Deep learning and the prediction of human disease risk

Nicholas A. Furlotte[1], Babak Alipanahi[1] and David A. Hinds[1]

[1] 23andMe, Inc., Mountain View, CA, USA

Abstract

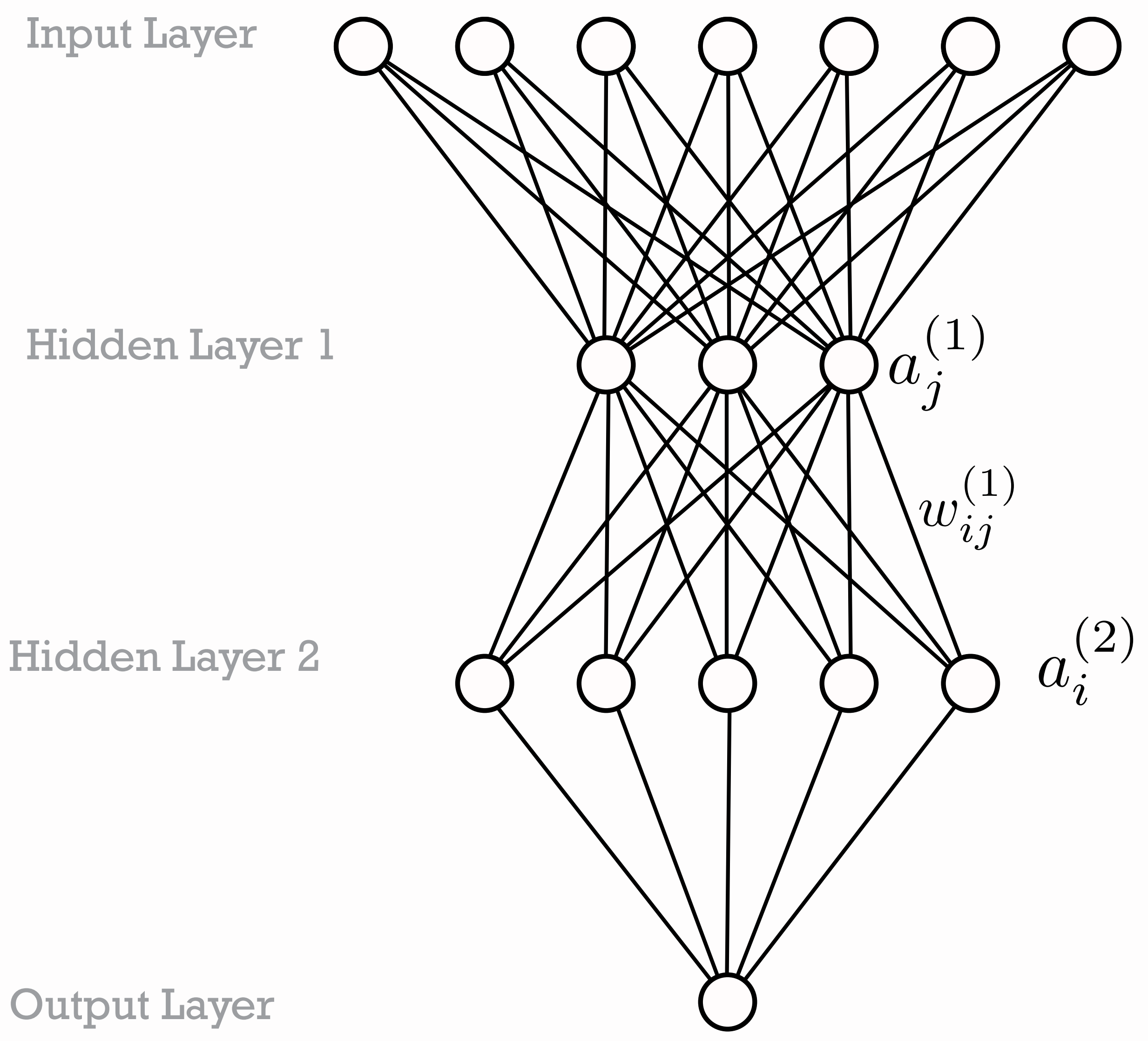
The accurate prediction of disease risk using genetic data remains one of the key challenges in human genetics research. Regression and whole-genome based methods that model the phenotype of interest as a linear function of a set of genetic variants have shown much promise, but generally their predictive power remains limited. One of the frequent criticisms of these methods is that they do not account for higher order interaction effects or more generally that they assume a restrictive genetic architecture when constructing the predictive model.

Artificial neural networks (ANNs) have experienced a resurgence in interest due to the success of deep learning in image and speech processing. These machines do not enforce a rigid relationship between the predictors and the target and can be thought of as general function approximations. Motivated by this feature and the general success of ANNs, recent work has suggested that ANNs might have an advantage over traditional methods of phenotype prediction and promising results have been demonstrated for genome-enabled trait prediction in cattle. However, there has been little recent work in developing and applying neural network-based approaches to the problem of predicting phenotypes and disease risk in human populations.

The development of deep learning based methods for human phenotype prediction has a number of challenges. First, deep learning methods require a large amount of training data. Second, there is not a standard methodology for applying deep learning techniques to prediction tasks or more practically there is not a well-defined network structure given the input data. Finally, large multi-layered networks can be computationally cumbersome to train and it can be difficult to control for overfitting.

In this project, we investigate these challenges and assess the ability of deep learning and ANN based methods to predict phenotypes with genetics by utilizing the large-scale 23andMe database of more than one million customers, 80 percent of whom consent to participate research. We compare the performance of standard methods like linear, logistic and ordinal regression with whole-genome-based and ANN-based predictive methods and evaluate performance across a spectrum of morphological and disease-related traits. In addition, we compare performance across different network architectures. Our results suggest a potential for applying deep learning methods to improve disease risk prediction.

Artificial Neural Network Structure



Variables such as SNPs and demographic info are inputs to the network.

The value of a hidden node is determined by a weighted sum of all nodes in the preceding layer.

$$a_i^{(2)} = \sum_j w_{ij}^{(1)} a_j^{(1)}$$

The final output is a function of all nodes from the final hidden layer. For example, the final output node is often a logistic function.

$$sigmoid(\sum_j w_{ij}^{(2)} a_j^{(2)})$$

Potential for Neural Nets in Phenotype Prediction

Universal Function Approximator

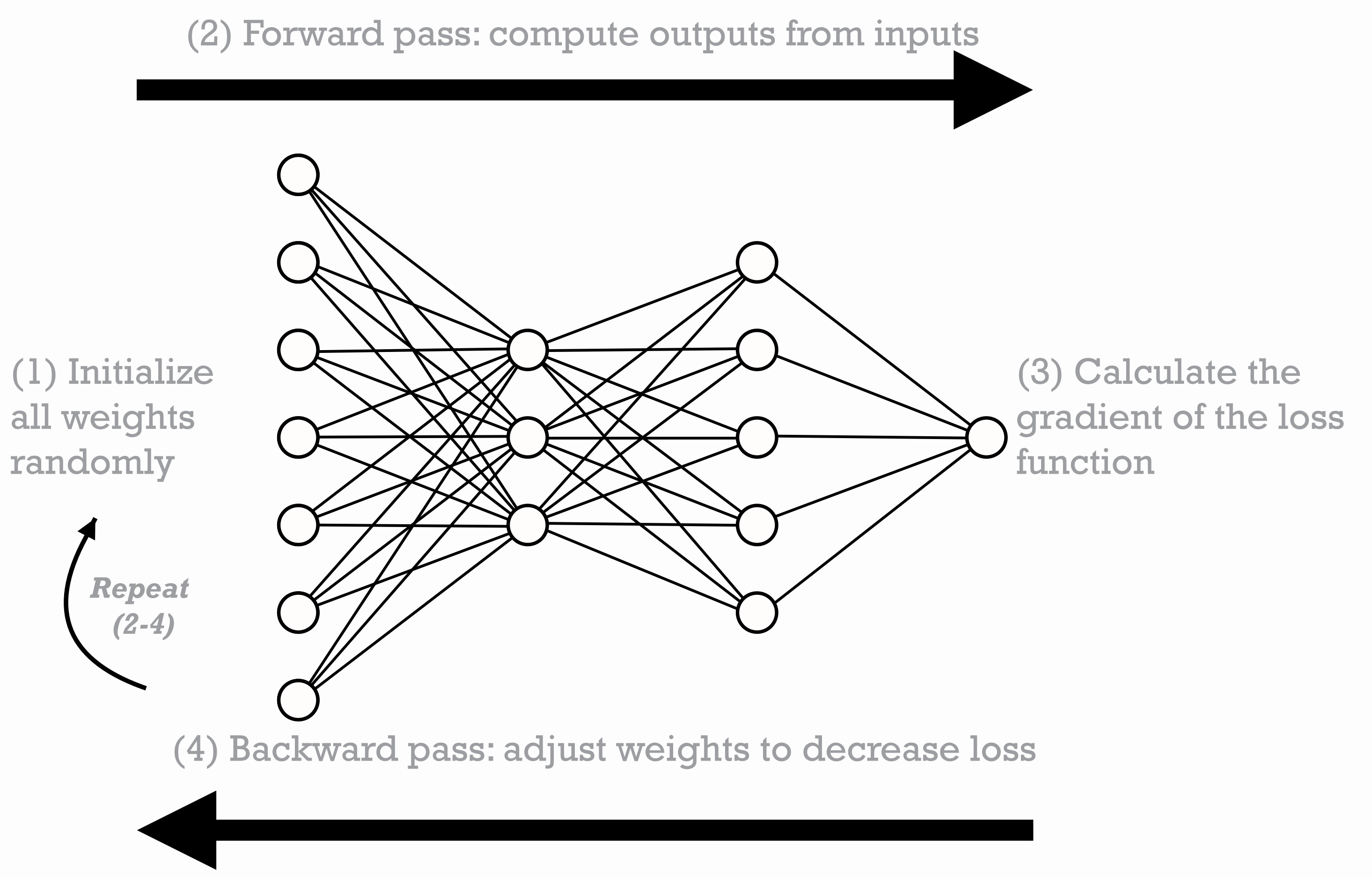
A multi-layered neural net can learn a diverse set of functions without explicit structure

Complex Interactions and Non-Linearity

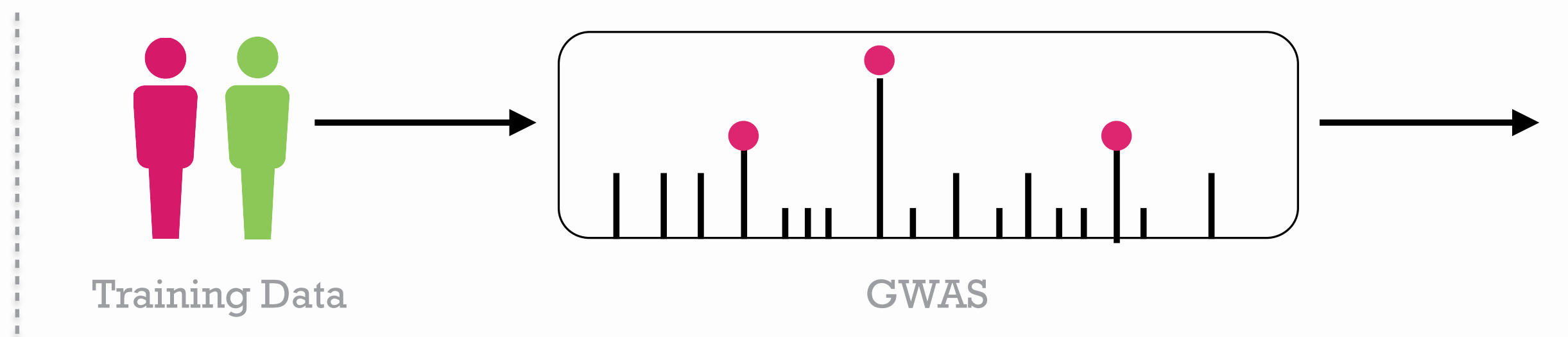
Handles complex gene-gene or gene-environment interactions as well as highly non-linear relationships between input and output

Methods

Back Propagation: Training the Network

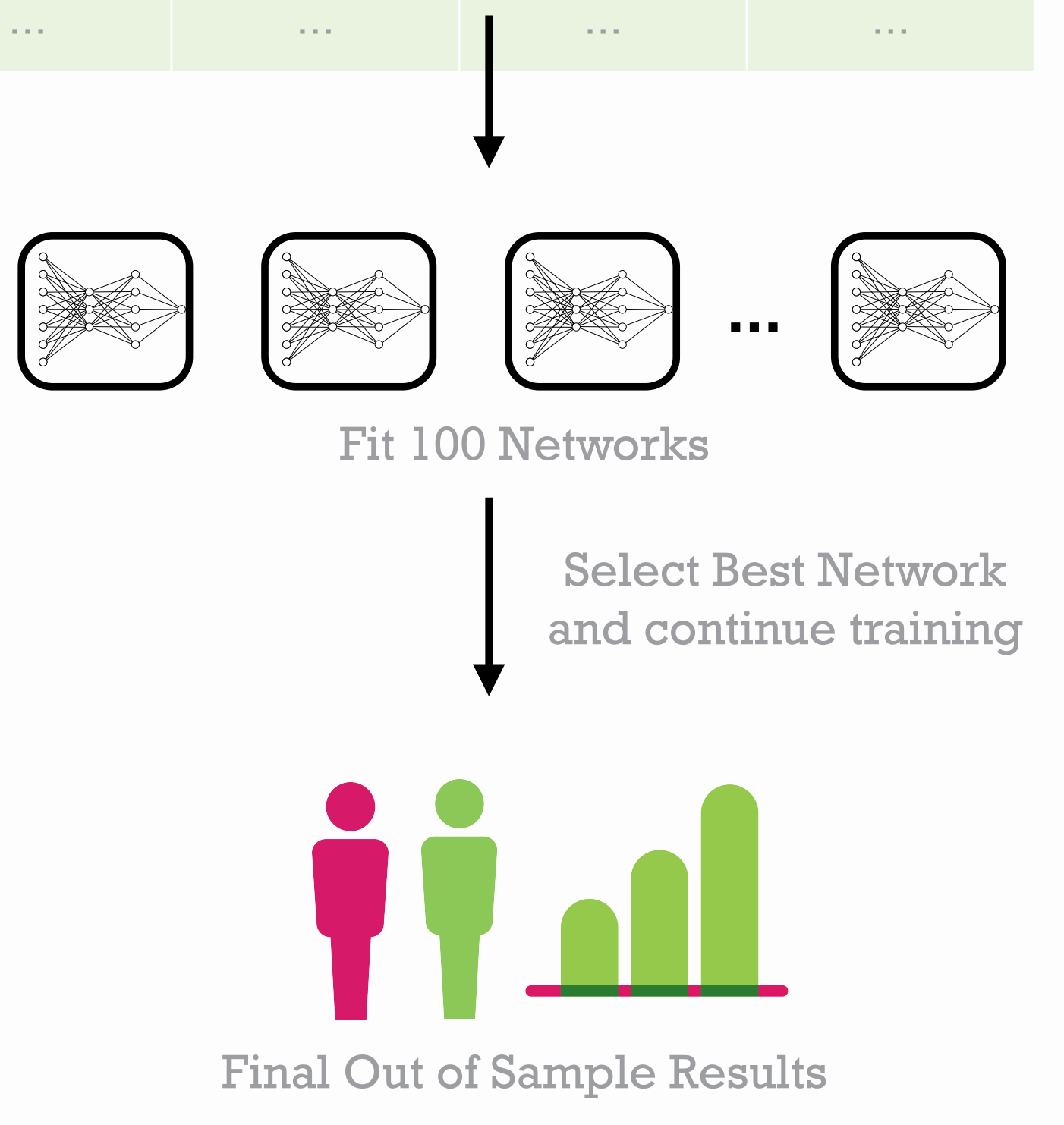

(2) Forward pass: compute outputs from inputs
(1) Initialize all weights randomly
Repeat (2-4)
(3) Calculate the gradient of the loss function
(4) Backward pass: adjust weights to decrease loss

Training Pipeline



Sample 100 Hyperparameter Settings

Learning Rate	Hidden Nodes	l1/l2	Momentum
1E-03	100	1E-05	0.25
1E-04	1000	1E-06	0.50
1E-05	1500	1E-07	0.75
…	…	…	…


Fit 100 Networks
Select Best Network and continue training
Final Out of Sample Results

Pipeline Overview

- We chose 1 disease phenotype, Type 2 Diabetes (T2D), and 8 non-disease appearance related traits for testing.
- Each data set was divided into 3/5 for training, 1/5 for validation and 1/5 for final testing.
- GWAS was run in the training set and we selected all SNPs with a p-value <= 1e-4.
- We sampled 100 hyperparameter settings, where the hyperparameters are learning rate, number of hidden nodes, momentum and l1 and l2 penalty. We did not employ drop out.
- A separate single layer neural net was fit for each hyperparameter setting. Training was done over 500 iterations. We selected the best setting and trained this model for an additional 1000 iterations. (Note: in some cases we continued for 5000 iterations)
- Age and sex were used as predictors in each model
- The final model is then evaluated in the test set.

Results

Comparison of Logistic Regression and Neural Net

Phenotype	European Cohort Size	Logistic AUC	Best Neural Net AUC
Sweet vs. Salty	143,600	0.573 (0.007)	0.538 (0.007)
Chin Dimple	77,400	0.682 (.012)	0.564 (0.011)
Attached Earlobes	62,000	0.661 (0.01)	0.622 (0.01)
Freckles	154,700	0.716 (0.007)	0.689 (0.007)
Dimples	73,200	0.597 (0.01)	0.590 (0.008)
Photic Sneeze	123,453	0.649 (0.008)	0.64 (0.007)
New Born Hair	55,798	0.610 (0.01)	0.592 (0.01)
Widows Peak	73,075	0.607 (0.01)	0.595 (0.01)
T2D	322,000	0.776 (0.003)	0.782 (0.003)

Table 1: Comparison of Logistic Regression AUC with the best Neural Net AUC.
We computed AUC for each binary trait in the final test set using a logistic regression classifier as well as the best neural net classifier chosen by validation AUC. The neural net classifiers either fell short or had similar performance to the logistic. This may indicate that the relationships between the input and output is not very complex or that the single layer neural net we employed is not able to learn this complexity. The best performing models had 0 drop out and high momentum (0.75-0.99).

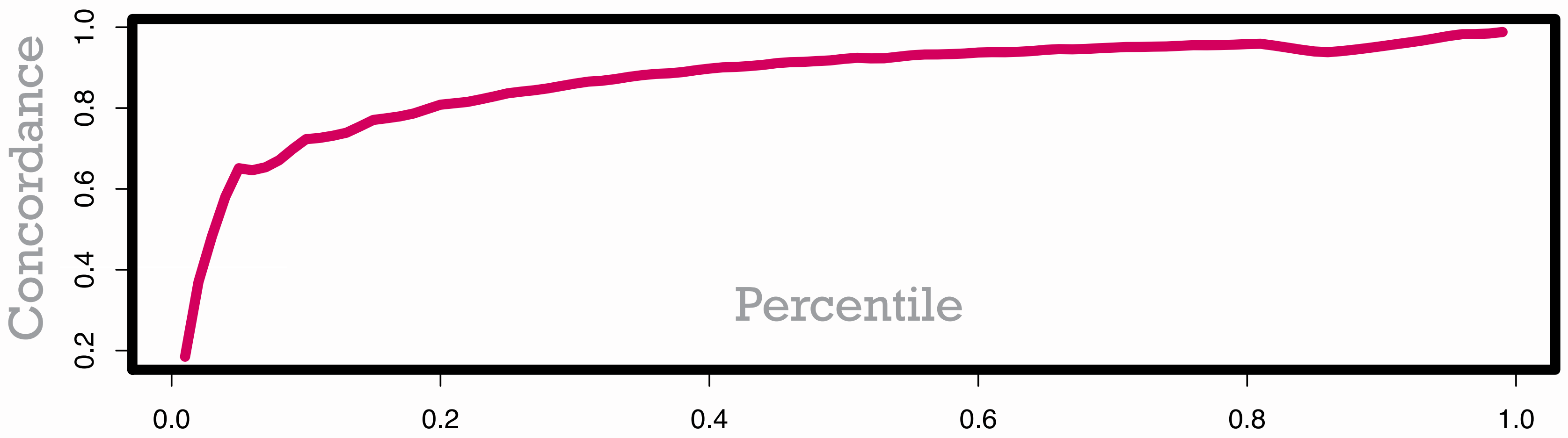


Figure 1: Concordance between risk for T2D as predicted by logistic regression and the best neural net model.
Although both approaches result in nearly the same AUC, there is a substantial difference between the ranking of individuals with respect to their risk score in the top 5% and beyond.

Future Work and Challenges

Given the current results, we believe that the exploration of more complex neural net structures is necessary. In theory, the single layer neural net structure with a logistic output layer should yield results at minimum equivalent to logistic regression. We did not observe this for all traits. In addition, we often find that the validation error fails to increase even after a large number of iterations, indicating that the network is failing to overfit to the training data, which could be due to limited complexity of the data or the network.

23andMe Data

Research participants were drawn from the customer base of 23andMe, Inc., a personal genetics company. Customers were genotyped on a custom Illumina HumanOmniExpress+ genotyping chip. Participants provided informed consent and answered research questions online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (E&I Review).

Acknowledgments

We thank 23andMe customers who consented to participate in research for enabling this study. We also thank employees of 23andMe who contributed to the development of the infrastructure that made this research possible.

References

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley and Y. Bengio. "Theano: new features and speed improvements". NIPS 2012 deep learning workshop.

https://github.com/Lasagne/Lasagne