equitable? More interpretable? More accountable?

Previous work on interpretability

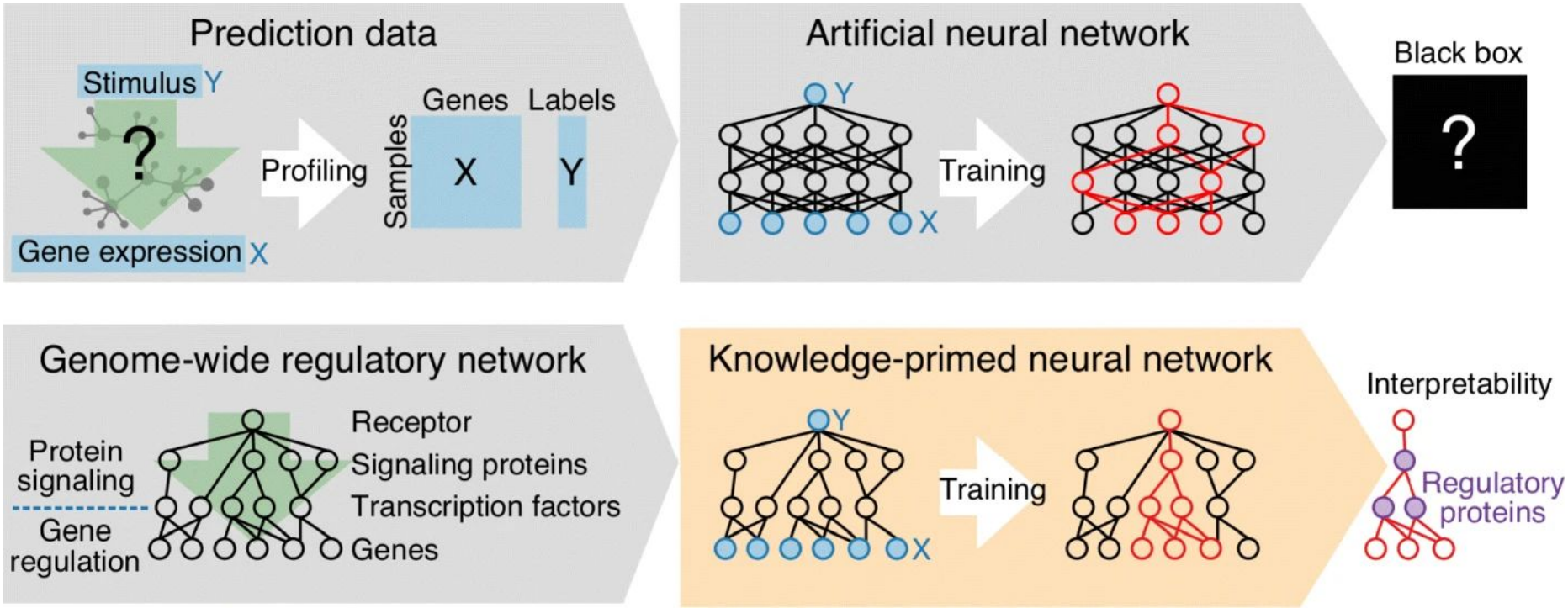
- Post-hoc (interpret a specific prediction after it's been made)
- What features were important for this prediction?



Problems with previous approaches

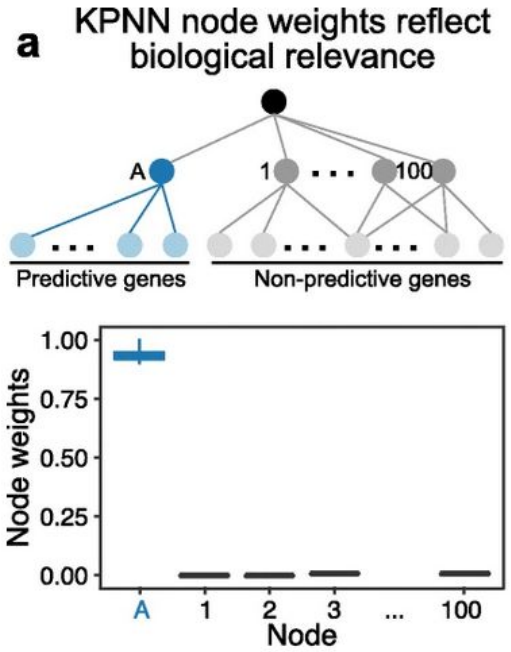
- Doesn’t help when you have lots of features (e.g. genes) or a hierarchy of concepts (e.g. genes → pathways)
  - Not super useful for biological discovery
- Post-hoc methods can be “tricked” with adversarial examples
  - Are these explanations meaningful?

KPNNs - knowledge-primed neural networks
Vs
ANNs - artificial neural network



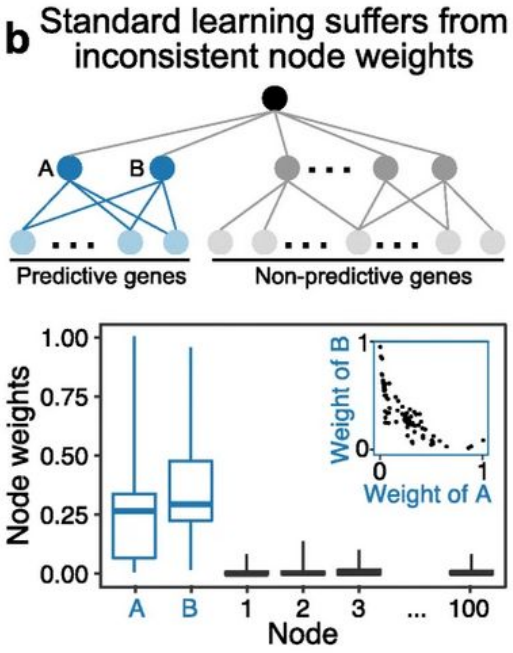
Experiment 1a: Simulated data

- One set of predictive genes connected to intermediate node (A)
- Other genes not predictive
- KPNN consistently gives A a much higher weight



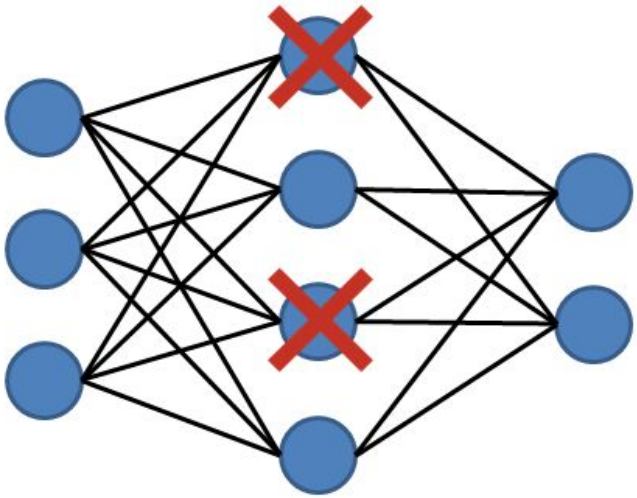
Experiment 1b: Simulated data

- Biological networks have redundancy in the real-world
- Multiple intermediate nodes connected to predictive genes
- Model weights are lower + have high variance :(



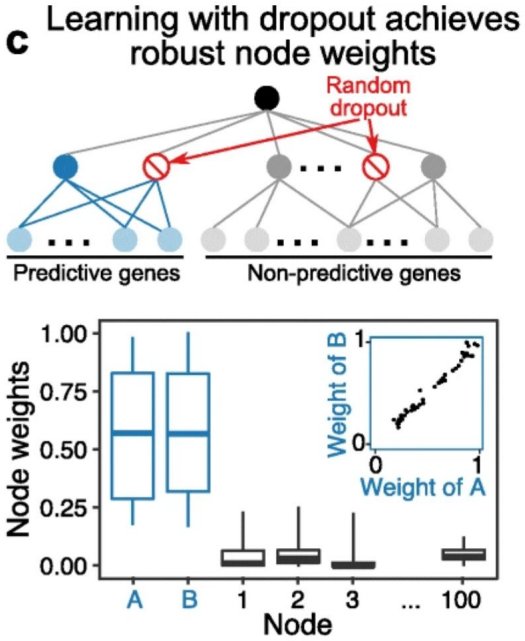
Solution: Dropout

- During training, zero-out nodes randomly
- Stops model from just fitting to one particular input → output relationship
- More likely to capture all relevant relationships

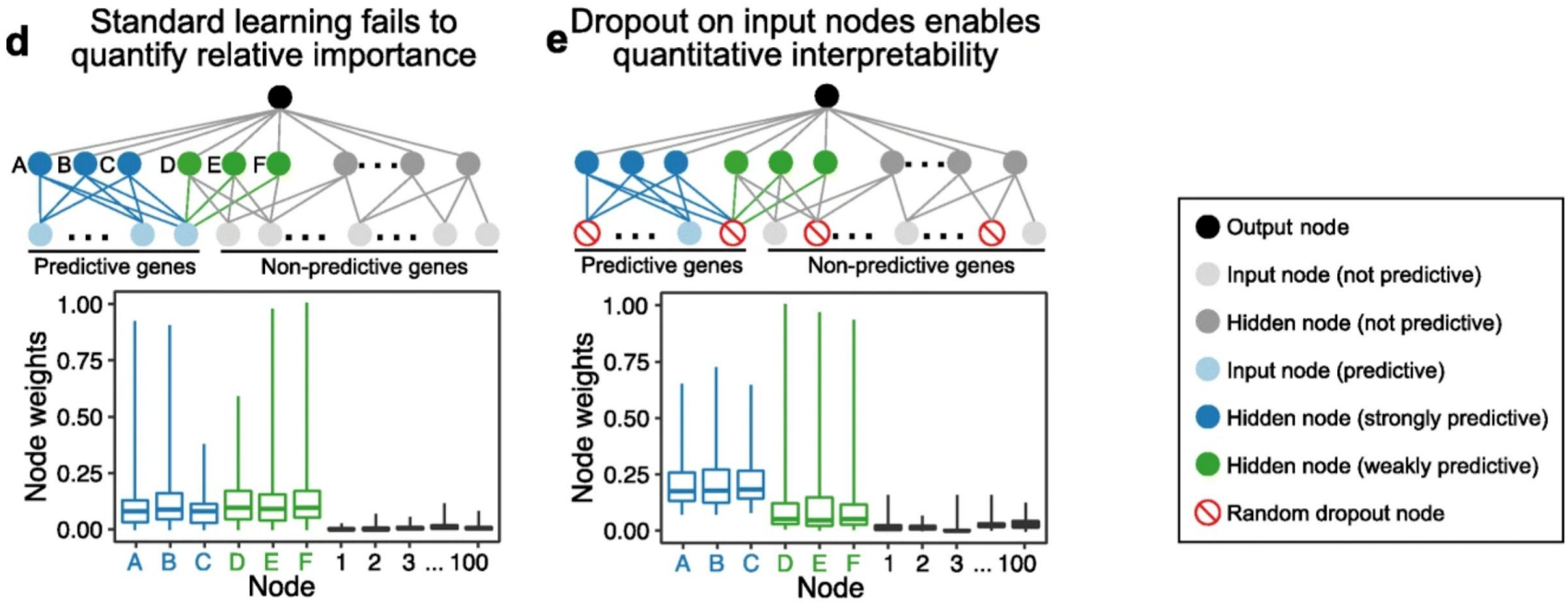


Dropout results

- Dropping-out intermediate nodes leads to multiple relationships being captured

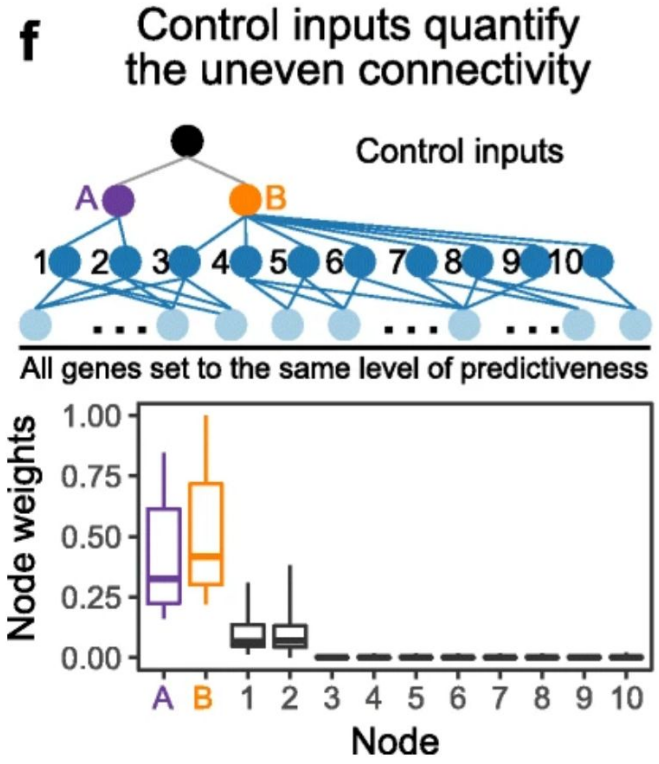


Dropout results



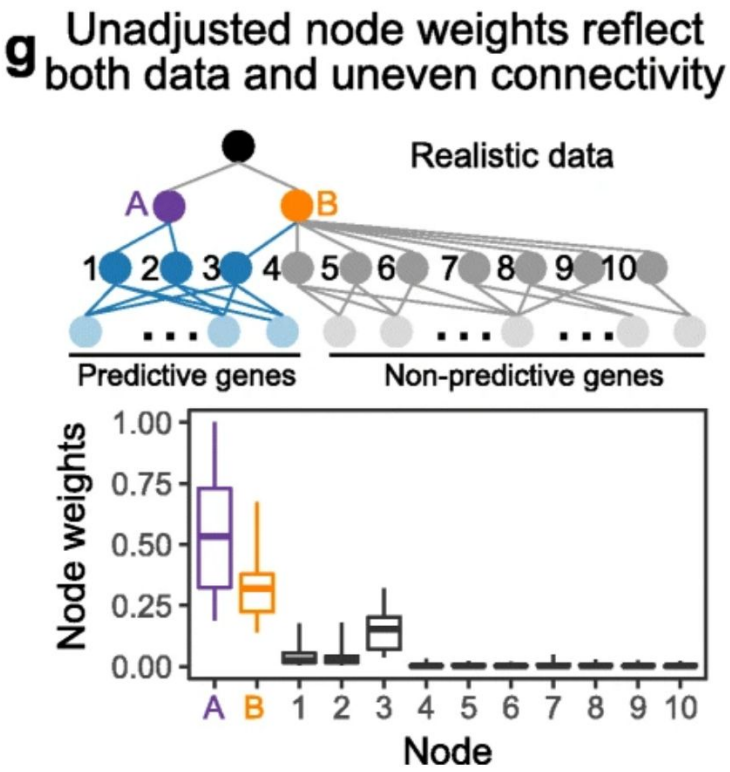
One more problem: Uneven connections

- Node weights might reflect connectivity rather than predictiveness
- Experiment on “control” genes with same amount of predictiveness

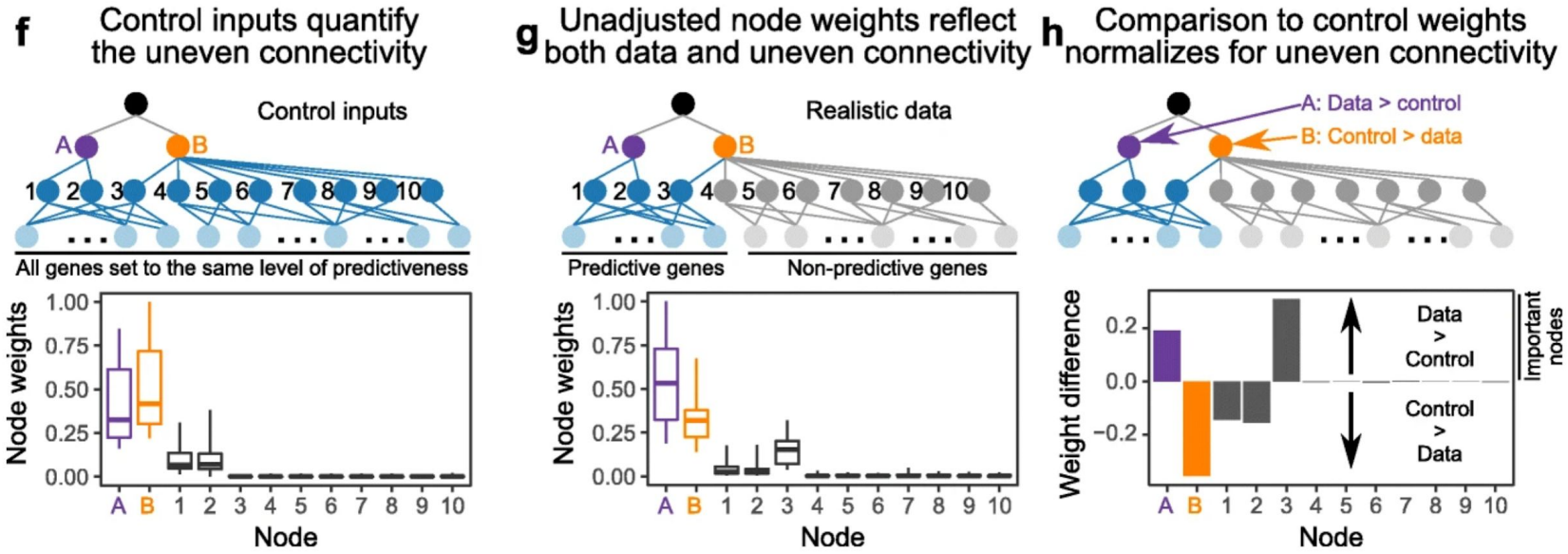


One more problem: Uneven connections

- “Non-predictive” intermediate node still has non-zero weight
  - Would expect near-zero given lack of predictivity of input genes



Node normalization



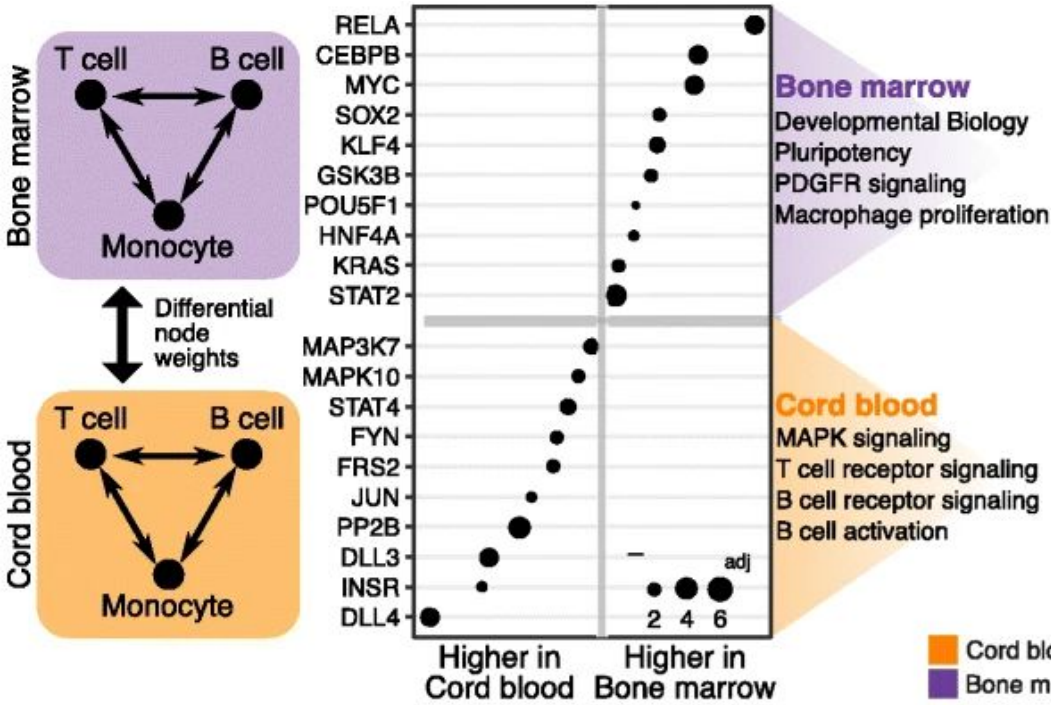
Validation on more complex datasets

Human cell atlas

- 500,000 transcriptomes
- 3 cell types
- 2 organs

Takeaway:

Could accurately predict cell type from gene expression in an ***interpretable*** way that corresponds to known biology



Discussion

- Weighing tradeoffs of accuracy vs. interpretability
  - What are the scenarios appropriate for each method?
  - Will this method inherently be less accurate?
    - Compared to other ML?
    - Compared to ground truth?
- Database problems (painful to set up, painful to sanity check as a biologist)
  - Are there problem sets with not enough biological data yet?
  - So far, there haven't been huge validation experiments (i.e. with high-throughput CRISPR screens), will we see different behavior?
- Are there problems this highly labeled node and edge structure will struggle with?
- Are we convinced KPNNs are the way forward for interpretable ML?