Deep integrative models for large-scale human genomics

Arnór I. Sigurdsson[1], David Westergaard[1], Ole Winther[2,3], Ole Lund[4], Søren Brunak[1], Bjarni J. Vilhjálmsson[5,6,7], and Simon Rasmussen[1]

[1]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen N, Denmark
[2]Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark
[3]Bioinformatics Centre, Department of Biology, University of Copenhagen
[4]Danish National Genome Center, Ørestads Boulevard 5, 2300 Copenhagen S, Denmark
[5]National Centre for Register-Based Research (NCRR), Aarhus University, Aarhus, Denmark
[6]Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Aarhus, Denmark
[7]Bioinformatics Research Centre (BiRC), Aarhus University, Denmark

Corresponding authors: Simon Rasmussen (simon.rasmussen@cpr.ku.dk) and Bjarni J. Vilhjalmsson (bjv@econ.au.dk).

ABSTRACT

Polygenic risk scores (PRSs) are expected to play a critical role in achieving precision medicine. PRS predictors are generally based on linear models using summary statistics, and more recently individual-level data. However, these predictors generally only capture additive relationships and are limited when it comes to what type of data they use. Here, we develop a deep learning framework (EIR) for PRS prediction which includes a model, genome-local-net (GLN), we specifically designed for large scale genomics data. The framework supports multi-task (MT) learning, automatic integration of clinical and biochemical data and model explainability. GLN outperforms LASSO for a wide range of diseases, particularly autoimmune disease which have been researched for interaction effects. We showcase the flexibility of the framework by training one MT model to predict 338 diseases simultaneously. Furthermore, we find that incorporating measurement data for PRSs improves performance for virtually all (93%) diseases considered (ROC-AUC improvement up to 0.36) and that including genotype data provides better model calibration compared to measurements alone. We use the framework to analyse what our models learn and find that they learn both relevant disease variants and clinical measurements. EIR is available at https://github.com/arnor-sigurdsson/EIR.

Keywords: Deep Learning, Machine Learning, Genomics, Precision Medicine, Polygenic Risk Scores, Data integration

INTRODUCTION

Polygenic risk scores (PRSs) are becoming increasingly relevant to public health due to larger cohorts and the development of more powerful prediction algorithms. Today, accurate PRS predictors have been trained to predict various human diseases such as type 2 diabetes, coronary artery disease and breast cancer[1–3]. Such PRS predictors are expected to become pervasive in clinical human health and decision making, hence playing a fundamental role in achieving personalized medicine[4–6]. PRS predictors can generally be placed in two categories based on the type of training data used, those using summary statistics from genome-wide association studies (GWAS) and those using individual-level data[7]. Today, the combined GWAS approach is more prevalent due to larger sample sizes. However, this is rapidly changing with individual-level human genetic variation data increasing in size, with cohorts comprising hundreds of thousands and even millions[8–12]. These large individual-level cohorts increasingly offer the opportunity of training accurate predictors for estimating PRSs which can outperform the combined GWAS approach[7]. Today many established methods exist for training predictors on summary statistics[13–18] and individual-level data[19–23], but these predictors generally explore linear relationships.

Artificial Intelligence (AI) and Deep Learning (DL) has revolutionized several scientific fields and is changing our society. At the end of 2019 it was selected as one of ten scientific events that shaped the last decade[24]. Within life sciences, DL has gained pace within the recent years[25–27]. In particular for imaging, but also notably single-cell sequencing[28,29], protein localization and protein folding[30,31]. For genome sequence data, the efforts have mainly been focused on identifying motifs such as ChIP-seq or identifying genomic variation[32]. Simultaneously, DL frameworks for large discrete data, such as genome-wide data, have not been extensively developed in the field. A potential advantage of DL based methods for PRS prediction is capturing complex non-linear effects, such as epistasis. However, previous work using neural networks (NNs) for predicting human traits and diseases directly from large-scale genomics has shown worse performance for NN models compared to linear ones[33,34]. The results indicate that the NNs were not able to capitalize on significant interaction effects or that no significant interaction effects are present in the data. The latter explanation is in contrast with studies focusing on model organisms, where significant interaction effects have been found[35–37]. However, there remains both doubt and controversy regarding the role of complex interaction effects in human traits and diseases[38–41]. Therefore, there could be many explanations for why the non-linear models have not shown consistent gains over linear models in predicting human traits and diseases. These include (a) linear models are able to capture the upper bound of the genetic variance explained by genotyped SNPs, (b) for the traits tested, there are either not enough samples or not enough SNPs measured to find the interaction effects, (c) there is a significant risk of mislabelled samples in the case of diseases which could affect the model's ability to learn subtle interaction effects, (d) interaction effects are present, but their effects are non-significant and (e) interaction effects are present and significant, but the non-linear models are not able to capture the interactions.

Despite significant non-linear interactions not having been widely shown, there are still advantages of developing NN based models for the prediction of disease risk. One factor is the flexibility of NN based models that can be developed to accept multi modal inputs and predict more than one trait at the same time. Some of these inputs might even be unstructured and highly non-linear, such as images or text. This in itself is a clear advantage over linear models. Additionally, even though the separate modalities could mostly be modelled with linear models, there is always the possibility that complex interaction effects, e.g. between clinical measurements and other biological data exist. Using NNs allows us to discover such

relationships.

However, there are many challenges with building complex NN models that can be applied to human health data. A key challenge is the immense scale of biological data. For example, genomics data often contain millions of genetic variants genotyped for large sample sizes[10,42,43]. Another challenge with fully leveraging health data is that it is often multi modal. Supervised machine learning tasks are often trained to accept one type of input, for instance classifying the main object in a given image. By contrast, health data can be compromised of multi-omics data such as genomics, transcriptomics and proteomics data coupled with targeted biochemical and clinical data, and even include ultra-high resolution imaging. In order to provide a comprehensive disease risk assessment, methods that can account for genetic, environmental and other risk factors can be advantageous.

Therefore, we developed a DL framework that supports large-scale genomics data and can integrate it with other omics or clinical data. Among the features of the framework is a new neural network model, genome-local-net (GLN), that we specifically developed for large scale genomics data. GLN is based on a custom locally-connected layer (LCL)[44–46] we developed, and it was able to extract genetic information from genome-wide data with equal or better performance compared to the other NN models we tested. Compared to our implementation of least absolute shrinkage and selection operator (LASSO)[47], we found GLN to perform statistically better overall on 338 diseases, disorders and traits. Of particular interest were autoimmune diseases such as type 1 diabetes (T1D) and rheumatoid arthritis, which previously have been shown to have complex interaction effects[48,49]. All models in the framework automatically extend to multi task (MT) learning, which we showcase by training one GLN model to predict 338 diseases at the same time. Furthermore, we integrate genotype data, age and sex covariates, blood measurements, urine measurements and various anthropometrics when training GLN based models on 290 diseases in the UK Biobank (UKBB). The integration of these measurements shows a clear improvement for almost all traits, highlighting the potential of integrative models for health based predictions. Using explainable AI we identify relevant SNPs and clinical measurements concordant with disease literature and finally we show that integrating genotype and clinical data gives better calibrated models for precision medicine.

RESULTS

A locally-connected layer for genome-wide data

We initially established the ability of different candidate models to capture additive and non-linear XOR (interaction) effects when using simulated genotype data to predict a continuous output. As expected, we found that the NN-based multilayer perceptron (MLP)[50] and convolutional neural network (CNN)[51,52] models were able to capture and model non-linear interaction effects with R^2 of 0.95-0.98. However, the linear LASSO model had R^2 of 0.75 for a mix of additive and XOR effects, whereas it completely failed to model pure XOR effects with an R^2 of -0.03 (Table S1). When scaling the NN based models to genome-wide genotype data or even to whole-genome sequencing data, the number of parameters when using fully connected (FC) layers increases dramatically. For instance, a FC layer with an input of one million one-hot encoded SNPs (i.e. 4 elements per SNP) would require roughly 400 million weights to be connected to a hidden layer of 100 neurons. While convolutional layers can be much more parameter efficient, the computational complexity of training them on very high dimensional inputs can rival or exceed that of FC layers[46]. Therefore, in order to have a model that was both parameter efficient and could take advantage of the local positional variance in genomics data, we implemented a locally connected layer (Figure S1). The layer is sparsely connected through groups, which greatly reduces the

number of parameters in comparison to a FC layer. The sparse connection allows for a larger intermediate representations while still keeping the parameter count relatively low. The GLN model is composed of multiple LCLs (Figure 1a), and as was the case with the MLP and CNN models, it effectively captured both additive and non-linear effects in the simulated data ($R^2 = 0.98$) while using fewer parameters (1.6X and 5.1X fewer than CNN and MLP respectively).

GLN based Genome-wide Polygenic Scores can outperform LASSO in the UKBB cohort

We then trained and validated PRSs using either LASSO, a MLP, CNNs or the GLN model for eight selected traits on 413,736 individuals with British, Irish, or any other Western European background in the UKBB cohort (Figures 1b, S2 and Data S1). Interestingly, we found that GLN was superior to using our LASSO model for T1D, Gout and Acute Myocardial Infarction with improvements of 0.04, 0.02 and 0.01 ROC-AUC on a held-out test set (Methods). For the remaining traits the differences were less than 0.01 ROC-AUC. Additionally, the GLN had better performance compared to the MLP and CNN with average improvements of 0.01 and 0.04 ROC-AUC, respectively. This replicates previous results where CNN based models did not show a consistent advantage for human trait prediction[33,34]. To examine whether the gain of 0.04 ROC-AUC for T1D was due to the chosen hyperparameters for the LASSO model, we retrained the LASSO with various combinations of hyperparameters but did not find it to match the performance of the GLN model (Table S2). To investigate and explain what the models had learned, we determined the SNPs that had the highest effects and cross referenced them to known associations for a particular trait. Specifically, for the T1D model we found that both LASSO and GLN were activated at the HLA region of chromosome 6 (Figure 1c,d) – a region that has previously been associated with T1D[53]. However, the GLN model was also highly activated by SNPs at multiple sites across the genome (Figure 1d and Table S3). Interestingly, among the ten most highly activated SNPs, three of them were not on chr6. One of these (rs3842753) had previously been associated with T1D[54] and was detected by our model, indicating that GLN can detect a wide array of associated SNPs for a given disease or trait.

GLN identifies disease relevant variants

When expanding the activation analysis to the other 7 traits, we found that in all cases a known association was found among the top 10 SNPs or the genes they reside in (Figures S3–S10 and Tables S3–S10). This is a strong indication of the models learning biologically relevant associations and that complex neural networks can be interpreted when modelled on extremely high dimensional genotype data. Even for diseases such as Acute Myocardial Infarction and Gout where the covariates (age, sex and first 10 genotype principal components) had a better performance (Figure S11 and Data S2) we found that the GLN model was activated by numerous relevant SNPs and genes for both diseases (Tables S4 and S8). The better performance of the covariate based models could be due to the covariates having much larger effects than the genotyped SNPs, e.g. if a disease is strongly affected by age or population stratification. Including the high-dimensional genotype data increases overfitting, which inflicts a performance trade-off against the much lower dimensionality of using only the covariates. Hence, a higher case count might be required to capture the SNP effects to such a degree that it boosts performance over the covariate based models[55].

Improved PRSs for autoimmune diseases

Knowing that the GLN model was competitive with the LASSO implementation on the 8 traits, we applied a more data driven approach of training GLN, LASSO and two covariate based models on 338 binary disease traits with at least 1,000 cases in the UKBB cohort (Data S3). Among the four

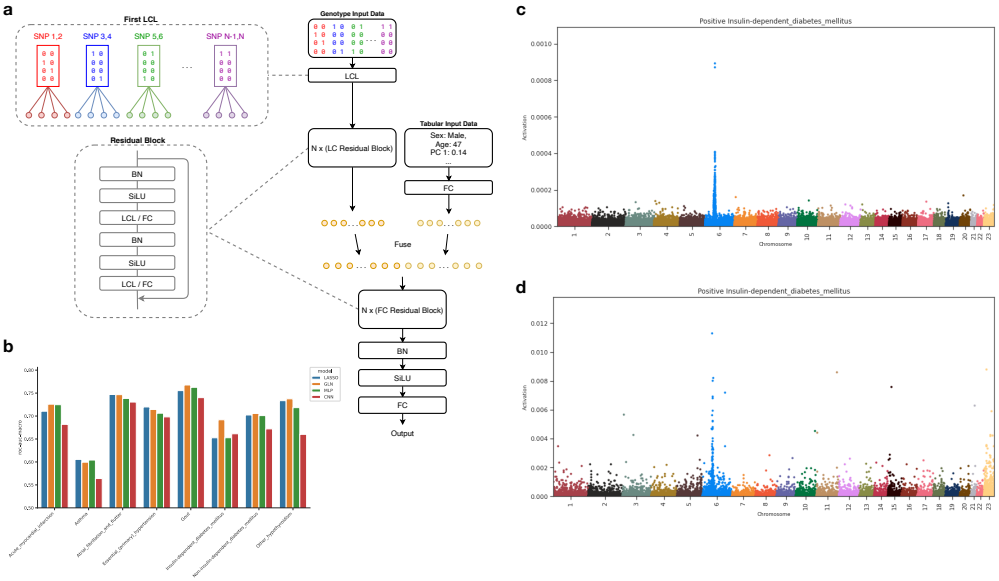

Figure 1. Genome-local-net (GLN) model architecture and performance. (a) Model architecture. The model uses a locally connected layer (LCL) with a kernel width covering two SNPs and four weight sets as the first layer. The output of the first layer subsequently goes through residual blocks composed of LCLs. The genomic representation is then fused with the tabular representation, which then is propagated through FC based residual blocks. A final set of `BN-SiLU-FC` layers is used to compute the final output. **(b)** Model performance on the test set across 8 traits in AUC-ROC. All models were adjusted for age, sex and the first 10 genomic PCs. **(c)** SNP activation distribution for type 1 diabetes for LASSO, showing high activations in the and around the HLA region. **(d)** SNP activation distribution for type 1 diabetes for GLN, showing high activations in and around the HLA region.

models tested, the GLN performed best on 59 diseases whereas the LASSO model performed best on 42 diseases. Comparing the ROC-AUCs between GLN and LASSO, we found GLN to perform better overall (Wilcoxon signed-rank test, one-sided, $P = 4.7 \times 10^{-14}$). Interestingly, using only covariates had the best performance for the remaining 237 traits, and overall it performed better when compared to GLN (Wilcoxon signed-rank test, one-sided, $P = 4.6 \times 10^{-15}$ and $P = 0.0015$ for linear and NN based covariate models respectively) (Data S4). The covariate based models performing better could be due to the low effective sample size (ESS) and the nature of some traits being more driven by environmental factors (Figures S12 and S13). When subsetting disease traits for where GLN and LASSO had better performance compared with covariates and difference of at least 0.01 ROC-AUC we found 16 and 9 disease traits where GLN and LASSO had the best performance, respectively (Figures S14 and S15). Interestingly, the GLN model performed markedly better on T1D, rheumatoid arthritis, multiple sclerosis, psoriasis and ulcerative colitis, all autoimmune traits in which studies have shown indication of interaction effects[48,49,56–59]. For instance, for rheumatoid arthritis, the GLN model had a ROC-AUC of 0.665 while the LASSO had a ROC-AUC of 0.624 on the test set and the covariate only models achieved a ROC-AUC of 0.622 and 0.634 for the LASSO and NN based models, respectively (Figure S16). When examining GLN and LASSO activations for rheumatoid arthritis, we found, as above, the models to be activated by relevant SNPs (Figures S17 and S18 and Tables S12 and S13). Taken together, our results therefore show an improvement of using NNs compared to LASSO for predicting disease risk from genome-wide genomics data.

GLN is fast and robust to missing data

To simplify calculation of PRSs we implemented the models, including the LASSO, to automatically handle missing genotype data and thus removing the need to impute data prior to training. In this work, the genotype data was not pre-processed extensively prior to modelling (denoted "NO-QC"). To investigate whether our results were consistent when using traditional pre-processing (denoted "QC"), we also trained GLN and LASSO on QC data. Besides reducing the number of SNPs and samples considered, the QC approach additionally results in a different train/test split. The NO-QC approach gave slightly better results on our eight benchmark traits (Figure S19 and Data S5). However, we did see the overall trends remain consistent whether using QC or NO-QC, e.g. with GLN performing markedly better on T1D (Figure S19). Finally, we investigated the computational costs of training the PRS models. For the eight benchmark traits we found training the GLN to be slightly faster (32 hours) compared with our LASSO (34 hours) (Figure S20 and Data S6). Even though the training latency of the LASSO model was lower than any of the NN based models, the total training time was higher due to using more steps before model convergence. Therefore, the framework was able to train large and deep neural networks on high dimensional individual-level genotype data in a reasonable time.

Multi-task learning offers a trade-off between performance and complexity

In multi-task (MT) learning, a single model is trained to solve multiple objectives at the same time, such as predicting height, disease liability and ethnicity. This can lead to improved predictive performance, reduced training time and better parameter efficiency[60,61]. We therefore hypothesized that predicting multiple outcomes simultaneously could regularize and potentially improve prediction performance. Using Type 2 Diabetes (T2D) for comparison we trained MT models to predict two, eight and 338 diseases jointly and found that maximum validation performance got progressively worse when increasing the number of tasks (Figure 2a and Data S7). This indicates that the model capacity was not high enough to effectively capture the variance of multiple traits as well as the single task model or that negative

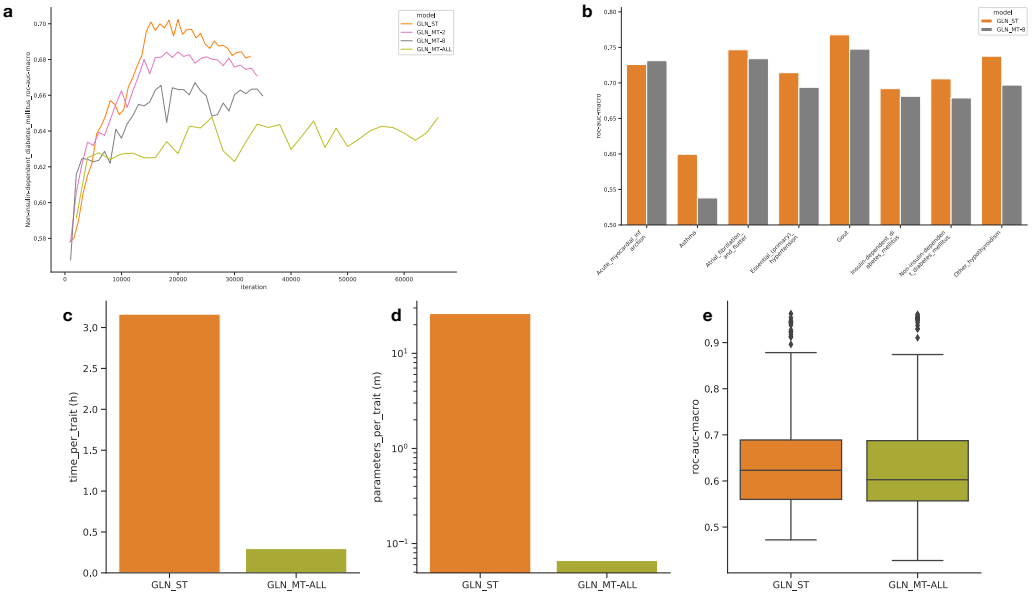

Figure 2. Genome-local-net (GLN) multi-task (MT) predictions. (a) Comparison of validation curves in ROC-AUC for type 2 diabetes as more tasks are added alongside the type 2 diabetes prediction task. The two task model is trained jointly on T2D and hypertension. The 8 task model is trained on the 8 benchmark traits showed in figure 1b. The ALL model is trained the total set of 338 diseases considered in this work. The single and two task runs show signs of overfitting as validation performance peaks and starts to deteriorate around 20K iterations. The eight and ALL multi task runs do not show as clear signs of overfitting, but overall performance is worse. **(b)** Comparison of single task and MT performance for the 8 benchmark traits. **(c)** Comparison of training time per trait for the ALL MT model and single task training. **(d)** Comparison of number of parameters per trait for the ALL MT model and single task training. **(e)** Overall performance of the ALL GLN MT model and single task training.

transfer between tasks degraded performance[62]. Similarly, when comparing test set performance the respective single task models and a MT model trained on the 8 benchmark traits, we found that the MT model was slightly worse for 7 diseases (average 0.023 ROC-AUC lower), with Acute Myocardial Infarction being the exception (0.0054 ROC-AUC improvement) (Figure 2b). However, despite being slightly worse for most of the traits, the MT models were remarkably effective. For example, the 8 trait MT model had a test ROC-AUC of 0.68 for T1D, which was considerably higher than the 0.58-0.59 ROC-AUC when using only covariates. In order to examine how well the framework scaled and whether we could effectively train very large scale MT models, we trained one GLN model to jointly predict 338 traits simultaneously. As expected, modeling on all traits jointly significantly reduced the training time (11X) and number of parameters per trait (395X) (Figure 2c,d). As in the other MT experiments, this came at the cost of reduced performance compared to the single task setting (Wilcoxon signed-rank test, one-sided, $P = 0.025$), however only with an average difference of 0.0056 ROC-AUC (Figure 2e and Data S8). Compared to the best performing covariate based models for each trait, the large MT model performed better for 61 traits, indicating that it was able to effectively capture genotype variance for some traits and not only using the covariates.

Integrating genomics with clinical data improves predictive performance

Although genetic data has proved to be a powerful predictor of various traits and diseases, there are other factors such as environmental effects that can play an important part[63]. With the increased digitization in the healthcare industry, clinical and electronic health data is only expected to become more

widely available. Among these are factors that are relatively easy and non-invasive to measure, such as anthropometrics, and other measurements included in the UKBB such as blood and urine measurements. To examine the benefit of using these measurements as part of our models for computing PRSs, we trained GLN models using only genotype data and covariates, and compared this with using genotype, covariates, physical, blood and urine sample measurements (denoted "Integrated"). Furthermore, to minimize data leakage we removed samples where the diagnosis occurred in the past compared to the measurements, ensuring that the model predicted future diagnoses and not previously diagnosed conditions. Therefore, when including the measurements the number of cases was for most traits reduced, leading to a trade-off between the gain of including measurements and the loss of removing samples. To examine this trade-off more precisely we compared to two genotype datasets, one where the censored individuals were removed (denoted "Genotype Filtered"), and another set where all individuals were included (denoted "Genotype") (Figure 3a and Data S9). For all eight benchmark traits, as expected, removing samples reduced performance with ROC-AUC of 0.014-0.092 (Wilcoxon signed-rank test, one-sided, $P = 0.0039$). Another contributing factor is that the sample removal is likely biased towards those individuals that have a high genetic load, and therefore diagnosed early. Compared to Genotype Filtered data, we found that using Integrated data greatly improved performance, with ROC-AUC increasing by 0.043-0.27 (Wilcoxon signed-rank test, one-sided, $P = 0.0039$) (Figure 3a). This was also the case when using Matthews correlation coefficient (MCC) as metric which improved between 0.010-0.35 (Figure S21). However, compared to the unfiltered Genotype data the results were more disease dependent. For instance, filtering hypothyroidism for time of diagnosis reduced case count from 16,894 to 4,663 in the training set, which was reflected in ROC-AUC performance reduction of 0.092. Including measurements therefore did not outweigh the performance reduction of discarding cases. Interestingly, we found that using Integrated data had superior ROC-AUCs for five of the traits compared to using measurements and covariates only (denoted "Measurements"), highlighting the benefit of including genotype data.

Integration of clinical and genomics data improve prediction of T2D

When investigating the activations of the integrative predictor we found that the model was activated by relevant clinical measurement features such as glycated haemoglobin (HbA1c) and blood glucose for T2D (Figure 3b). However, for some diseases, such as T2D, predictors using Measurements data and the Integrated data had very similar ROC-AUC performances. This does not necessarily indicate that the genetic component of the traits was low, perhaps the more likely explanation is that the measurements can act as a proxy for the genomics effects. For example, in the case of T2D, high genomics risk will in many cases manifest itself in high levels of glycated haemoglobin, and when it can be measured directly, there is perhaps not much extra variance gained when including the genotype data. However, when we investigated the genotype activations we found that the model was activated by relevant SNPs even when measurement data was included (Figure S22). This was a similar trend we observed above, where the addition of genotype data resulted in worse overall performance, despite the models being activated by disease relevant SNPs. When exploring this further, we found that including the genotype data resulted in a predictor for T2D with higher MCC (0.43) compared with using only the measurements (0.33) (Figure S21). Interestingly, this was particularly due to better classification of true negatives and indicates the usefulness of the integration.

Large scale integrative modeling

Knowing that the integrative models could increase performance and learn relevant features, we, as before, performed large scale analysis of 290 traits that included time of diagnosis. Integration of the

measurements showed a large increase in performance for almost all of traits compared with the Genotype Filtered predictor (Figure 3c and Data S10). Interestingly, we had expected improvements for endocrine, nutritional and metabolic diseases but found improvements in ROC-AUC and MCC for other classes of diseases such as mental and behavioral disorders (Figure S23). Compared with the Genotype predictors we observed the same overall trend that including measurements improved performance, but the effect was less pronounced due to the Genotype models using more samples (Wilcoxon signed-rank test, one-sided, $P = 6.1 \times 10^{-37}$ and $P = 9.9 \times 10^{-09}$ for ROC-AUC and MCC, respectively) (Figures S24, S25 and Data S11). To examine the effect of including genotype data when measurements were available, we compared the Measurements based models to models using Integrated data and found that the differences was small for most traits (Wilcoxon signed-rank test, one-sided, $P = 0.048$ and $P = 0.054$ for ROC-AUC and MCC, respectively) (Figures S26, S27 and Data S12). This could be due to low ESS for many of the traits, traits being driven more by environmental effects and/or high genomic risk being reflected in the measurements (Figures S12, S13 and S28).

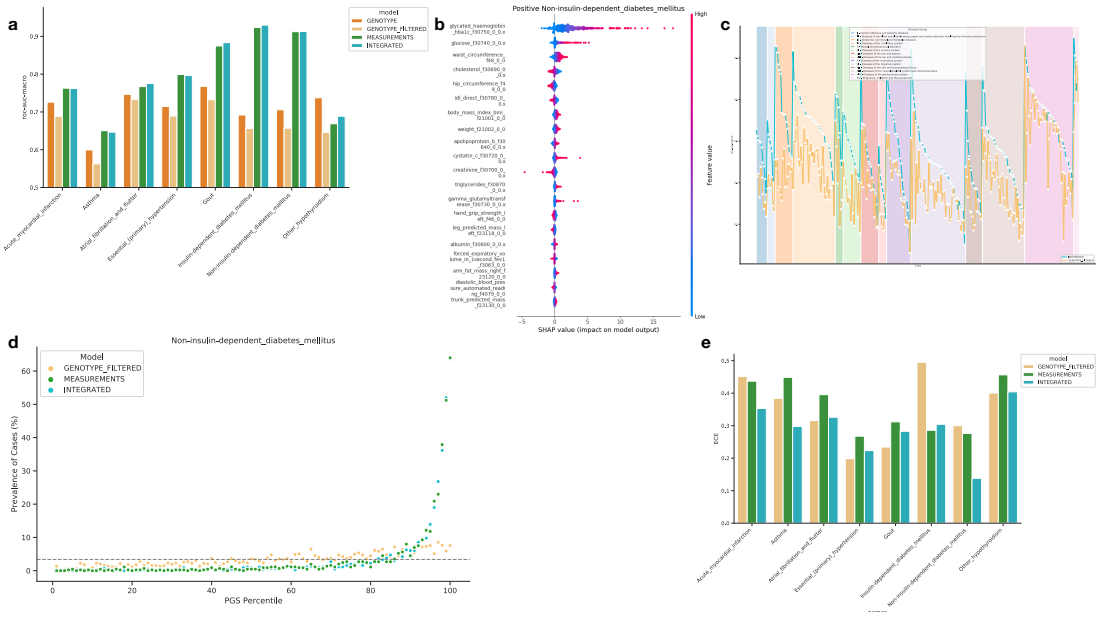

Figure 3. Integrating genotype and clinical data with genome-local-net (GLN). (a) The 8 benchmark traits with combinations of genomics and measurements compared using ROC-AUC. **(b)** Feature importance and impact of integration measurement values on the GLN model prediction for T2D. For a given feature, each dot represents a sample in the test set. The colors indicate the actual feature value. For example, a strong trend of high glycated haemoglobin value influencing the model to make a positive prediction for T2D can be seen. **(c)** Summary of ROC-AUC across all the 290 traits that had a time measured column associated with them, with all data compared with using only genotype and covariate data, filtered for time of diagnosis. **(d)** Prevalence plot of T2D prediction when using the GLN model on different data sources. **(e)** Comparison of Expected Calibration Error (ECE) for the different data sources when using the GLN model.

Genotype data calibrates PRS from measurement data

For clinical decision making, performance measurements such as the ROC-AUC and MCC are often insufficient to describe model performance[64,65]. We therefore examined the risk score distribution of our models, specifically investigating the relationship between the fraction of positives for a given percentile of the PGSs (Figures 3d and S29). When using Integrated data, we found that 5 of the 8 diseases had a prevalence of 10% or more in the top percentile, whereas models using Genotype Filtered or

Measurements data had the same performance for 2 and 5 diseases respectively. However, we also found that the models were poorly calibrated due to overconfidence, an issue that could be amplified by mislabeled data, where individuals have a high risk but have not yet been diagnosed with a disease (Figure S30). However, for 7 of the 8 traits, including T2D, we found the Integrated model to perform better than the Measurement model in terms of Expected Calibration Error (ECE) and Brier Score (BS) (Wilcoxon signed-rank test, one-sided, $P = 0.0078$ for both metrics) (Figure 3e), indicating more reliable probability estimates, the probability estimates being closer to the true labels and better uncertainty estimation. These results are particularly interesting for traits like Acute Myocardial Infarction, where the Integrated models and Measurement models had virtually the same ROC-AUC and MCC, but a 0.084 and 0.066 improvement in ECE and BS, respectively, when including genotype data (Figures 3e and S31). These results showcase the clinical potential of integrating genotype data with clinical measurements.

DISCUSSION

Here, by developing DL models specifically for large scale individual-level genotype data, we show that they can achieve competitive performance for a wide range of diseases. Particularly, they outperform our LASSO implementation for various autoimmune diseases which previously have been studied for interaction effects. By leveraging explainable AI, we confirm that our models identify disease relevant variants, showcasing that complex DL models are not strictly uninterpretable black boxes. We also show that our DL models can be trained on extremely high dimensional genotype data in a reasonable amount of time (32h for 8 traits) and they are robust to missing variants. Furthermore, we showcase the flexibility that DL architectures offer by training a single model to predict 338 disease traits at the same time with minimal loss in performance. Additionally, given that a predictor using the measurements can be trained on the same data as one using only genotype data, there is a clear advantage of including the measurement data for virtually all the diseases we considered. If including the measurements poses a data leakage risk and subsequent removal of samples, one must consider whether the trade-off in samples and additional features is acceptable. However, we see a strong indication that the inclusion of measurements outweighs the removal of samples for many disorders. Finally, we show that the inclusion of genotype data in addition to clinical and biochemical measurements provides better calibrated models, which is highly relevant for clinical decision making. The overall poor calibration however highlights the importance of examining it, especially when developing models for clinical applications. This is particularly important when using NNs, which tend be overconfident[66]. In those cases, it is common to calibrate predictions post-hoc with methods such as Platt scaling, temperature scaling or isotonic regression[67–69]. Additional factors to consider are uncertainty estimation and out-of-distribution detection, which post-hoc calibration methods are not guaranteed to solve. In this work we have considered data from individual-level cohorts, but it is straightforward to use the framework to integrate PRSs from predictors trained using summary statistics or genome-wide data, which is expected to improve performance even further. We believe that the development of accurate predictors that can model on various types of data, whether structured or unstructured, will be important for achieving precision medicine.

METHODS AND MATERIALS

Processing of UK Biobank genotype and clinical data

The genotype data was processed using Plink[70], version v1.90b6.10. After processing, the genotype data was converted to 459,576 one-hot encoded sample arrays of shape (4, 803,113) each, in the case of no quality control (NO-QC). The NO-QC approach might include signals from rare variants that otherwise

would be filtered out using a minor allele frequency threshold and previous work have shown negligible differences between kinship filtered and unfiltered sets of the UKBB data[71]. When applying quality control (QC), we used the following parameters in Plink `--maf 0.001 --geno 0.03 --mind 0.1` as well as removing samples with a kinship of more than 0.1. After applying QC there were 425,439 one-hot encoded sample arrays of shape (4, 662,143) each. X chromosomes were included in both cases. Unless otherwise specified, age, sex and the first 10 genotype principal components were included during training. For the tabular data, continuous columns were standardized using the training set statistics in all experiments, meaning that the values computed for the training set were applied to the validation and test sets. Missing values were imputed with the averages from the training set. Categorical columns were numerically encoded, missing values were marked as "NA" prior to numerical encoding. ICD-10 codes were used to derive the disease phenotypes. Only samples with a self-reported British, Irish, or any other Western European background were used for the experiments, which amounted to to 413,736 samples in training/validation sets and 45,840 in the test set in the NO-QC case. In the QC case, this resulted in 382,894 and 42,545 samples in the training/validation and test sets respectively. For the integration experiments, we removed samples where the the measurements used for integration were measured after a disease diagnosis. This was to avoid feature leakage, where e.g. a drug for a certain disease influences measurement values. An alternative could be to mark the measurements as missing and allow them to be subsequently imputed with the train set statistics. However, this might bias the case data towards having all of the measurements imputed, which the model might learn. Hence, it is not certain that such approaches would completely prevent feature leakage. Out of the 338 diseases, 48 did not have any time of diagnosis associated with them and we therefore excluded these from the integration experiments and analysis.

Training implementation and approach

All models were implemented using Pytorch[72], version 1.7.1. A held-out test set was used for all models in order to get a final performance after training and evaluating on train and validation sets, respectively. We used negative log likelihood loss during training for the classification tasks. All models were trained with a batch size of 64 except for the large multi task model (338 traits) which used a batch size of 32. During training, we used plateau learning rate scheduling to reduce the learning rate by a factor of 0.2 if the validation performance had not improved for 10 steps. The validation interval was calculated dynamically based on the number of cases for a given disease trait (C/B where C is case count and B is batch size, with thresholds of 100 and 2000 for the minimum and maximum intervals respectively), as was the number of validation samples used ($\max[10000, -1.5 \times C + 50000]$ where C is case count). We used early stopping to terminate training when performance had not improved for a certain number of validation steps. We used 16 and 20 steps for traits with less than and more than or equal to 2,500 cases, respectively. For the early stopping, we also used a buffer of a certain number of iterations before it was activated, using 1,000 iterations for the 8 trait benchmark and the MT experiments and 2,000 iterations for the large scale experiments. Weighted sampling with respect to the target variable was used in all runs during training. All models were trained with the Adam optimizer[73]. In the NN based models we used a weight decay of 1×10^{-3} with decoupled weight decay regularization[74]. All NN based models used a learning rate of 1×10^{-4}, while the LASSO models used a learning rate of 5×10^{-5}. We found that lower learning rate for LASSO gave better training stability and overall results. All neural network architectures used the SiLU[75,76] (also known as Swish[77]) activation function with a trainable parameter β inside the sigmoid function. When using weight decay, we did not apply it to the β parameter. For the neural network models, we augmented the input by randomly setting 40% of the SNPs as missing in

the one-hot encoded array, this is similar to input dropout[78] and we found it to be important to prevent overfitting in the NN models. For the LASSO we used L1 regularisation with $\lambda = 1 \times 10^{-3}$ for traits that had more than 2,500 cases and $\lambda = 1 \times 10^{-2}$ for traits that had less. All models were trained on a single 16GB NVIDIA® V100 Tensor Core GPU.

Architectures

This section details how the model architectures were implemented, which are broadly depicted in Figures S32 and S33. The MLP feature extractor was one FC layer with 10 output nodes. The main building blocks of the convolutional feature extractor were residual blocks with the first block using full pre-activation[79,80]. We added squeeze-and-excitation (SE) blocks[81] to the residual blocks, which we found both stabilized training and improved performance with minimal computational overhead. We used a dropout[78] of 0.5 between the convolutional layers in the residual blocks as recommended in prior work[82]. Before the residual blocks, the feature extractor used a single convolutional layer with a kernel size of (4, 39), a stride of (1, 10) along and 64 output channels. All the residual blocks used 64 input and output channels, a kernel size of either (1, 20) or (1, 19) and a stride of (1, 10) in the first convolutional layer and when downsampling the identity. The feature size after the convolutional blocks was 576, which went through `BN-ACT-FC` layers with a output feature size of 256. The feature extractor of the GLN model was similar to that of the CNN model where the main difference was that we used LCLs instead of convolutional layers, only two residual blocks instead of four and no SE blocks. In the first LCL, we used a kernel width of 8 (covering two SNPs per group) and 4 output sets and in the subsequent residual layers, we used a larger kernel width of 32 and 4 output sets. The final output dimension from the feature extractor was 396. The tabular feature extractor used in all models used embeddings for categorical inputs and left continuous inputs unchanged. The tabular inputs were concatenated and passed through a single FC layer. The fusion model aggregated the intermediate representations from the individual feature extractors by simply concatenating them. For the CNN and GLN NN predictors, we used the fused features from the fusion module as input and propagated them through FC residual blocks. For the CNN and GLN models, the predictors used four residual blocks with 256 nodes in the FC layers and a dropout of 0.5 between the FC layers. After the final residual blocks there was a `BN-ACT-DO-FC` which computed the final output for a given task. In the MLP case, we did not use residual blocks but rather a classic feed forward network. The intermediate representation from the fusion model was propagated thorough 5 sets of `BN-ACT-DO-FC` layers. Excluding the last, all FC layers had 256 output nodes. We used a dropout of 0.5 before the FC hidden layers.

Multi-task prediction

For each task, the NN predictor was a sequence of 4 residual blocks with FC layers composed of 256 nodes in the two and eight task models, but 64 nodes in the 338 task model. We used homoscedastic uncertainty to weight the loss contributions from individual tasks[83]. The technique we use for our MT learning is known in the as hard-parameter sharing, where all tasks share a subset of the model parameters throughout the entire training procedure. To examine how well the default GLN model performed in MT learning compared to other NN models, we compared it with a MLP model and a GLN based model using a Multi-gate Mixture-of-Experts (MGMoE)[84] as the predictor on the 8 benchmark traits. We found the default GLN model to perform best overall (Figure S34 and Data S13).

Framework

We present a framework that offers training with both linear models and various DL architectures on one-hot encoded omics and tabular input data, with the option of training on both continuous and categorical target variables. These models include linear and logistic regression models, MLP, CNN and GLN. Note that as the data is encoded in a one-hot format, the linear models should be able to detect some non-additive effects such as dominance. When modelling with different datasets of omics data, it is very common for datasets to differ in the number of genotyped SNPs. This can present a challenge when applying NN models designed for one dataset to another, as the network might have to be reconfigured to match the genotype dimensionality of the new dataset. To reduce the manual work that comes with such reconfiguration, the architecture of all models in the framework is automatically configured according to the size of the omics data so that no changes are needed whether the model has, for instance, 2,000 or 803,113 genetic variants as input. The framework attempts to strike a good balance between usability and flexibility by offering various model configurations that work with the automated architecture setup. For example in the case of CNN models, the number of channels, number of layers, kernel width, down stride and other hyperparameters can be configured when setting up the model. Users are not constricted to the predefined models however and can when using the Python interface define their own models which can be passed to the framework. Besides configuring model architecture, various other settings are available. For learning rate scheduling, the model offers various schedules such as cosine annealing with and without warm restarts and plateau scheduling. Additionally, the number of warm-up steps can be configured. When training predictive models on extremely high dimensional data, care must be taken to avoid overfitting. This is especially true when working with complex DL models, and hence the framework offers various regularisation strategies. These include weight decay, L1 regularisation, dropout and input dropout for omics data. Additionally, the framework offers data augmentation strategies in the form of Mixup[85] and two CutMix[86] variants, which are applied both to omics and tabular data when used. We found the mixing augmentations strategies as a promising way to reduce overfitting and increasing model generalization performance when applied to omics and tabular data (Figure S35 and Data S14). In addition, the framework offers weighted sampling with respect to arbitrary columns in the label file, which we found to greatly speed up training when applied to the target labels. The framework automatically extends to multi-task settings when more than one target is used for the modelling task. Additionally, the framework automatically combines information from omics and tabular data when using both as inputs. To gain insights into what the models learn during training, the framework offers automated activation computation and analysis for both omics and tabular inputs, using the SHAP framework[87]. We confirmed the models learning disease relevant variants and genes using LitVar and DisGeNet[88,89]. Lastly, the framework offers early stopping and saving of the model state that performed best on the validation set across the whole training run. The functionality to use an exported model to predict on a test set or other new samples is included in the framework.

ACKNOWLEDGMENTS

We thank Ingi Thor Sigurdsson for helpful discussion and contribution to the software. This research has been conducted using the UK Biobank Resource application 31823. A.I.S., D.W., S.R. and S.B. were supported by the Novo Nordisk Foundation (grants NNF18SA0034956, NNF14CC0001, and NNF17OC0027594). B.J.V. was supported by a Lundbeck Foundation Fellowship (R335-2019-2339) and by the Danish National Research Foundation (Niels Bohr Professorship to Prof. John McGrath), and by the Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH (R102-A9118,

R155-2014-1724 and R248-2017-2003). O.W was supported by by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (grant no. NNF20OC0062606). Computations described in this paper were performed using the National Life Science Supercomputing Center - Computerome at DTU and UCPH, www.computerome.dk.

AUTHOR CONTRIBUTIONS

S.R. conceived the study and guided the analysis with D.W., S.B., O.W., O.L. and B.J.V. A.I.S. wrote the software and performed the analyses. A.I.S, B.J.V. and S.R. wrote the manuscript with contributions from all coauthors. All authors read and approved the final version of the manuscript.

COMPETING INTEREST

S.B. has ownerships in Intomics A/S, Hoba Therapeutics Aps, Novo Nordisk A/S, Lundbeck A/S, and managing board memberships in Proscion A/S and Intomics A/S. The other authors declare no competing interests.

CODE AVAILABILITY

The code and binaries for EIR are available under the AGPL-v3 license and can be found at https://github.com/arnor-sigurdsson/EIR and https://github.com/RasmussenLab/EIR.

REFERENCES

1. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**, 1219–1224 (2018). URL http://www.nature.com/articles/s41588-018-0183-z.
2. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults. *Journal of the American College of Cardiology* **72**, 1883–1893 (2018). URL https://linkinghub.elsevier.com/retrieve/pii/S0735109718369493.
3. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics* **104**, 21–34 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S0002929718304051.
4. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**, 581–590 (2018). URL https://www.nature.com/articles/s41576-018-0018-x.
5. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Human Molecular Genetics* **28**, R133–R142 (2019). URL https://doi.org/10.1093/hmg/ddz187.
6. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine* **12**, 44 (2020). URL https://doi.org/10.1186/s13073-020-00742-5.
7. Albiñana, C. *et al.* Leveraging both individual-level genetic data and GWAS summary statistics increases polygenic prediction. *bioRxiv* 2020.11.27.401141 (2020). URL https://www.biorxiv.org/content/10.1101/2020.11.27.401141v1.
8. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. preprint, Genomics (2019). URL http://biorxiv.org/lookup/doi/10.1101/531210.
9. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. preprint, Genomics (2019). URL http://biorxiv.org/lookup/doi/10.1101/563866.

10. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). URL http://www.nature.com/articles/s41586-018-0579-z.
11. Pedersen, C. B. *et al.* The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Molecular Psychiatry* **23**, 6–14 (2018). URL http://www.nature.com/articles/mp2017196.
12. Saunders, G. *et al.* Leveraging European infrastructures to access 1 million human genomes by 2022. *Nature Reviews Genetics* **20**, 693–701 (2019). URL https://www.nature.com/articles/s41576-019-0156-9.
13. Vilhjálmsson, B. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics* **97**, 576–592 (2015). URL https://linkinghub.elsevier.com/retrieve/pii/S0002929715003651.
14. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* **41**, 469–480 (2017). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22050.
15. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications* **10**, 5086 (2019). URL http://www.nature.com/articles/s41467-019-12653-0.
16. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications* **10**, 1776 (2019). URL http://www.nature.com/articles/s41467-019-09718-5.
17. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2020). URL https://doi.org/10.1093/bioinformatics/btaa1029.
18. Privé, F., Vilhjálmsson, B. J. & Mak, T. S. H. lassosum2: an updated version complementing LDpred2. *bioRxiv* 2021.03.29.437510 (2021). URL https://www.biorxiv.org/content/10.1101/2021.03.29.437510v1.
19. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015). URL http://www.nature.com/articles/ng.3190.
20. Moser, G. *et al.* Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLOS Genetics* **11**, e1004969 (2015). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004969.
21. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018). URL https://academic.oup.com/bioinformatics/article/34/16/2781/4956666.
22. Qian, J. *et al.* A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genetics* **16**, e1009141 (2020). URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009141.
23. Li, R. *et al.* Fast Numerical Optimization for Genome Sequencing Data in Population Biobanks. *bioRxiv* 2021.02.14.431030 (2021). URL https://www.biorxiv.org/content/10.1101/2021.02.14.431030v1.
24. Nature. The scientific events that shaped the decade. *Nature* **576**, 337–338 (2019). URL http://www.nature.com/articles/d41586-019-03857-x.
25. Eraslan, G., Avsec, Z., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* **20**, 389–403 (2019). URL https://www.nature.com/articles/s41576-019-0122-6.

26. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface* **15** (2018).
27. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* (2021).
28. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* **10**, 390 (2019). URL https://www.nature.com/articles/s41467-018-07931-2.
29. Grønbech, C. H. *et al.* scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics (Oxford, England)* **36**, 4415–4422 (2020).
30. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* **37**, 420–423 (2019).
31. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020). URL https://www.nature.com/articles/s41586-019-1923-7.
32. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018).
33. Bellot, P., de los Campos, G. & Pérez-Enciso, M. Can Deep Learning Improve Genomic Prediction of Complex Human Traits? *Genetics* **210**, 809–819 (2018). URL https://doi.org/10.1534/genetics.118.301298.
34. Xu, Y. *et al.* Learning polygenic scores for human blood cell traits. *bioRxiv* 2020.02.17.952788 (2020). URL https://www.biorxiv.org/content/10.1101/2020.02.17.952788v1.
35. Mackay, T. F. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics* **15**, 22–33 (2014).
36. Clark, A. G. & Wang, L. Epistasis in measured genotypes: Drosophila p-element insertions. *Genetics* **147**, 157–163 (1997).
37. Elena, S. F. & Lenski, R. E. Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390**, 395–398 (1997).
38. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genetics* **4**, e1000008 (2008).
39. Aschard, H. *et al.* Inclusion of Gene-Gene and Gene-Environment Interactions Unlikely to Dramatically Improve Risk Prediction for Complex Diseases. *The American Journal of Human Genetics* **90**, 962–972 (2012).
40. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
41. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics* **47**, 702–709 (2015). URL http://www.nature.com/articles/ng.3285.
42. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. Preprint, Genetics (2019).
43. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020). URL https://www.nature.com/articles/s41586-020-2853-0.
44. Ngiam, J. *et al.* Tiled convolutional neural networks. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R. & Culotta, A. (eds.) *Advances in Neural Information Processing Systems*, vol. 23 (Curran Associates, Inc., 2010). URL https://proceedings.neurips.cc/paper/2010/file/01f78be6f7cad02658508fe4616098a9-Paper.pdf.
45. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. DeepFace: Closing the Gap to Human-Level

Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708 (2014). ISSN: 1063-6919.

46. Chen, Y.-h. *et al.* Locally-connected and convolutional neural networks for small footprint speaker recognition. In *Sixteenth Annual Conference of the International Speech Communication Association* (2015).
47. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288 (1996). URL https://www.jstor.org/stable/2346178.
48. Lenz, T. L. *et al.* Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nature Genetics* **47**, 1085–1090 (2015). URL http://www.nature.com/articles/ng.3379.
49. Wei, W.-H. *et al.* Major histocompatibility complex harbors widespread genotypic variability of non-additive risk of rheumatoid arthritis including epistasis. *Scientific Reports* **6**, 25014 (2016). URL https://www.nature.com/articles/srep25014.
50. Ivakhnenko, A. & Lapa, V. *Cybernetic Predicting Devices*. JPRS 37, 803 (Purdue University School of Electrical Engineering, 1966). URL https://books.google.no/books?id=l38DHQAACAAJ.
51. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* **36**, 193–202 (1980). URL https://doi.org/10.1007/BF00344251.
52. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).
53. Pociot, F. & McDermott, M. F. Genetics of type 1 diabetes mellitus. *Genes & Immunity* **3**, 235–249 (2002). URL https://www.nature.com/articles/6363875.
54. Wang, S., Flibotte, S., Camunas-Soler, J., MacDonald, P. E. & Johnson, J. D. A new hypothesis for type 1 diabetes risk: The at-risk allele at rs3842753 associates with increased beta-cell ins messenger rna in a meta-analysis of single-cell rna-sequencing data. *Canadian Journal of Diabetes* (2021). URL https://www.sciencedirect.com/science/article/pii/S1499267121000812.
55. Kim, H., Grueneberg, A., Vazquez, A. I., Hsu, S. & Campos, G. d. l. Will Big Data Close the Missing Heritability Gap? *Genetics* **207**, 1135–1145 (2017). URL https://www.genetics.org/content/207/3/1135.
56. Moutsianas, L. *et al.* Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nature Genetics* **47**, 1107–1113 (2015). URL https://www.nature.com/articles/ng.3395.
57. Strange, A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics* **42**, 985–990 (2010). URL https://www.nature.com/articles/ng.694.
58. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nature Genetics* **47**, 898–905 (2015). URL https://www.nature.com/articles/ng.3353.
59. Wang, M.-H. *et al.* Gene–gene and gene–environment interactions in ulcerative colitis. *Human Genetics* **133**, 547–558 (2014). URL https://doi.org/10.1007/s00439-013-1395-z.
60. Zhang, Y. & Yang, Q. A Survey on Multi-Task Learning. *arXiv:1707.08114 [cs]* (2018). 1707.08114.
61. Vandenhende, S., Georgoulis, S., Proesmans, M., Dai, D. & Van Gool, L. Revisiting Multi-Task Learning in the Deep Learning Era. *arXiv:2004.13379 [cs]* (2020). 2004.13379.
62. Wu, S., Zhang, H. R. & Ré, C. Understanding and Improving Information Transfer in Multi-Task Learning. *arXiv:2005.00944 [cs]* (2020). URL http://arxiv.org/abs/2005.00944. ArXiv: 2005.00944.

63. Sami, W., Ansari, T., Butt, N. S. & Hamid, M. R. A. Effect of diet on type 2 diabetes mellitus: A review. *International Journal of Health Sciences* **11**, 65–71 (2017). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426415/.
64. Shah, N. D., Steyerberg, E. W. & Kent, D. M. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA* **320**, 27–28 (2018). URL https://doi.org/10.1001/jama.2018.5602.
65. Van Calster, B. *et al.* Calibration: the Achilles heel of predictive analytics. *BMC Medicine* **17**, 230 (2019). URL https://doi.org/10.1186/s12916-019-1466-7.
66. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 427–436 (2015). ISSN: 1063-6919.
67. Platt, J. C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, 61–74 (MIT Press, 1999).
68. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 1321–1330 (PMLR, 2017). URL http://proceedings.mlr.press/v70/guo17a.html.
69. Zadrozny, B. & Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, 694 (ACM Press, Edmonton, Alberta, Canada, 2002). URL http://portal.acm.org/citation.cfm?doid=775047.775151.
70. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics* **81**, 559–575 (2007). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/.
71. Lello, L. *et al.* Accurate Genomic Prediction of Human Height. *Genetics* **210**, 477–497 (2018). URL https://doi.org/10.1534/genetics.118.301267.
72. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* **32** (2019). URL https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.
73. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2017). URL http://arxiv.org/abs/1412.6980. ArXiv: 1412.6980.
74. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *arXiv:1711.05101 [cs, math]* (2019). URL http://arxiv.org/abs/1711.05101. ArXiv: 1711.05101.
75. Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv:1606.08415 [cs]* (2020). URL http://arxiv.org/abs/1606.08415. ArXiv: 1606.08415.
76. Elfwing, S., Uchibe, E. & Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks* **107**, 3–11 (2018). URL https://www.sciencedirect.com/science/article/pii/S0893608017302976. Special issue on deep reinforcement learning.
77. Ramachandran, P., Zoph, B. & Le, Q. V. Searching for Activation Functions. *arXiv:1710.05941 [cs]* (2017). URL http://arxiv.org/abs/1710.05941. ArXiv: 1710.05941.
78. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15**, 1929–1958 (2014).
79. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016). ISSN:

1063-6919.

80. He, K., Zhang, X., Ren, S. & Sun, J. Identity Mappings in Deep Residual Networks. In Leibe, B., Matas, J., Sebe, N. & Welling, M. (eds.) *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, 630–645 (Springer International Publishing, Cham, 2016).
81. Hu, J., Shen, L. & Sun, G. Squeeze-and-Excitation Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141 (2018). ISSN: 2575-7075.
82. Zagoruyko, S. & Komodakis, N. Wide Residual Networks. *arXiv:1605.07146 [cs]* (2017). URL http://arxiv.org/abs/1605.07146. ArXiv: 1605.07146.
83. Kendall, A., Gal, Y. & Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7482–7491 (2018).
84. Ma, J. *et al.* Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, 1930–1939 (Association for Computing Machinery, London, United Kingdom, 2018). URL https://doi.org/10.1145/3219819.3220007.
85. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412 [cs, stat]* (2018). URL http://arxiv.org/abs/1710.09412. ArXiv: 1710.09412 version: 2.
86. Yun, S. *et al.* CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6022–6031 (2019). ISSN: 2380-7504.
87. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* **30** (2017). URL https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
88. Allot, A. *et al.* LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Research* **46**, W530–W536 (2018). URL https://doi.org/10.1093/nar/gky355.
89. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48**, D845–D855 (2020). URL https://doi.org/10.1093/nar/gkz1021.