Caveats with prediction and interpretation

- Not all neural networks are equal. Be mindful of biases in datasets, problem formulation, loss functions and training strategies. These aspects matter way more than “architecture”.
- Not all interpretation methods are equal. Be aware of assumptions and caveats
- Interpretation is through the lens of a specific model
  - Test robustness to model design, data sampling, background sets
- Interpretation success depends on model performance
  - If you get your prediction wrong, your interpretation is going to wrong
  - You can get a prediction correct and still get a wonky interpretation (many models can explain the same output)