A deep learning based approach for genetic risk prediction

Raquel Dias, PhD.
Senior Staff Scientist
Scripps Research Translational Institute
raqueld@scripps.edu, @RaquelDiasSRTI
Ali Torkamani, PhD.
atorkama@scripps.edu, @ATorkamani

Scripps Research
Science Changing Life

Whole Genome Sequencing vs. Genotype array

Full Data
(whole genome sequencing)

```
0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 1 0 1 1 1 0 1 1 1 1 0 1 1 0 1
```

Sparse Data
(genotype array)

```
0 0 ? ? ? 0 0 1 1 ? ? 0 ? ? ? ?
0 0 ? ? ? 0 0 1 1 ? ? 0 ? ? ? ?
0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?
0 0 ? ? ? 0 0 1 1 ? ? 1 ? ? ? ?
0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?
0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?
```

Scripps Research
Translational Institute

Whole Genome Sequencing vs. Genotype array

Full Data
~80M genetic variants

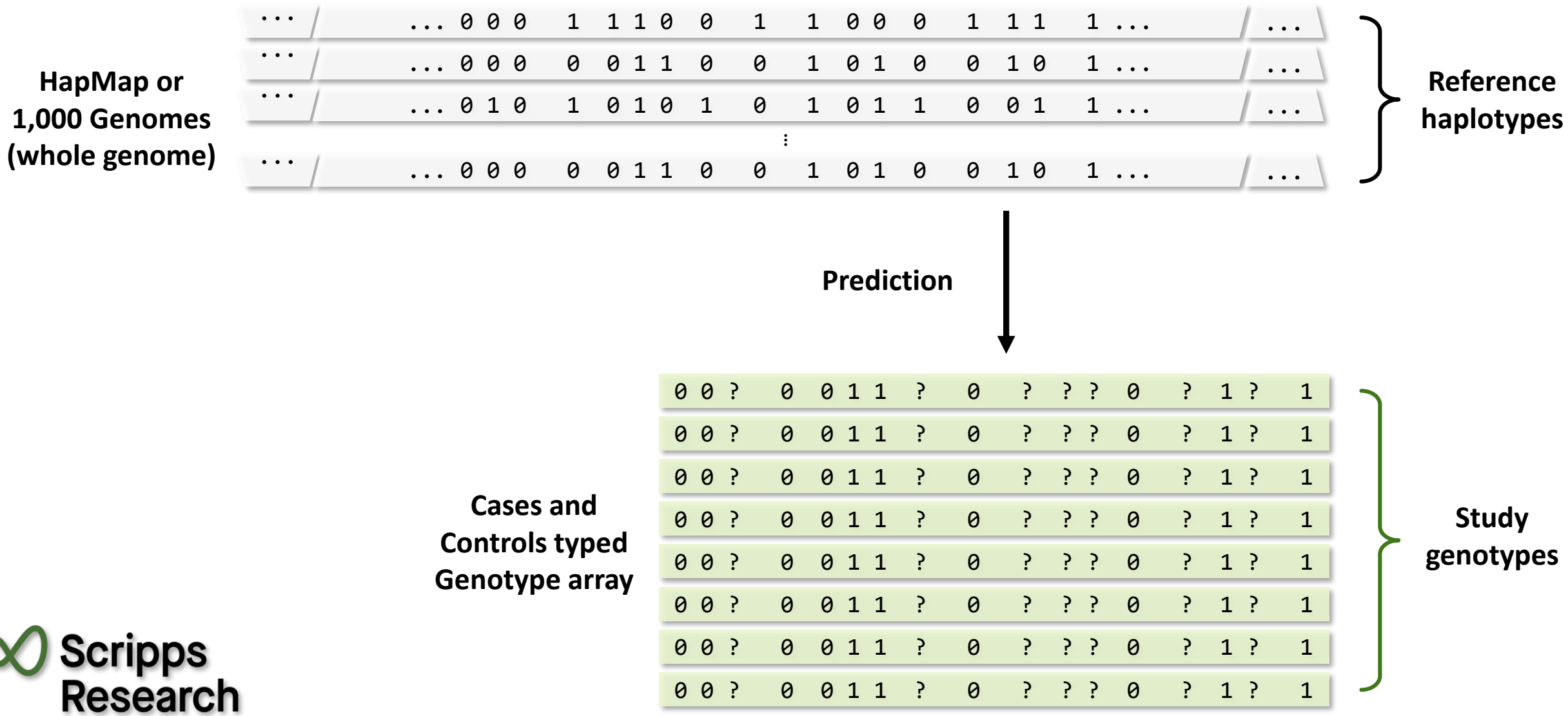
```
0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 1 0 1 1 1 0 1 1 1 1 0 1 1 0 1
```

Sparse Data
~4 million genetic

```
0 0 ? ? ? 0 0 1 1 ? ? 0 ? ? ? ?
0 0 ? ? ? 0 0 1 1 ? ? 0 ? ? ? ?
0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?
0 0 ? ? ? 0 0 1 1 ? ? 1 ? ? ? ?
0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?
0 1 ? ? ? 1 0 1 1 ? ? 0 ? ? ? ?
```

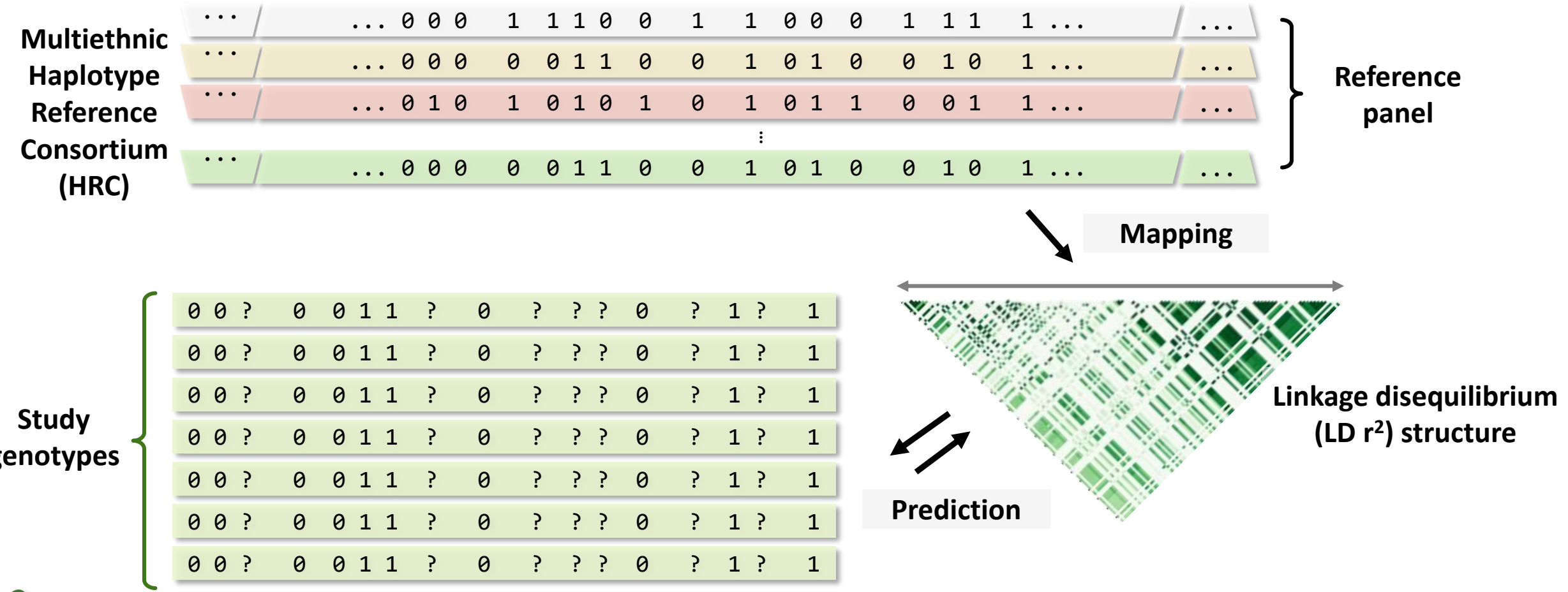
Scripps Research
Translational Institute

Genetic imputation problem



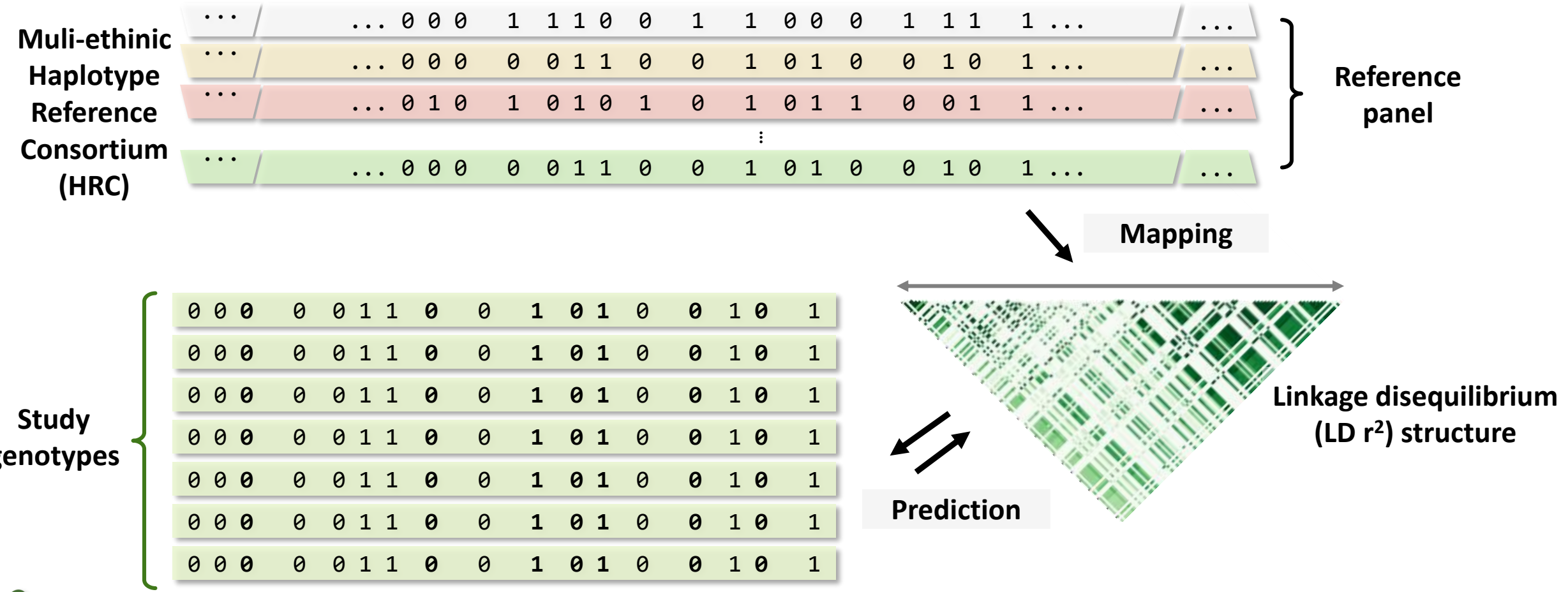
Scripps Research
Translational Institute

A typical imputation approach



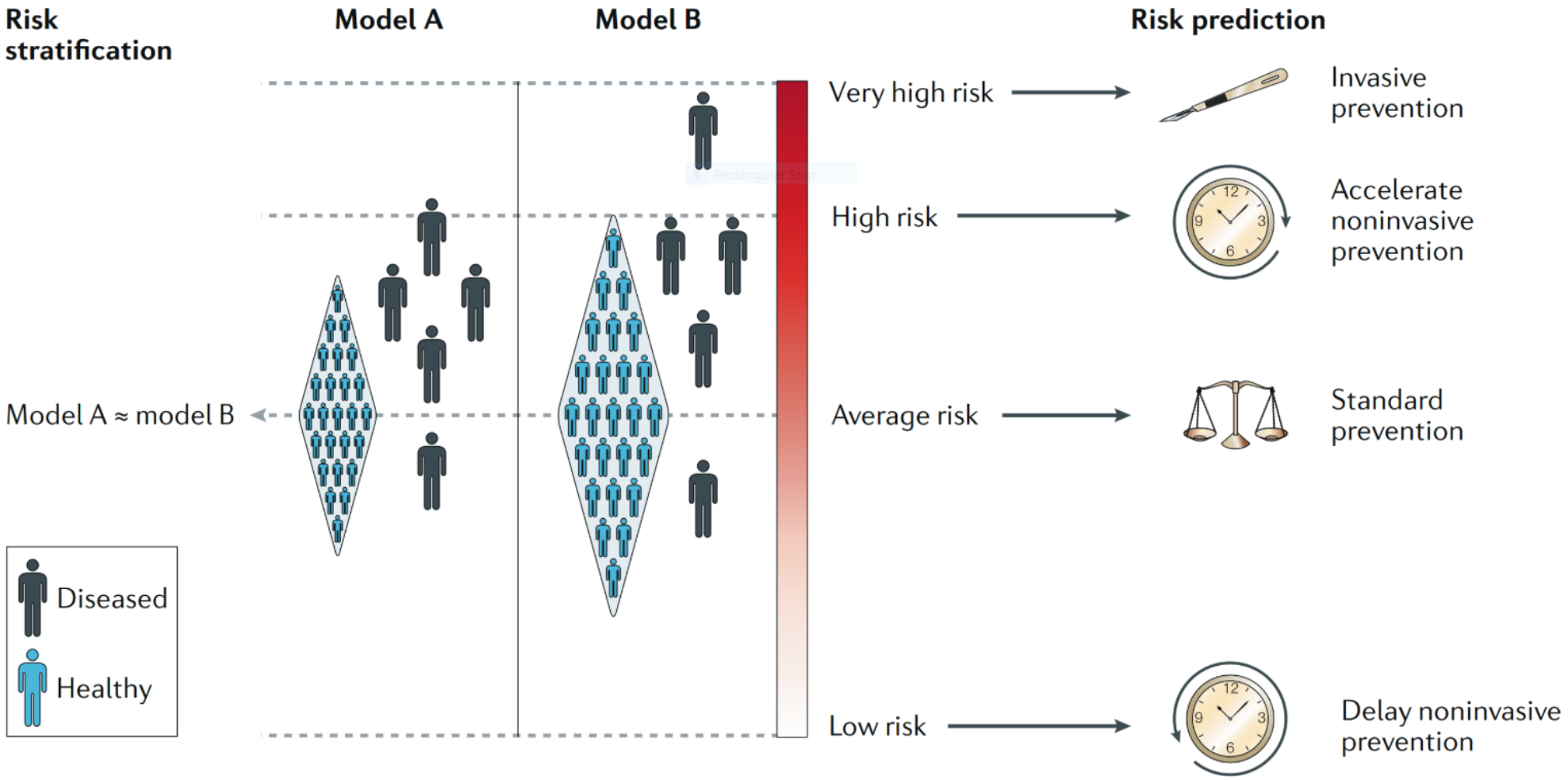
Scripps Research
Translational Institute

A typical imputation approach



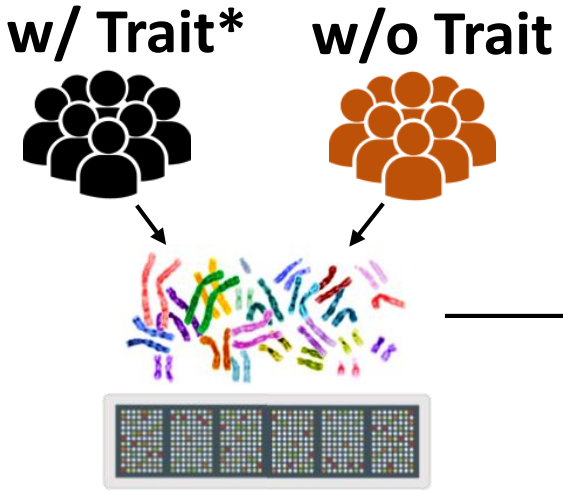
Scripps Research
Translational Institute

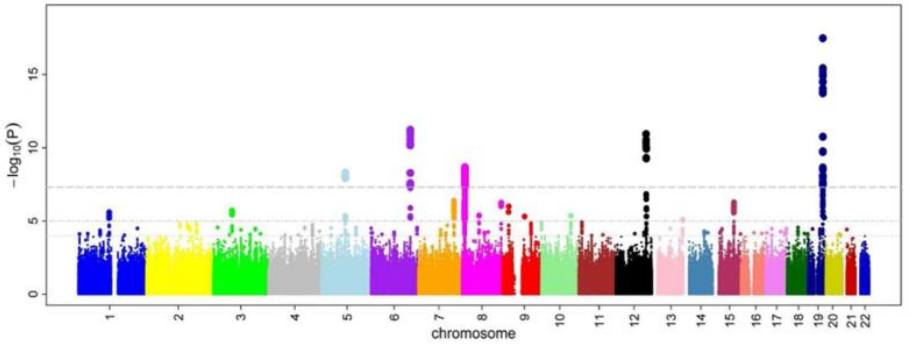
Polygenic Risk Score (PRS)

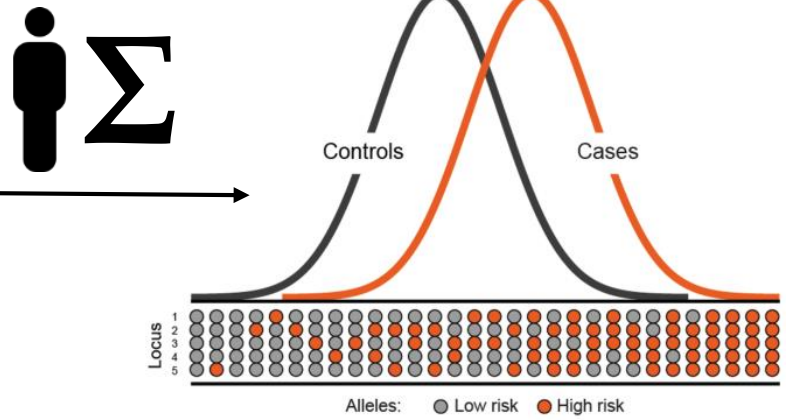


Scripps Research
Translational Institute

Polygenic Risk Calculation







Design
100,000+ subjects

Results
Millions of known variants

Polygenic Risk Score
Cumulative sum

***Trait** can often be heterogeneous
e.g. coronary artery = heart attack, stroke, bypass surgery, etc.

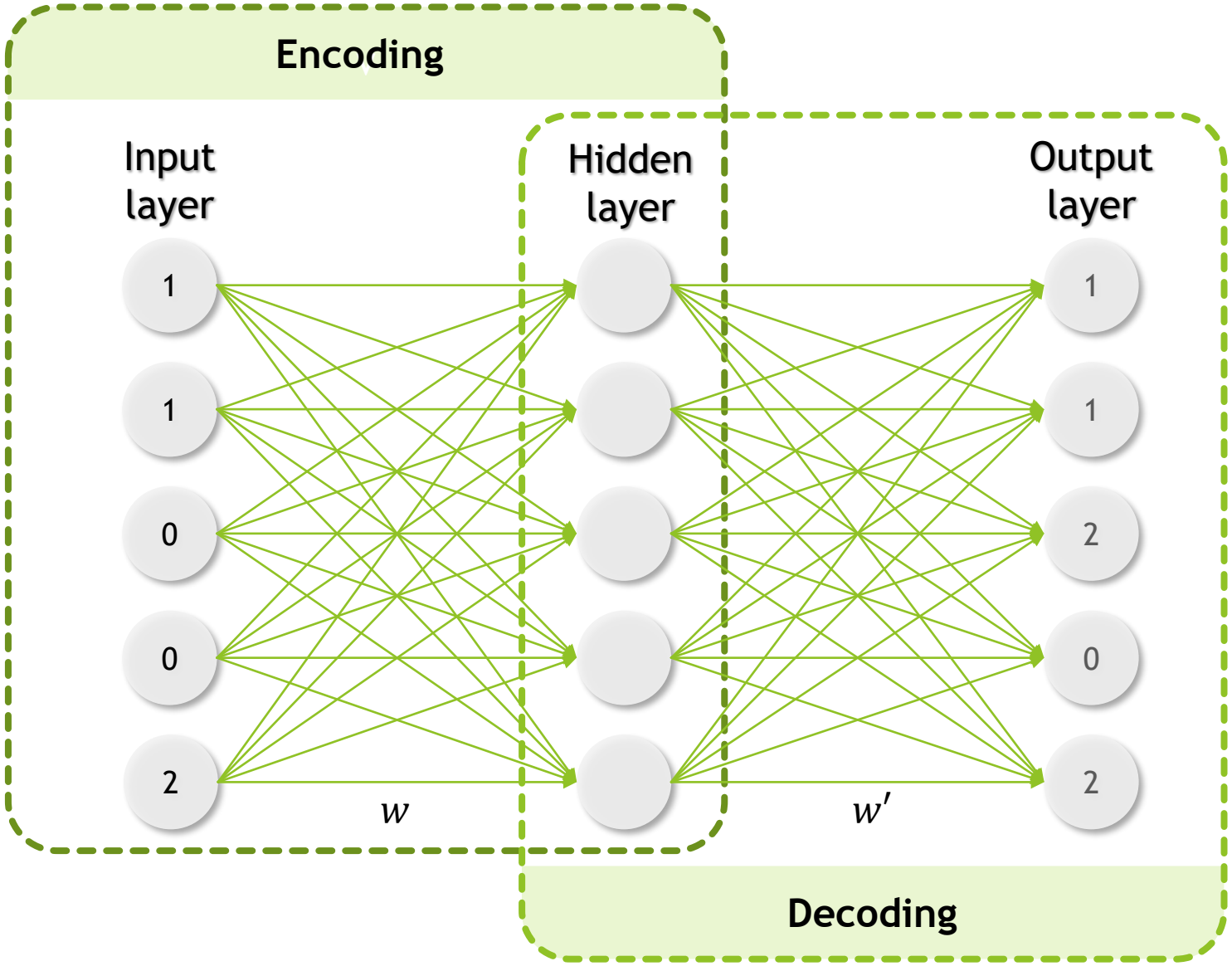
Scripps Research
Translational Institute

Objectives

1. More accurate and faster imputation
2. Find important genetic variants
3. Better polygenic risk score calculation

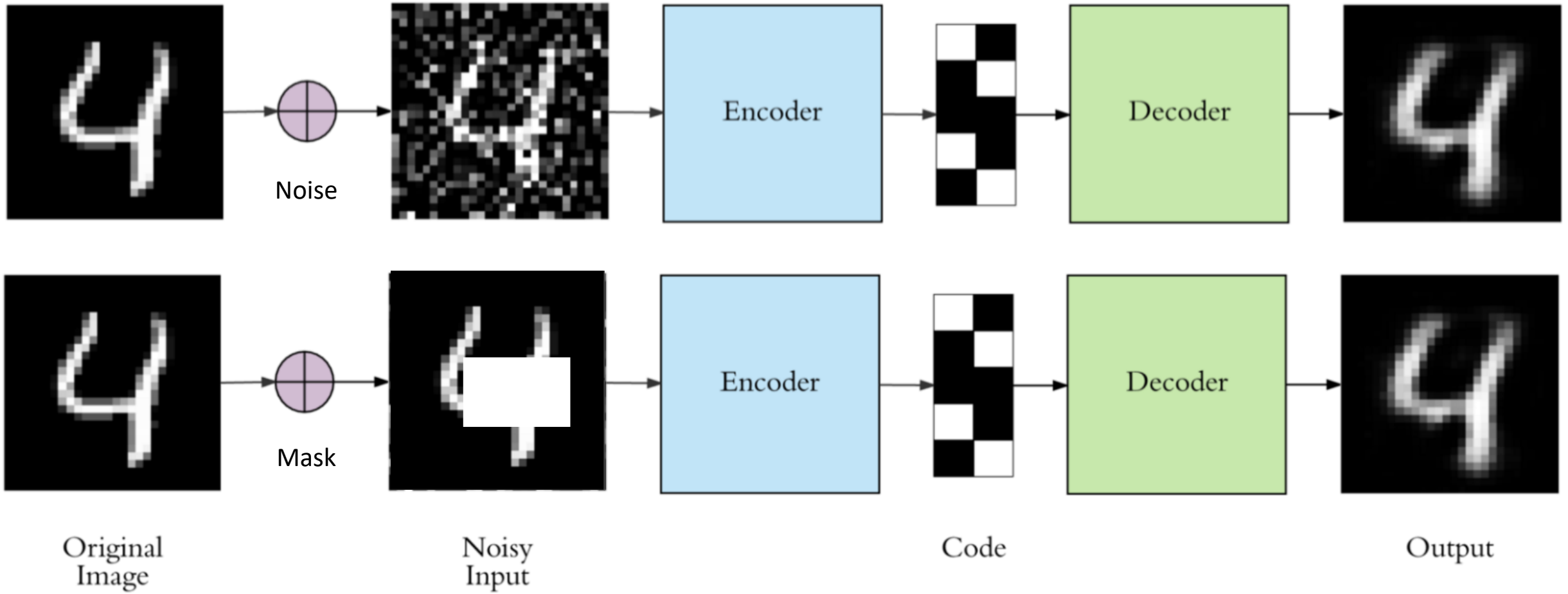
Scripps Research
Translational Institute

Our proposed approach



Scripps Research
Translational Institute

Denoising autoencoder for image restoration



Scripps Research
Translational Institute

Bigdeli, Siavash Arjomand, and Matthias Zwicker. "Image restoration using autoencoding priors." *arXiv preprint arXiv:1703.09964* (2017).

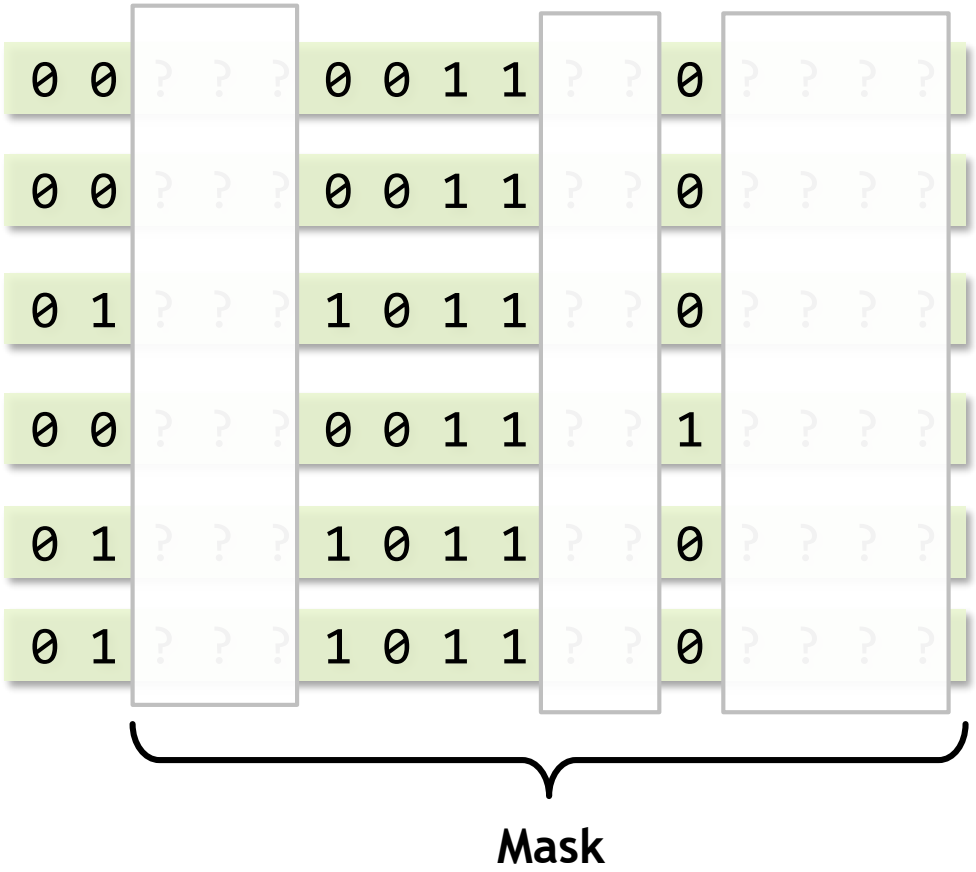
Wang, Ruxin, and Dacheng Tao. "Non-local auto-encoder with collaborative stabilization for image restoration." *IEEE Transactions on Image Processing* 25.5 (2016): 2117-2129.

Genotype imputation case study example

Ground truth
(whole genome sequencing)

```
0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 1 1 0 1 1 0 1
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0
0 1 0 1 0 1 0 1 1 1 1 0 1 1 0 1
0 1 0 1 1 1 0 1 1 1 1 0 1 1 0 1
```

Masked input
(genotype array)



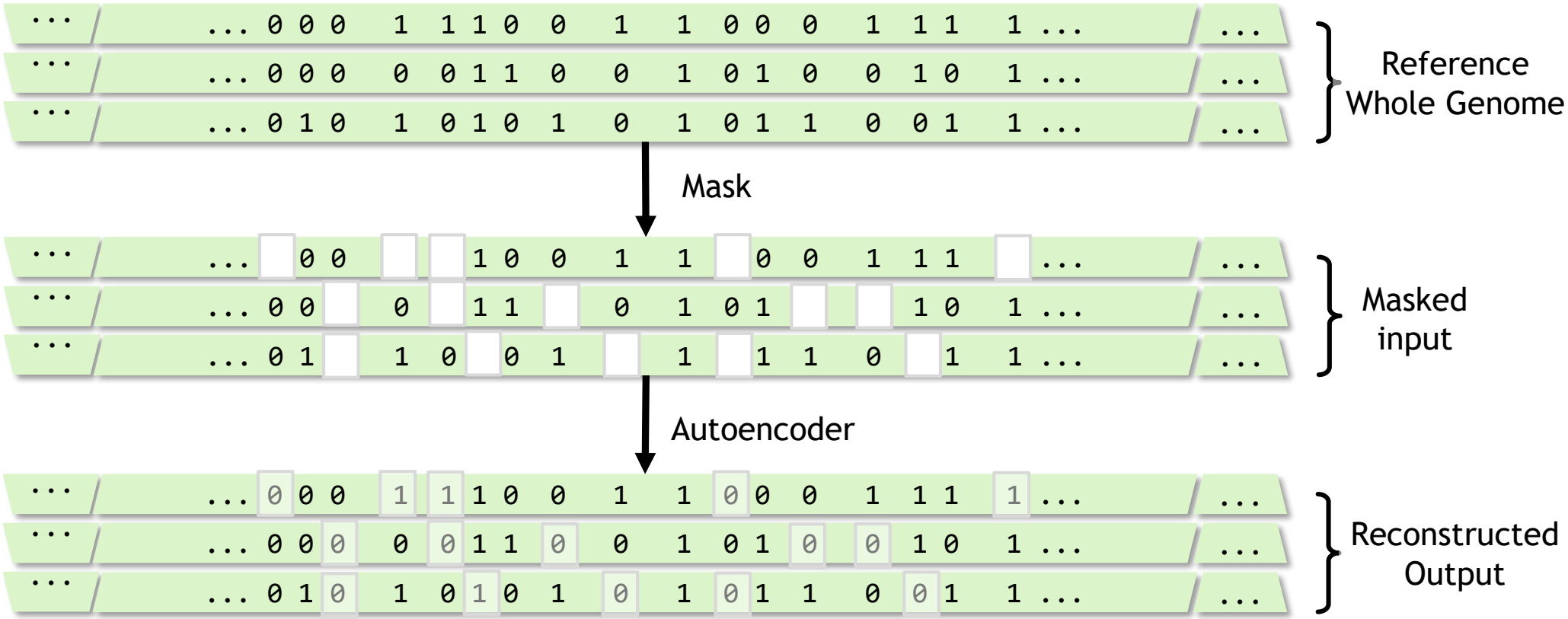
Scripps Research
Translational Institute

Case study: 9p21.3 region of the genome

- **Length: 59846 bp**
- **846 genetic variants in reference panel (whole genome data)**
  - Approx. 200 common variants
  - Approx. 600 rare variants
- **Only 17-47 variants in genotype array!!!**
- **Strong association to coronary artery disease (CAD)**
- **Genotyped and sequenced in many studies**

Scripps Research
Translational Institute

Training on the reference panel: Data augmentation strategy



Scripps Research
Translational Institute

Customized Sparsity Loss Function

- Sparsity loss with Kullback-Leibler (KL) / cross entropy element:

$$D_{KL}(\rho||\hat{\rho}) = \rho * \log\left(\frac{\rho}{\hat{\rho}}\right) + (1-\rho) * \log\left(\frac{1-\rho}{1-\hat{\rho}}\right)$$

- Customized loss adjusted for hidden activation sparsity:

$$loss = MSE + \beta * \sum_{i=1}^{n} D_{KL(i)}$$

- Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

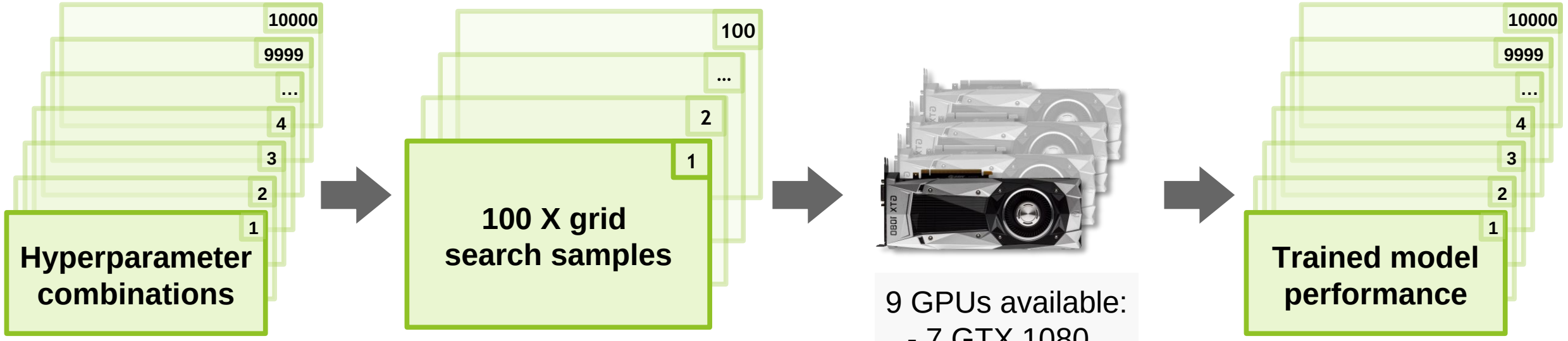
Scripps Research
Translational Institute

Hyper parameters to be optimized

- β
- ρ
- Activation functions
- L1/L2 regularizers
- Learning rate
- Batch size

Scripps Research
Translational Institute

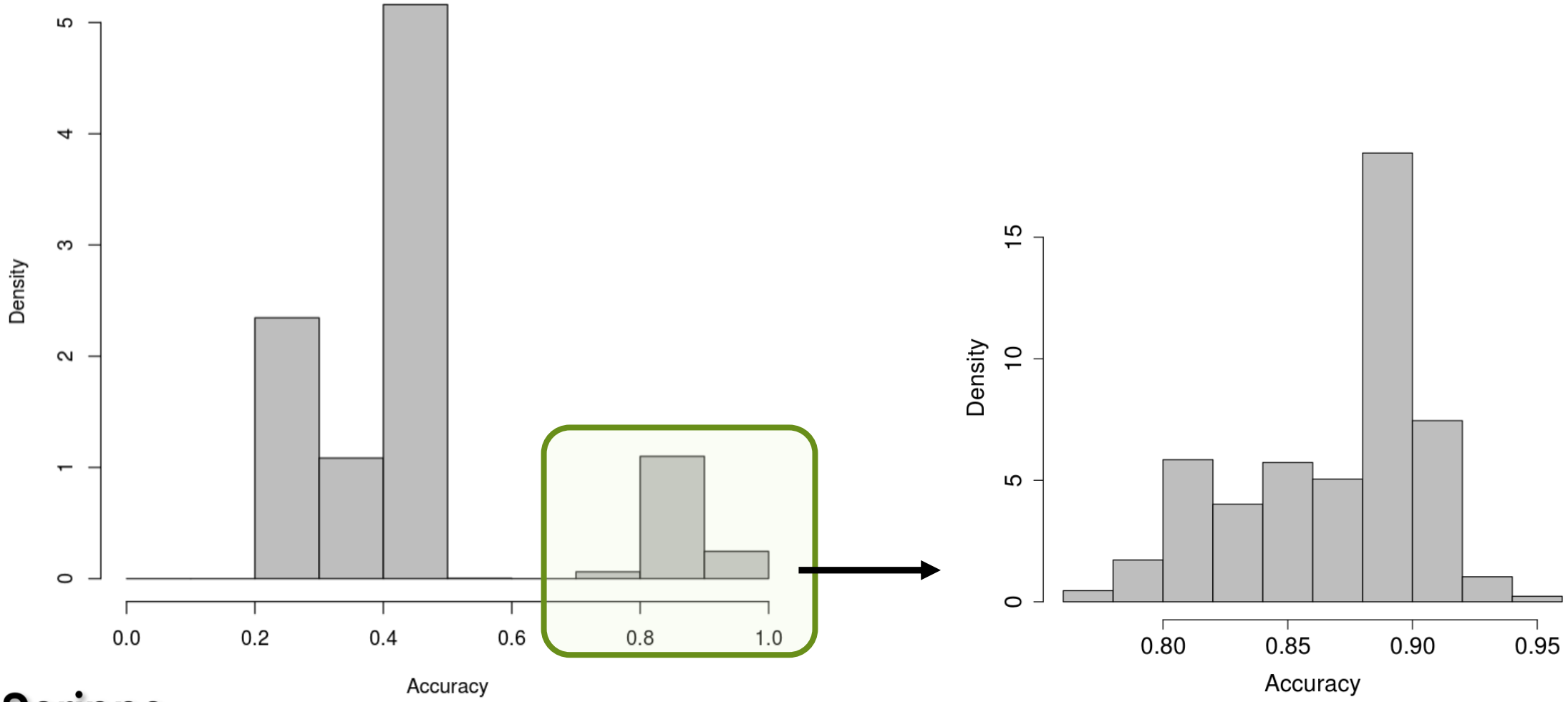
Parallel Grid Search
Hyperparameter optimization approach



860 hours
(sequential run,
100 epochs)

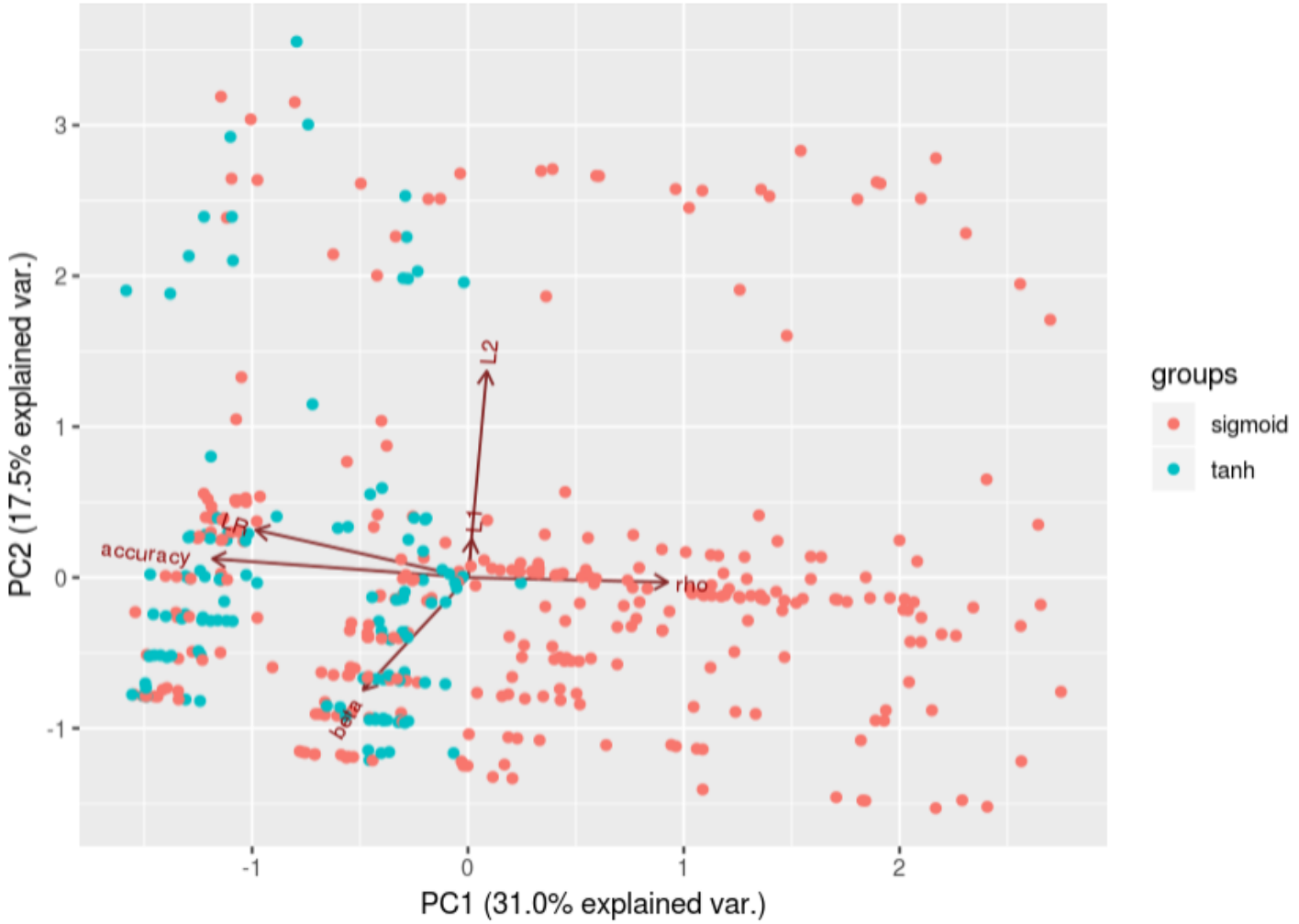
Scripps Research
Translational Institute

Grid Search Results: training accuracy



Scripps Research
Translational Institute

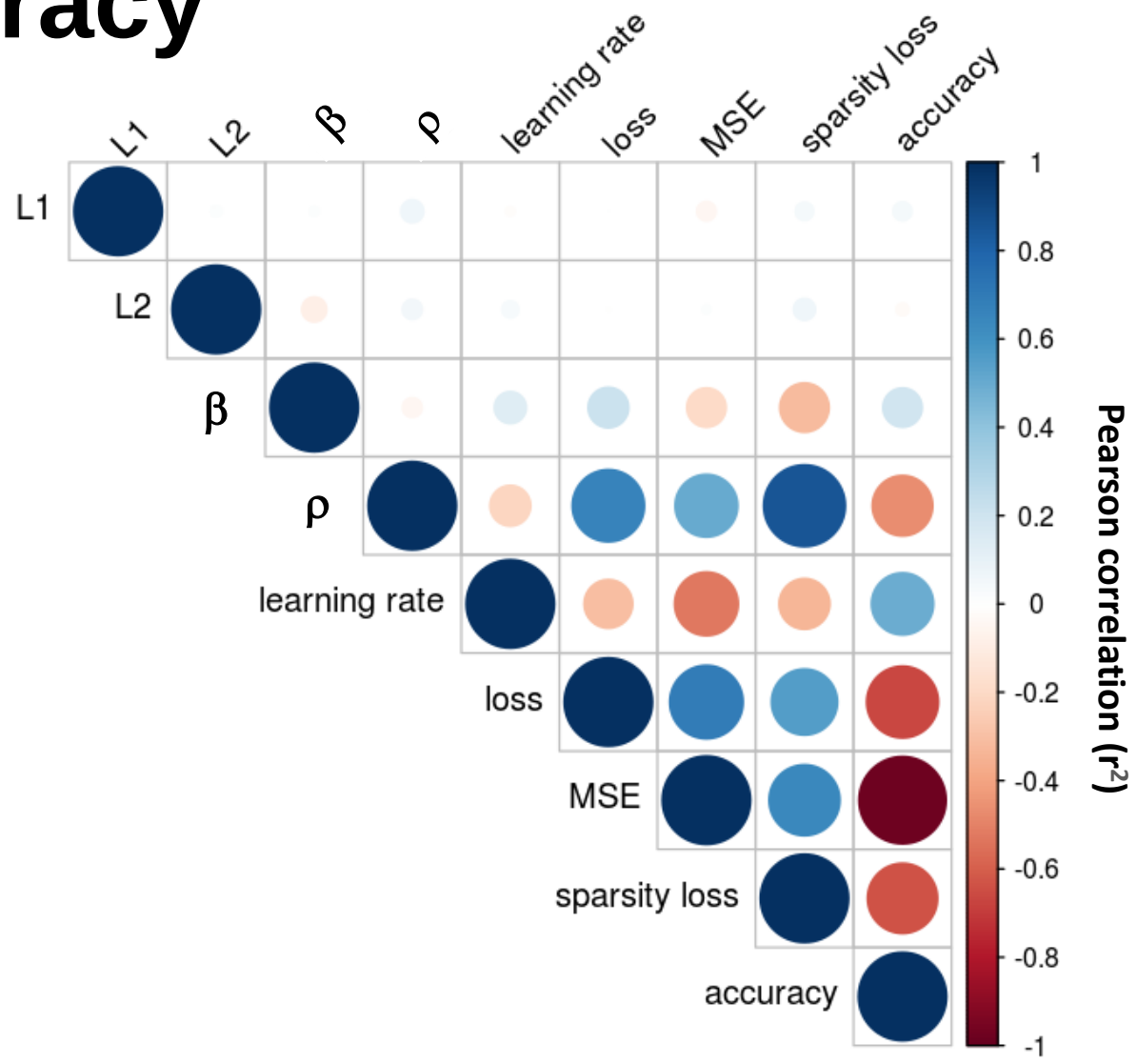
Grid Search results: assessing best hyperparameter values



Scripps Research
Translational Institute

Effect of hyper parameter values in training accuracy

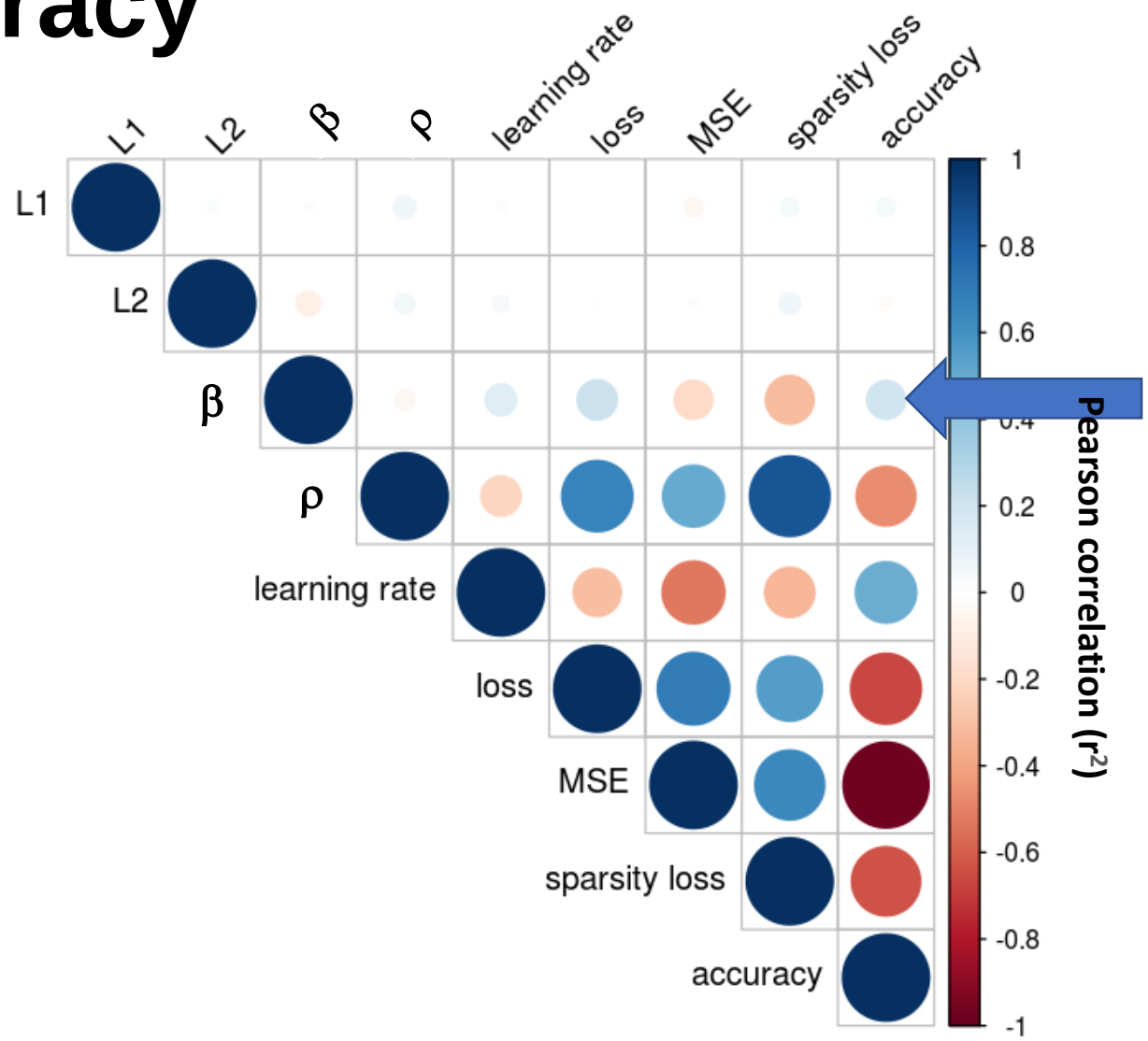
- β
- ρ
- Learning rate



Scripps Research
Translational Institute

Effect of hyper parameter values in training accuracy

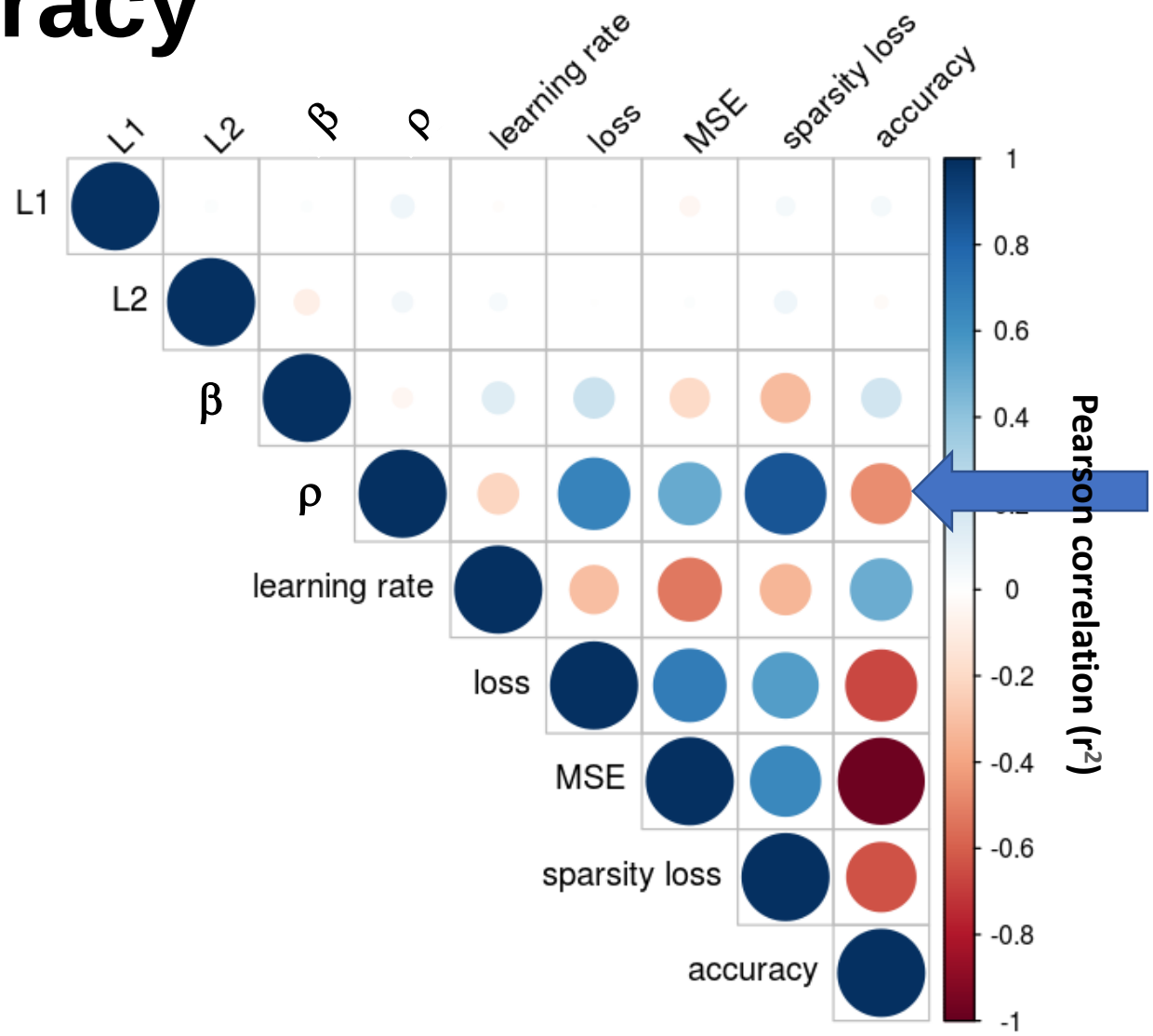
- β
- ρ
- Learning rate



Scripps Research
Translational Institute

Effect of hyper parameter values in training accuracy

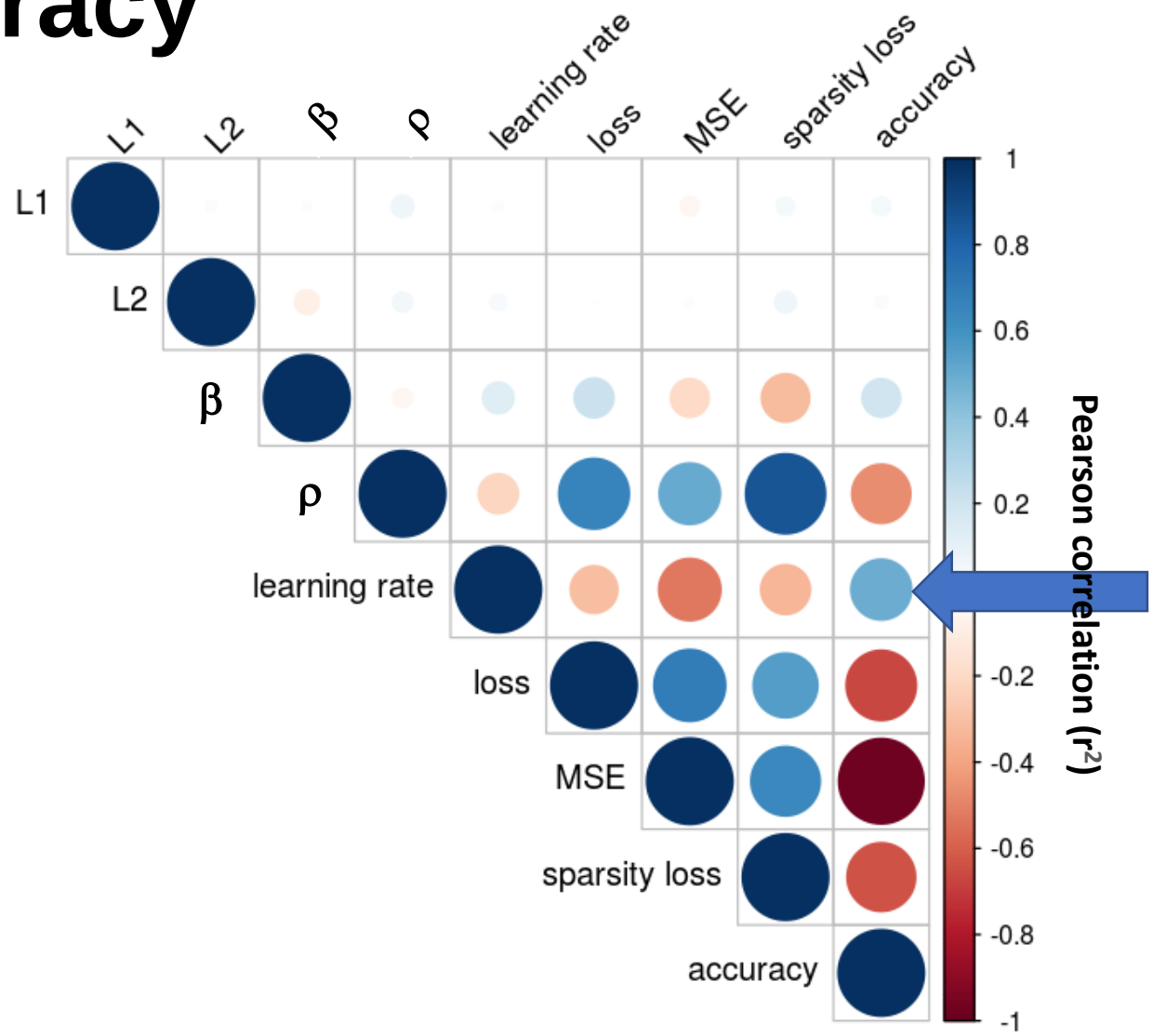
- β
- ρ
- Learning rate



Scripps Research
Translational Institute

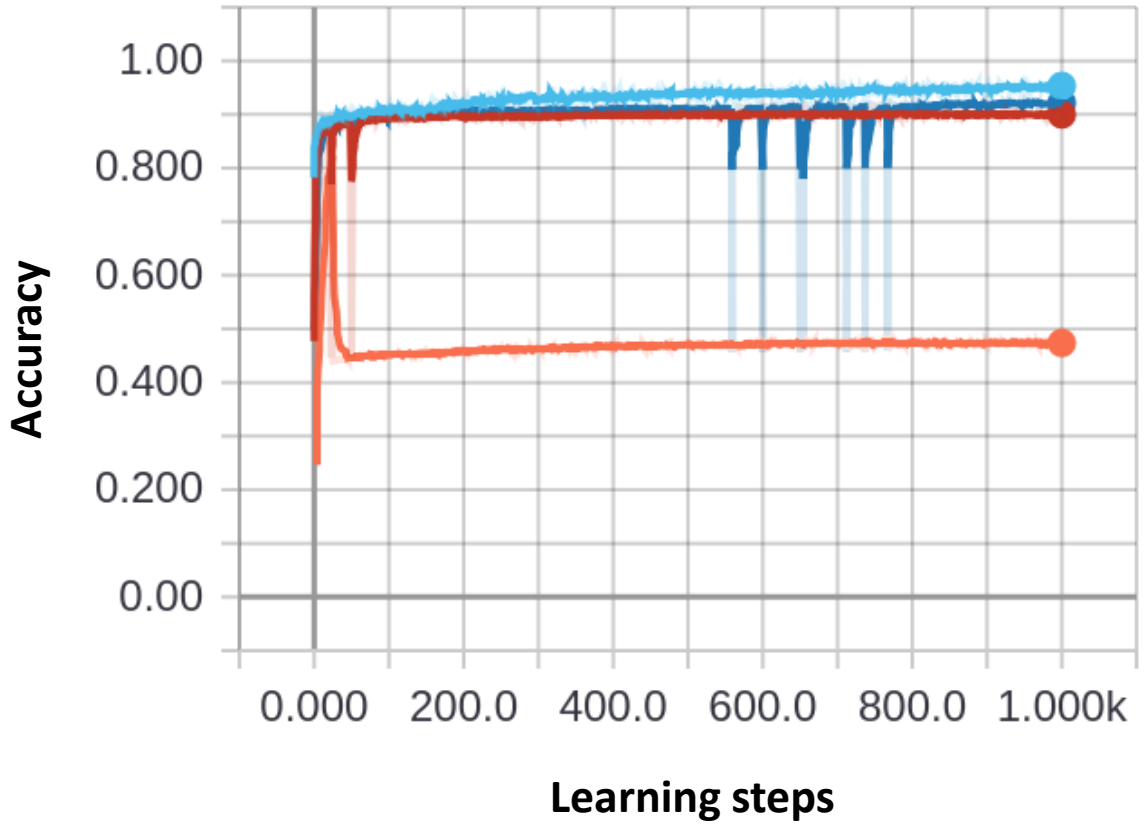
Effect of hyper parameter values in training accuracy

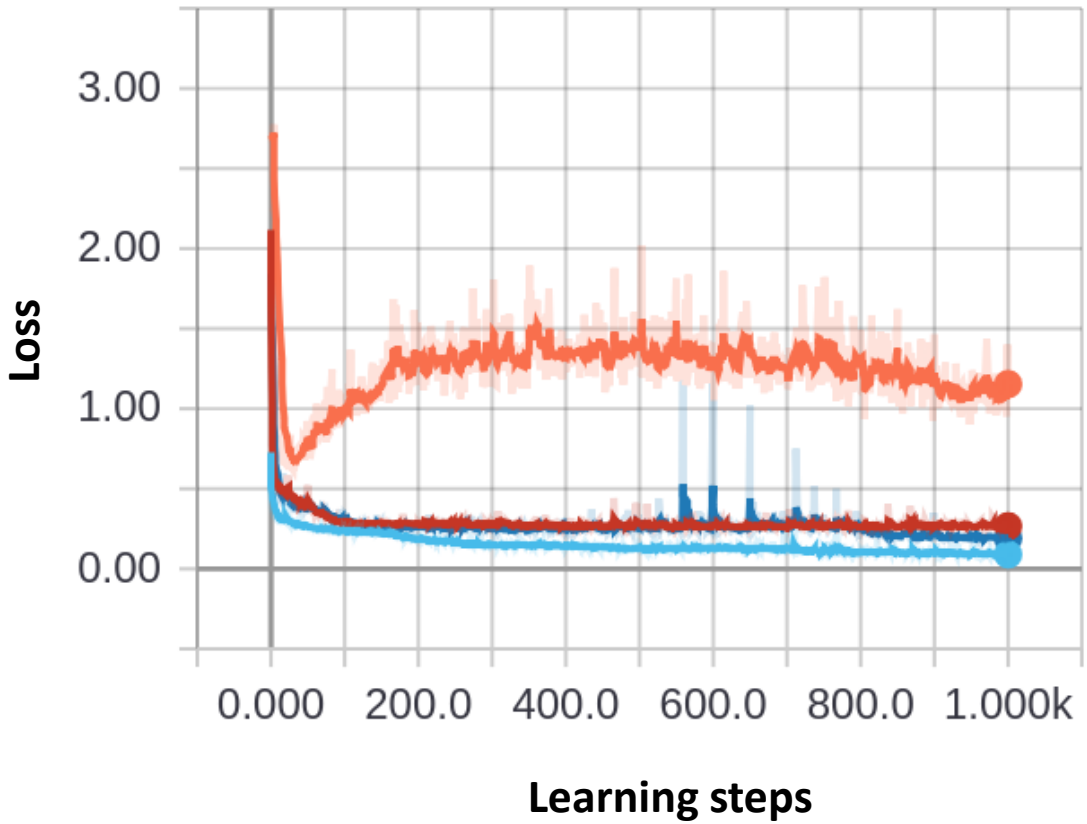
- β
- ρ
- Learning rate



Scripps Research
Translational Institute

Optimizing batch size: training accuracy

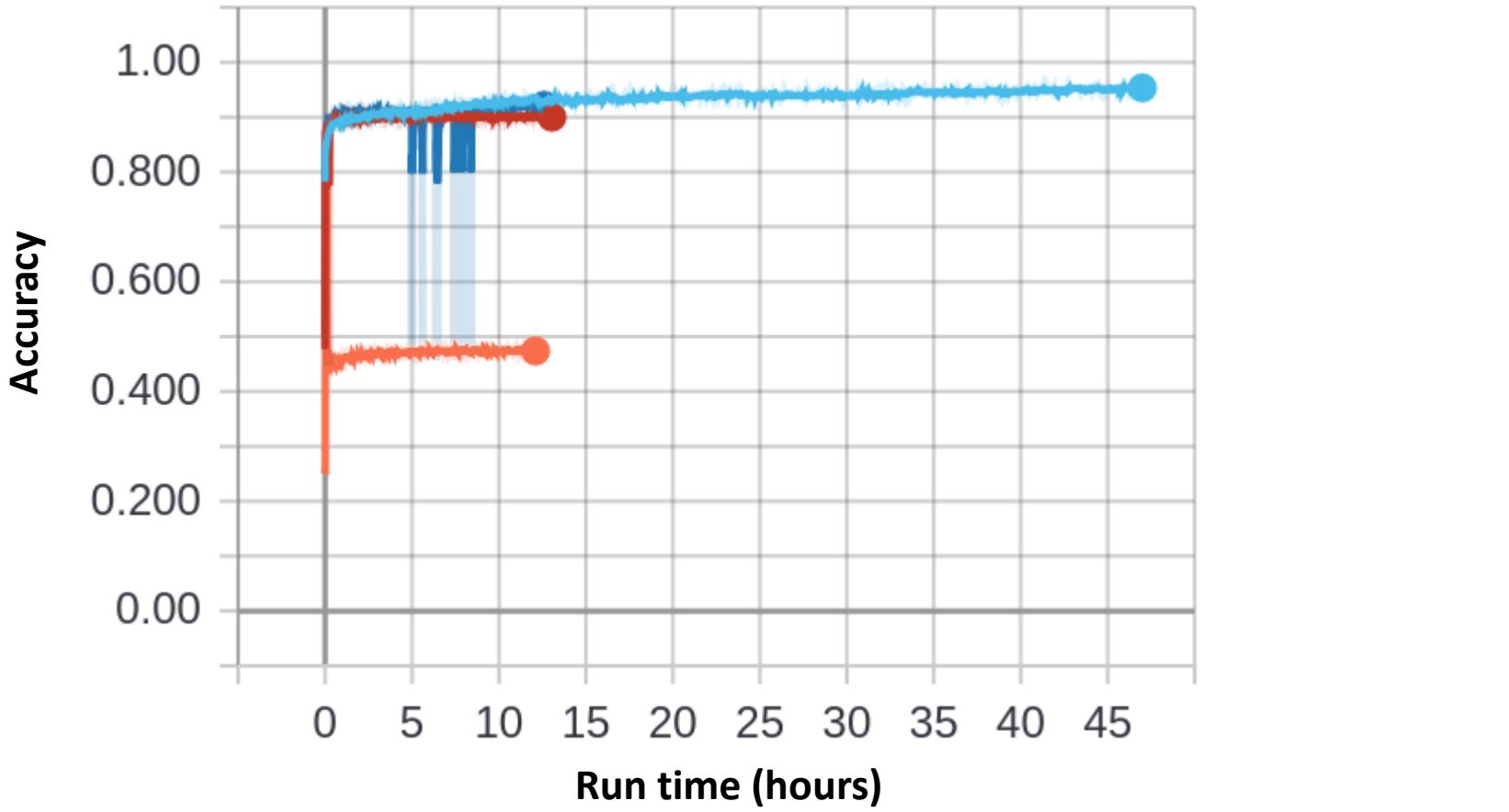




Scripps Research
Translational Institute

10 batches
50 batches
100 batches
1000 batches

Optimizing batch size: training run time



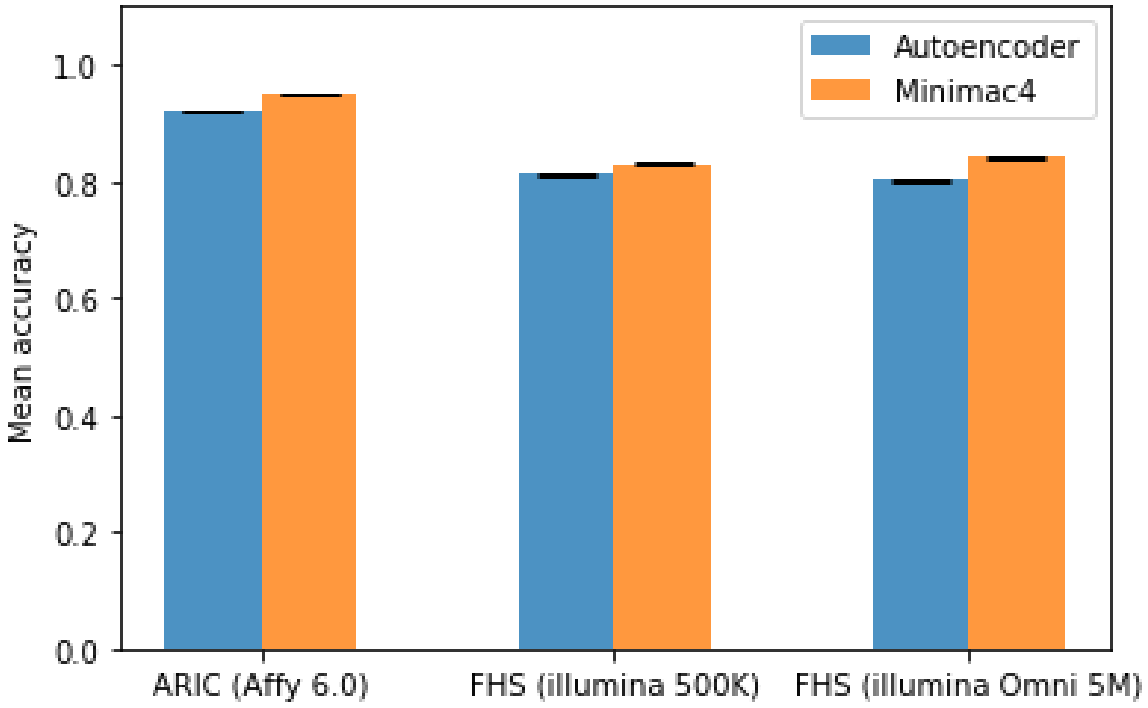
Scripps Research
Translational Institute

Testing on multiple case studies

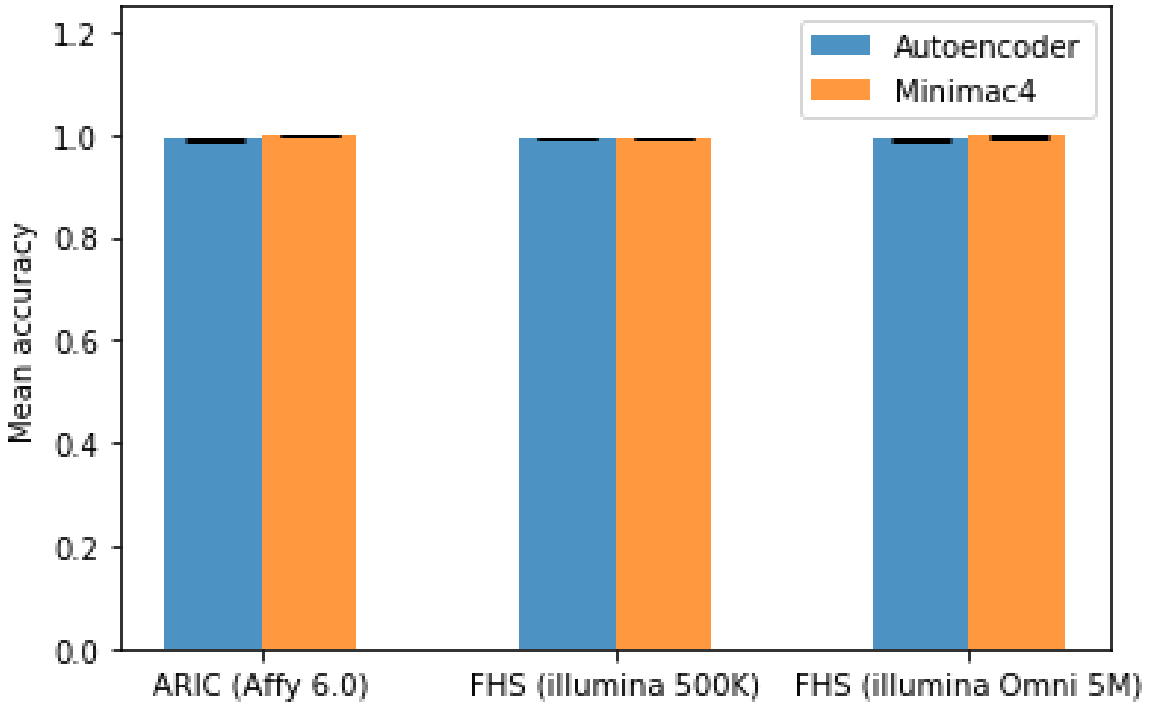
- **Atherosclerosis Risk in Communities (ARIC)**
  - More than 3000 samples
  - Whole genome sequencing (846 variants, 0% mask, ground truth)
  - Affymetrix 6.0 genotype array (17 variants, 98% mask, input data)
- **Framingham Heart Study (FHS)**
  - More than 500 samples
  - Whole genome sequencing (846 variants, 0% mask, ground truth)
  - Illumina 500K genotype array (47 variants, 95% mask, input data)
  - Illumina 5M (93 variants, 89% mask, input data)

Accuracy in additional case studies: Proposed approach versus common statistic methodology

Performance: all variants

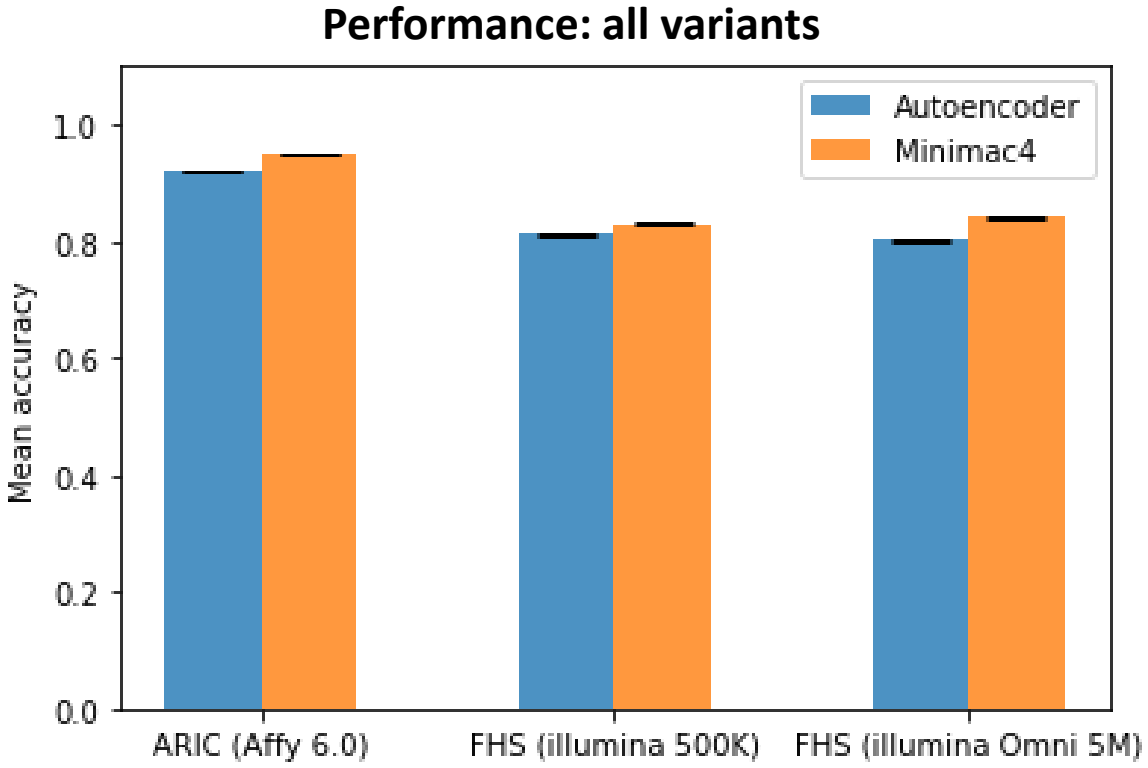


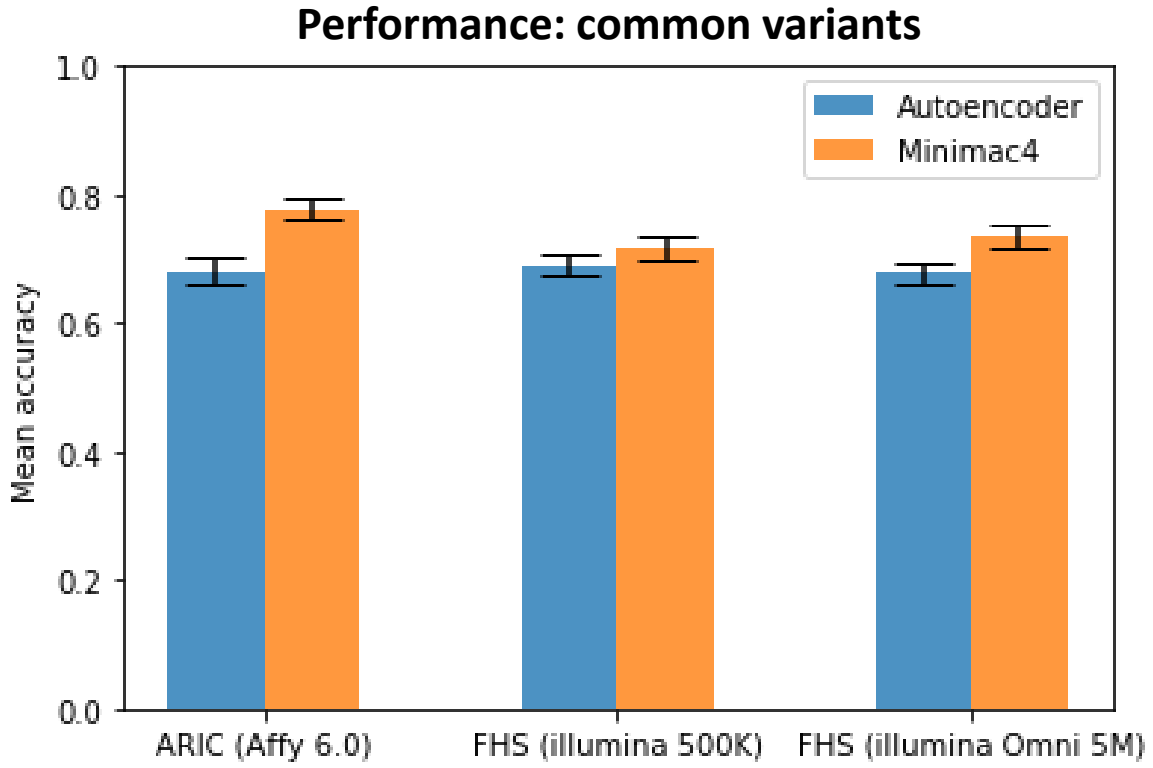
Performance: rare variants



Scripps Research
Translational Institute

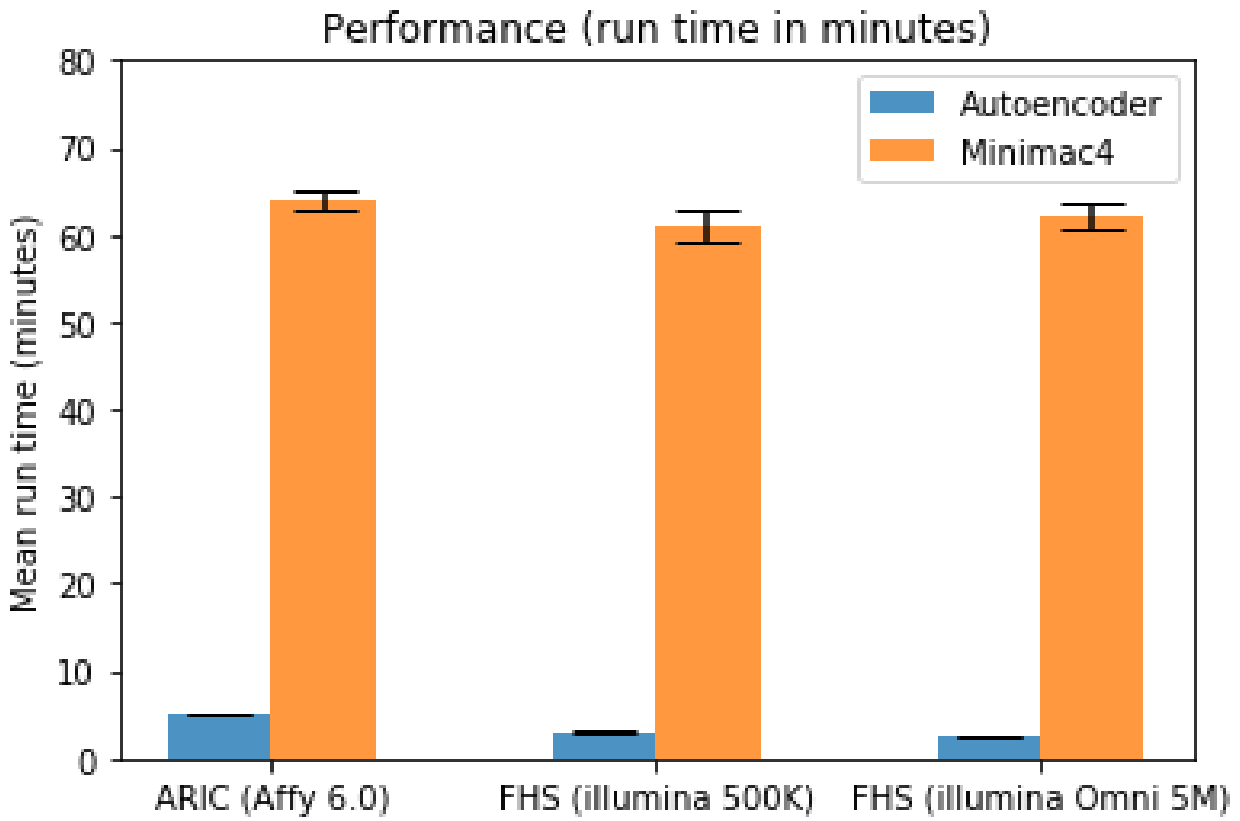
Accuracy in additional case studies: Proposed approach versus common statistic methodology





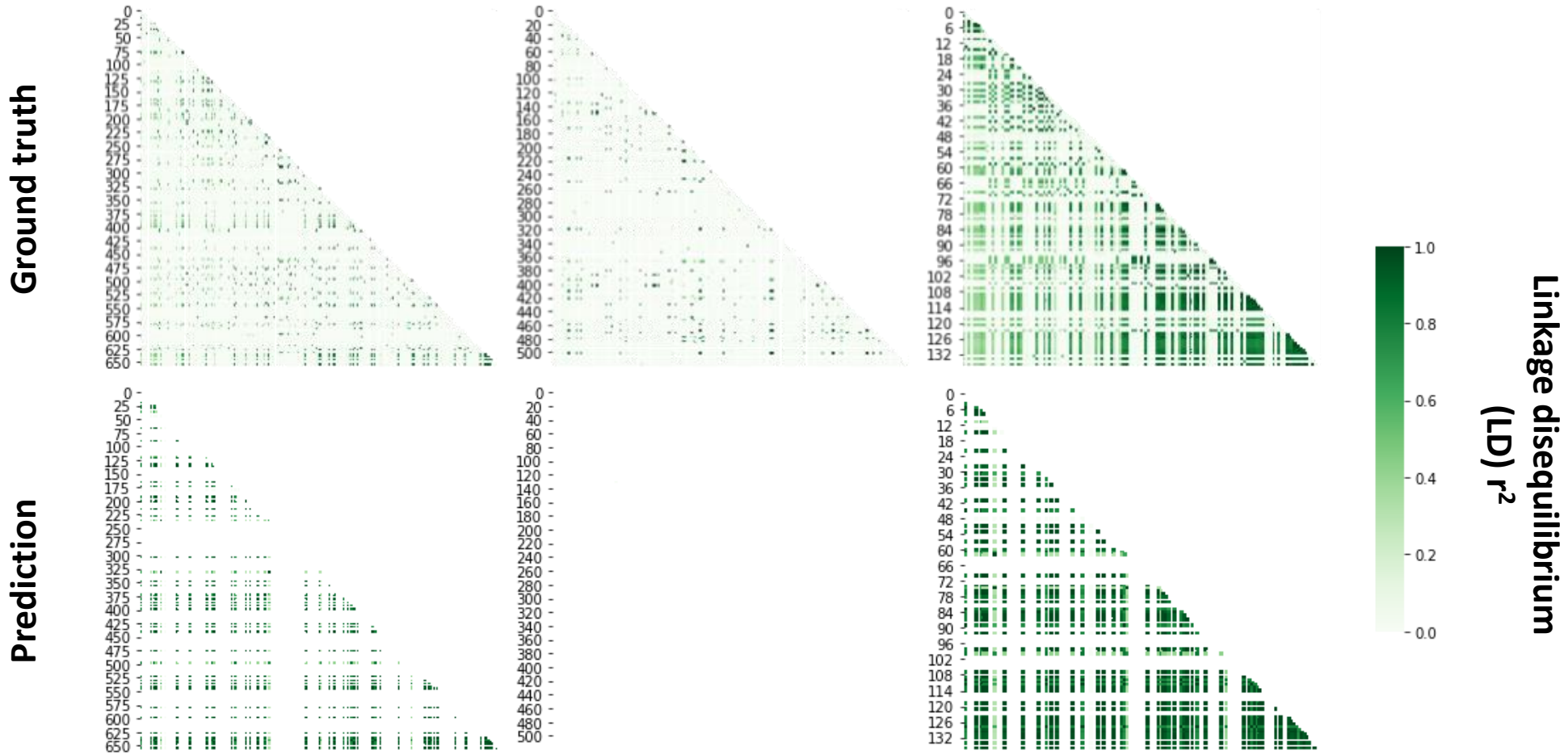
Scripps Research
Translational Institute

Run time: Proposed approach versus common statistic methodology



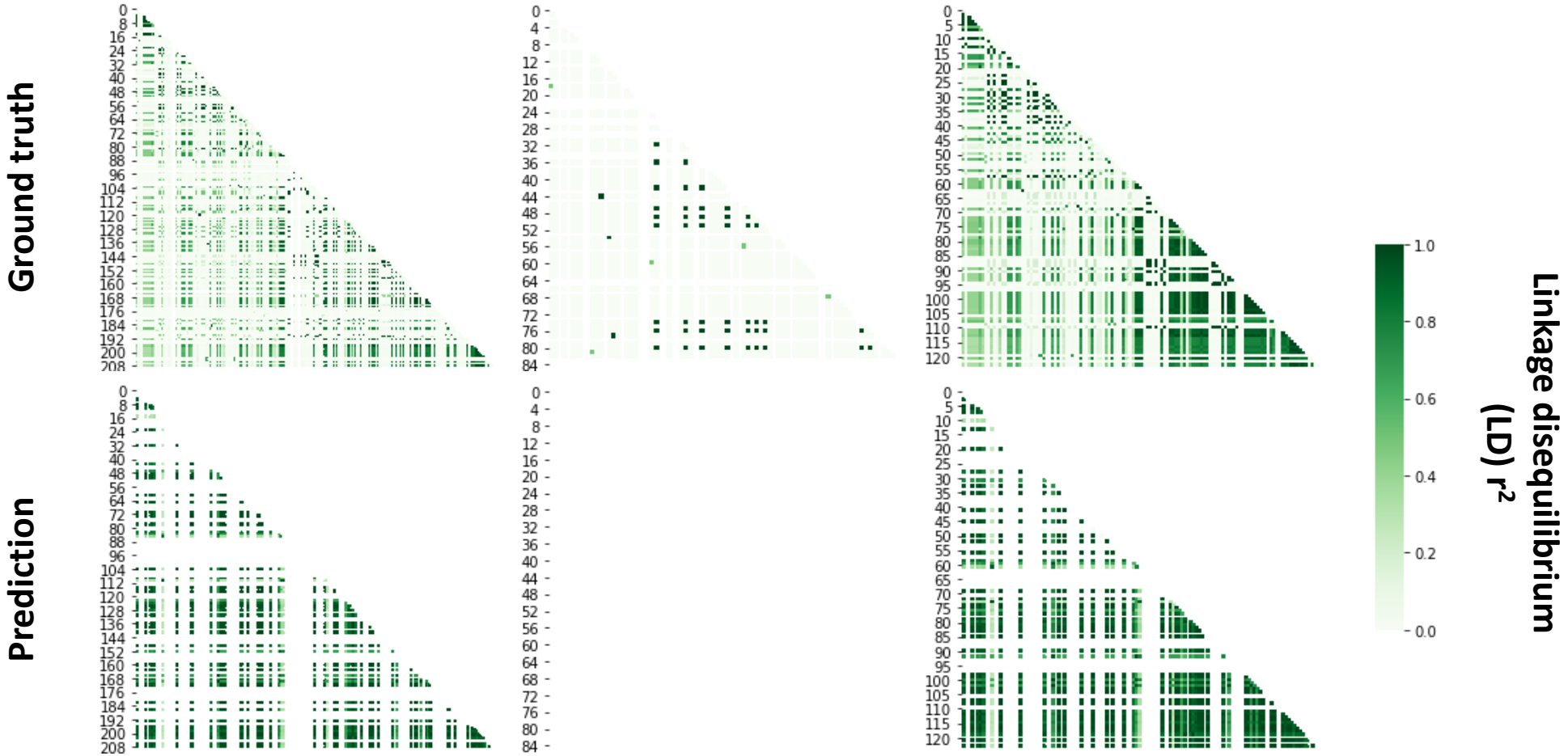
Scripps Research
Translational Institute

Linkage disequilibrium structure: ARIC



Scripps Research
Translational Institute

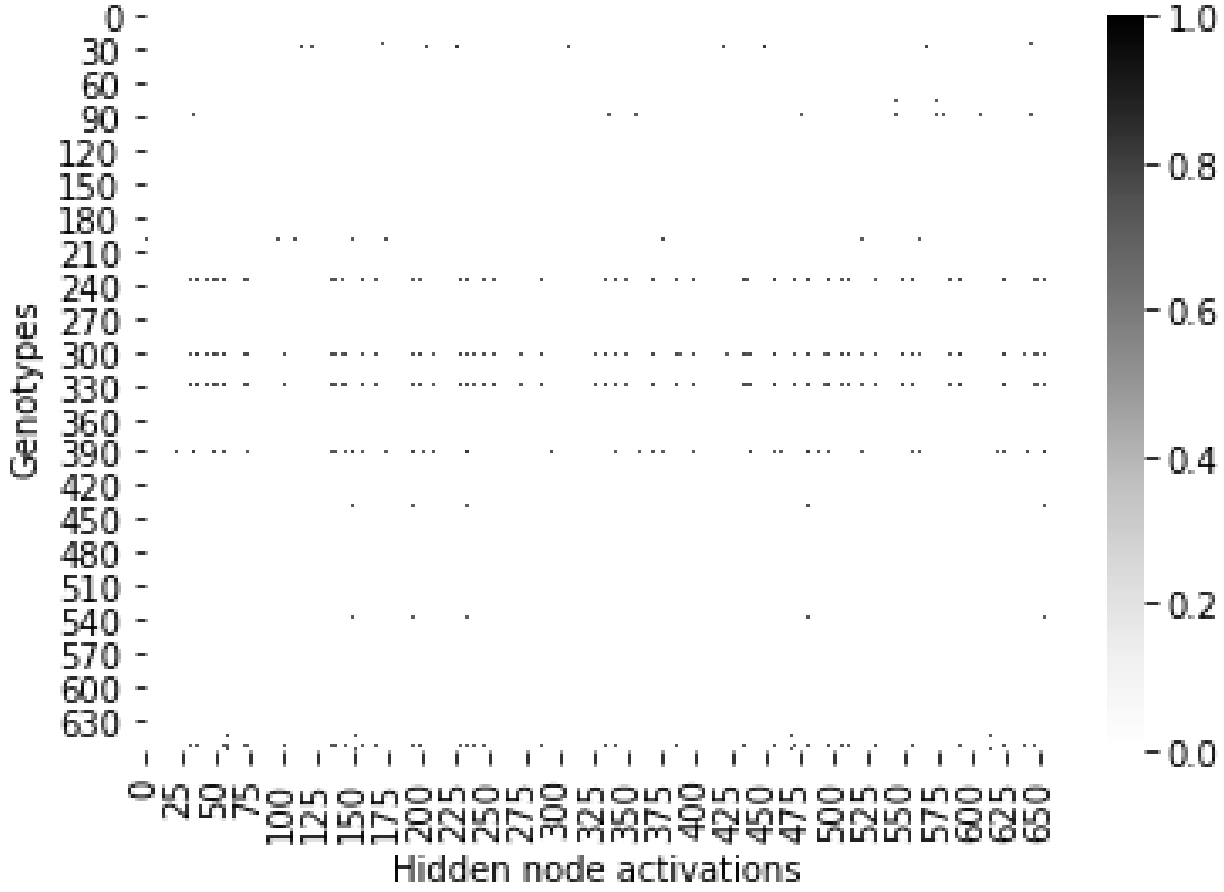
Linkage disequilibrium structure: FHS



Scripps Research
Translational Institute

Interpretability: identifying representative genetic variants

Maximal information criteria



Scripps Research
Translational Institute

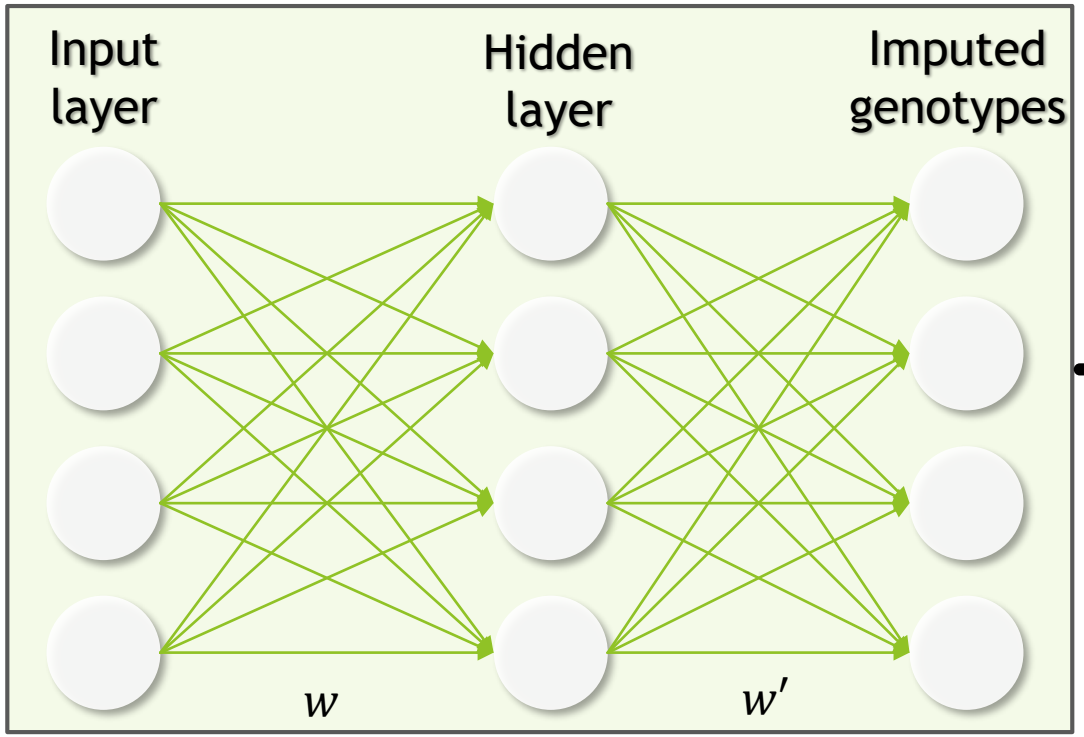
Conclusions

- **Grid search** was able to find high accuracy models (>0.90)
- **Hyperparameters** played an important role in training performance
- Reconstruction of genetic variants from very sparse data with **high accuracy** (>0.80)
- Superior computational performance, **faster predictions**
- **Fine parameter tuning** may be necessary

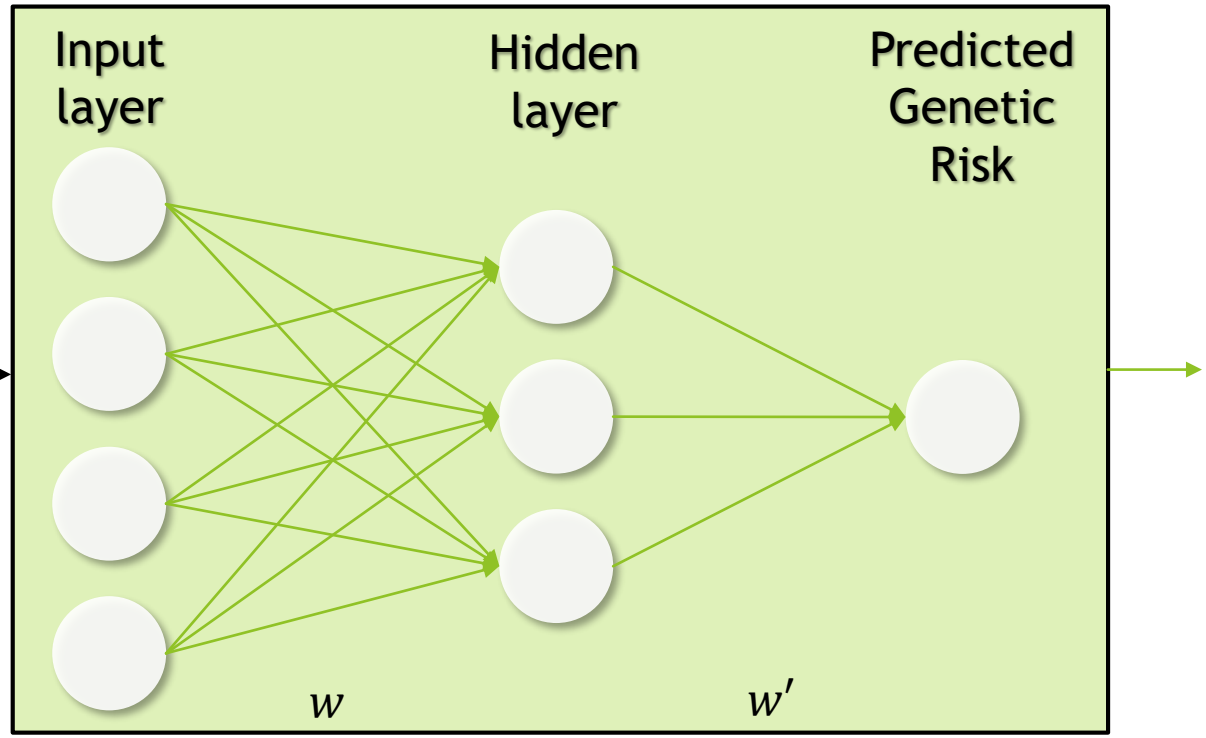
Scripps Research
Translational Institute

Future steps

- **Expand to other genomic regions, fine parameter tuning**
- **Use imputation autoencoder results as input for polygenic risk score calculation**



Imputation Autoencoder

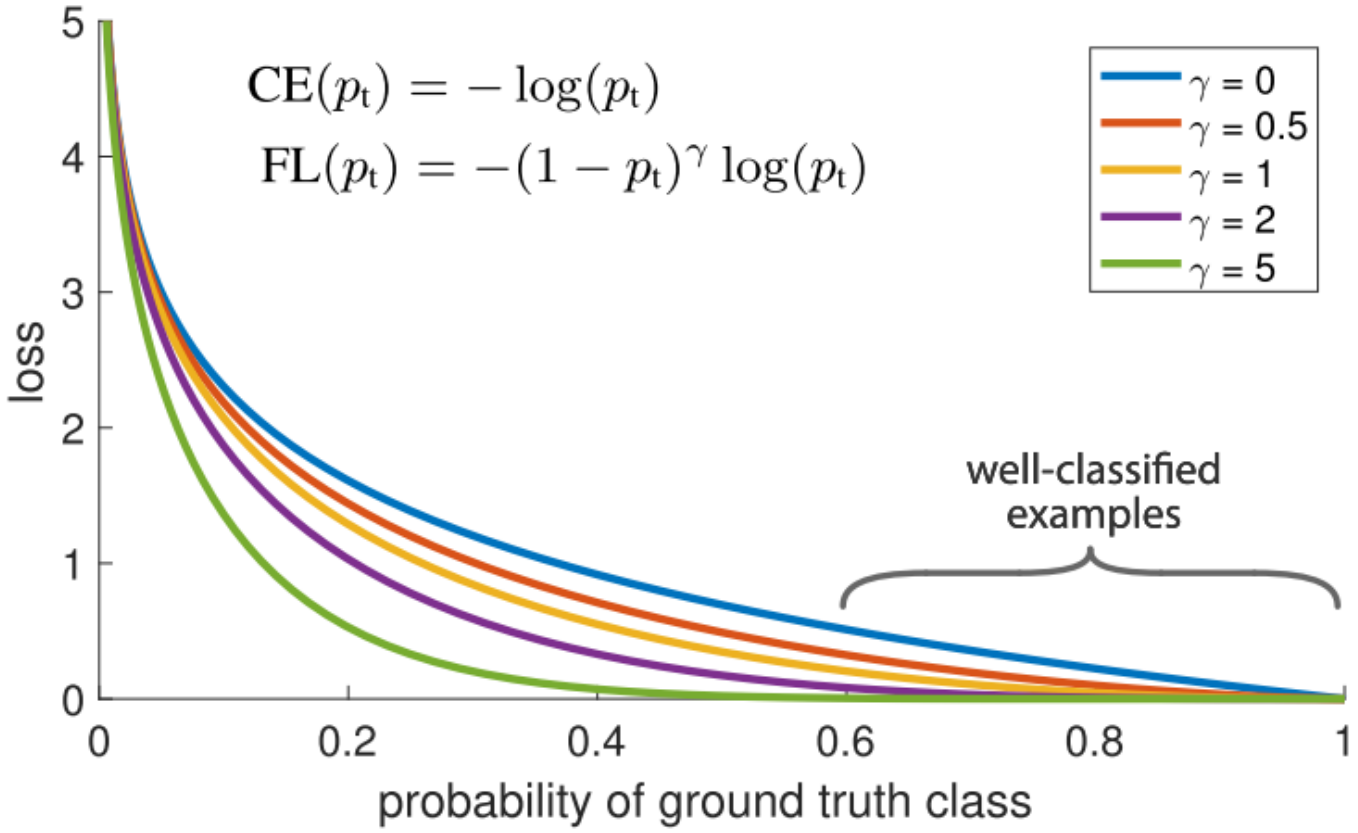


Feed Forward Neural Network

Scripps Research
Translational Institute

Future steps

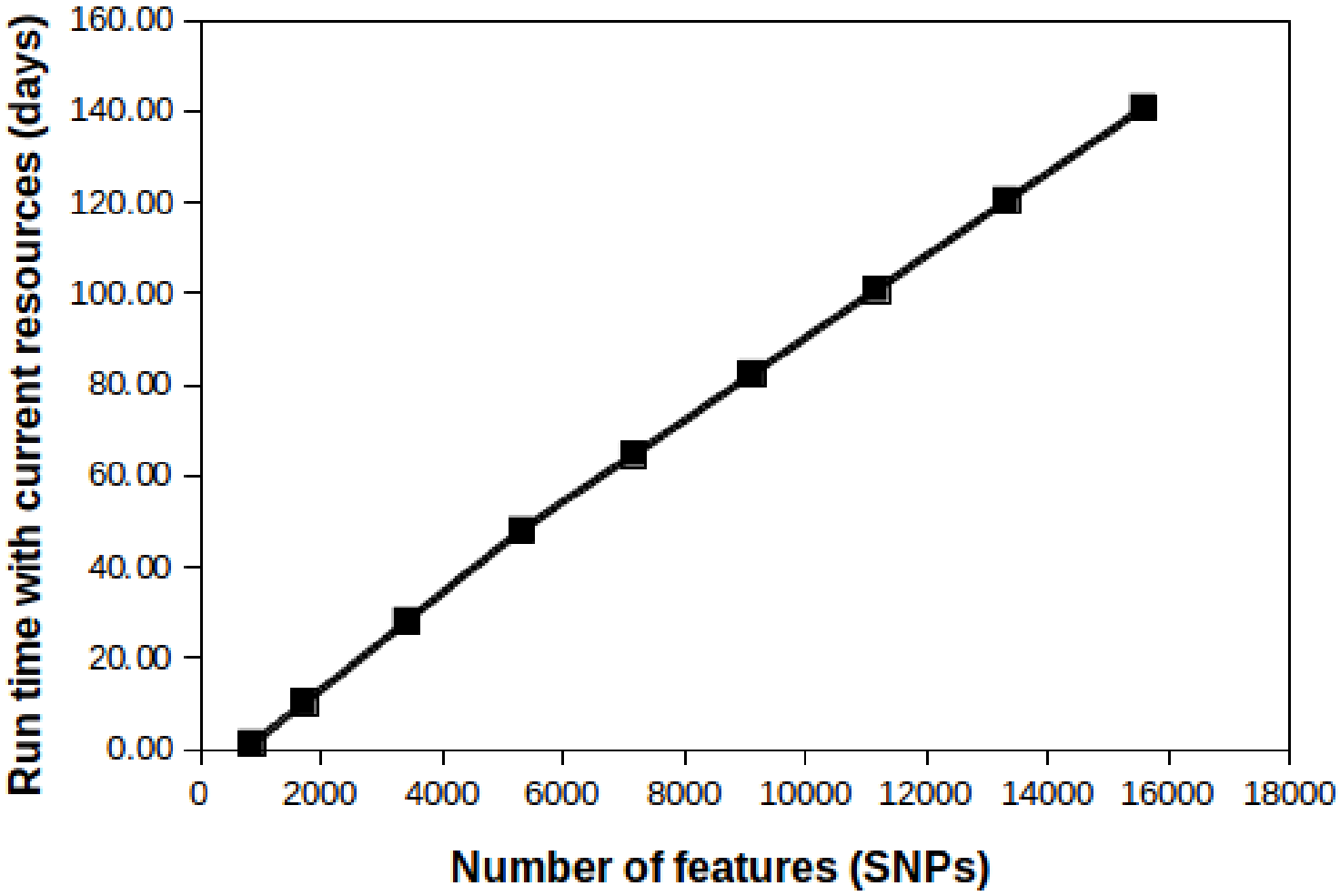
- **Focal loss to compensate for rare variants**



Scripps Research
Translational Institute

Lin, Tsung-Yi, et al. **"Focal loss for dense object detection."** *Proceedings of the IEEE international conference on computer vision*. 2017.

Limitations: expanding the methodology to other genomic regions



Scripps Research
Translational Institute

Acknowledgements

Scripps Research
Translational Institute

Ali Torkamani PhD
Shang-Fu Chen
Elias Salfati PhD
Doug Evans
Shuchen Liu
Alex Lippman
Nathan W. PhD
Emily Spencer PhD
Eric Topol MD

NVIDIA collaborators

Johnny Israeli
Carla Leibowitz
Fernanda Foertter
Brian Welker
David Nola

NIH National Institutes of Health

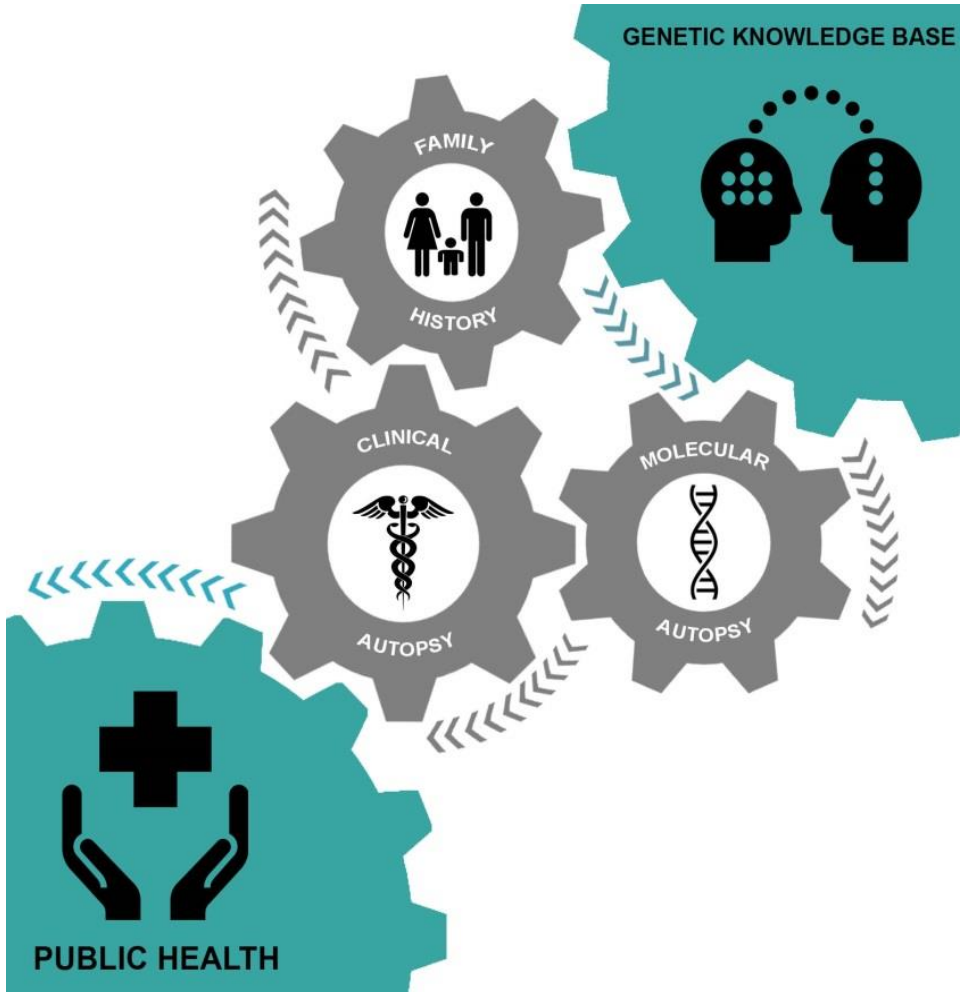
Funding

NIH/NCRR flagship CTSA Grant

Scripps Research

Contact:
raqueld@scripps.edu, @RaquelDiasSRTI
atorkama@scripps.edu, @Atorkamani

Thanks for your attention!!



Scripps Research