The Problem with ATAC-seq
- Chromatin is a compressed version of DNA. A gene cannot be expressed when it is compressed or closed.
- ATAC-seq is a sequencing technique used to determine what parts of chromatin are open or closed.
- The quality of ATAC-seq data is heavily dependent on many environmental factors, and therefore prone to noise.

Model Design and Selection

Architecture Search: Random Search
- **Initial Model:**
  - Random architecture search
  - 1 Conv2D layer into 1 Dense layer
  - Randomize:
    - Kernel size
    - Filters
    - Hidden units

Architecture Search: NASH
- **Update Best Model**
- **Iterations of Search**
- **Train Error**
- **Test Error**

Results
- We used an architecture search to develop a CNN architecture that is trained on high quality, low quality pairs of data.
- Our model can reliably convert noisy data into high quality data.
- An example is to the right, where the subsampled peak is shifted right, and is low score. The model shifts it left, and boosts the peak score.

Discussion
- Our model is able to map low quality to high quality ATAC-seq data using a CNNS developed through an Architecture Search.
- The results achieved are good, which is expected given the role of the Architecture Search to build a network with the best possible performance.
- Our computationally simulated noise, done via subsampling, is a reasonable estimate of ATAC-seq noise caused by environmental factors like amount of input DNA and sequencing depth, so our model should do a good job of denoising true low-quality data.

Future Work
- Develop pipeline for cell-line specific training.
- Confirm biological relevance of denoised data using properties of chromatin accessibility for evaluation.
- Replicate work with other sources of noise.

References