Deep learning genome-wide genetics through biregular sparse decoder layer: deriving new disease genetic associations and disease risk predictions

Mika Gustafsson (✉ mika.gustafsson@liu.se)
Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden
https://orcid.org/0000-0002-0048-4063

Sanjiv Dwivedi
University of Gothenburg/ Chalmers https://orcid.org/0000-0003-3400-4133

Sandra Hellberg
Linköping University

Leonid Padyukov
Karolinska Institutet https://orcid.org/0000-0003-2950-5670

Rebecka Jornsten

Article

Keywords: genetics, Deep learning, autoencoder, long-range dependencies, genome-wide, complex disease risk

Posted Date: December 11th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3486573/v1

Additional Declarations: Yes there is potential Competing Interest. Computational resources were granted by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), including the National Supercomputer Centre (NSC, Berzelius-2021-26 and Berzelius-2022-156) and the High Performance Computing Center North (HPC2N, SNIC 2021/5-131 and SNIC 2021/22-199). This work is supported by the Swedish Research Council (Grant 2019-04193) and the Wallenberg AI, Autonomous Systems and Software Program (WASP) and SciLifeLab and Wallenberg National Program for Data-Driven Life Science (DDLS) (WASPDDLS21-040/KAW 2020.0239).

Deep learning genome-wide genetics through biregular sparse decoder layer: deriving new disease genetic associations and disease risk predictions

Sanjiv K. Dwivedi[1,3], **Sandra Hellberg**[1], **Leonid Padyukov**[2], **Rebecka Jörnsten**[3], **and Mika Gustafsson**[1,*]

[1]Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden
[2]Division of Rheumatology, Department of Medicine Solna, Karolinska Institutet, Karolinska University Hospital, SE-171 76, Stockholm, Sweden
[3]Division of Applied Mathematics and Statistics Mathematical Sciences University of Gothenburg/ Chalmers, Gothenburg, Sweden
[*]mika.gustafsson@liu.se

ABSTRACT

Genomic biobanks provide valuable resources for studying how genetic variation affects phenotypic differences and disease susceptibility, driving advancements in human health. While unsupervised deep learning frameworks like autoencoders have gained attention in other omics domains with fewer features, the computational complexity of genetic data has hindered their application. To overcome this challenge, we propose a sparse biregular decoder layer for deep genetic autoencoders. This layer significantly reduces the computational burden by using parameters at the scale of output nodes rather than multiplying output and input nodes, as in a fully-connected layer. We trained the genetic autoencoders efficiently by storing and computing data in the internal memory, both independently for each chromosome and collectively for all autosomal chromosomes, creating a trans-chromosomal autoencoder. This approach enabled the discovery of modular associations at the gene level in the interactome and facilitated the prediction of disease-associated single-nucleotide polymorphisms (SNPs) that yielded additional highly significant variants to pairwise linkage disequilibrium score. Furthermore, we demonstrated the effectiveness of deep autoencoders in enhancing the prediction performance for patients' genetic risk scores by incorporating their embeddings into a LASSO model. This improvement was replicated in independent multiple sclerosis (MS) and rheumatoid arthritis (RA) cohort datasets, highlighting the potential of deep autoencoders in personalized medicine.

Keywords: genetics, Deep learning, autoencoder, long-range dependencies, genome-wide, complex disease risk

Introduction

Through the acquisition of large biobanks in national populations, researchers can nowadays assess terabytes of digital large-scale data covering hundreds of thousands of patients and the corresponding phenotypes using resources like the UK Biobank[1]. Ideally, such data could revolutionize digital health and lead to a holistic view of medicine[2–6]. Although non-linear and long-range relationships can be learned, many approaches for this data have relied on simple and linear models that use the co-localization of neighboring sites on DNA strands, known as linkage disequilibrium (LD). In recent years, the deep learning (DL) framework has been proposed as a way to enhance the predictive power using available large-scale human genomics data and the computational resources. In theory, DL models can capture both nonlinear and long-range relations embedded in data given enough training samples, which potentially increase the generalized accuracy of the predictions. For example, DL can play a key role in resolving the common issue of unperceived epistatic interactions among gene variants that are limitation in the linear models.[7,8]. In fact, DL has already had a significant impact on various fields, like image recognition[9], sentiment learning[10] and is now also progressing into biological fields like protein structure prediction[11]. In these fields, researchers can access millions of samples, and while supervised machine learning models have a rather high accuracy, they heavily rely on man-made labels which can be sparse, biased, and noisy, resulting in unreliable models[12]. Unsupervised DL models can be more successful and generalizable as they avoid the inherent problems that come with predefined labels[13].

However, a key problem for DL is to find a good representation that is both highly accurate and functionally meaningful, making the representation generalizable to new problems[14]. In the genomics context, the large number of SNPs would, despite large sample size, still pose an infeasible problem as the number of variables would vastly exceed the number of samples. To address this challenge, prior knowledge associating nearby SNPs to similar features has been integrated structurally[15,16] and also by constraining non-linearities between gene co-localizing into similar pathways[16]. This approach has the advantage of creating models that are easier to interpret and constrains the representation domain substantially, making the predictions more robust. Further, these specialized layers heavily limit the search space and the memory usage in the training process.

Inter-chromosomal interactions, i.e., trans-effects, are generally regarded as weaker compared to so called cis-effects, which have a local effect to its own chromosome. However, despite weak effect size, trans-effects could cumulatively explain differences in the heritability of gene expression[17,18]. Furthermore, trans-effects offer the potential to go beyond the local regulatory machinery and provide important insights into the gene-regulatory network underlying complex diseases. In this work, we created a specialized layer allowing us to train sparse trans-genetic deep auto-encoders (trans-chromosomal deep AE), that efficiently decodes the cis and trans genetic variation integrated through the preceding layers. Further, we interpreted the learned representations from the trans-DAE using existing network biology and found it to code for a modular gene-gene representation. This was achieved using the *light-up* procedure, as previously introduces by us[19]. Next, we developed a method that translates the trans-DAE to derive the additional disease associated variants. We compared the trans-DAE derived methods with a more classical approach that associated SNPs to LD blocks and by that asssociated SNPs to diseases. Finally, we demonstrated the sustainable role of trans-DAE under the frame-work of transfer-learning via three methods in genomics. The two are based on first and second-layer compressed features. The third one utilises the discovered new additional variant features of trans-DAE with the disease variants of r^2 measure of LD using GWAS variants with their p-values. We showed an increase in disease classification accuracy in the combined scores of individuals via trans-DAE based and LD-based method. The key aspect among these methods is that the trans-DAE reduces the complexity of the UK-biobank data through its embeddings and also prioritizes disease variants. Moreover, we independently replicated our classification approach in two Swedish case-cohorts of patients with multiple sclerosis (MS) and rheumatoid arthritis (RA) respectively(Figure 1).

Results

A deep genome-wide AE more accurately predicts human genomic data compared to independent autosomal AEs

In order to create a reusable genetic association map that could both reduce the degrees of freedom in genome-wide analyses and capture nonlinear and long-range SNP relationships, we trained and tested autoencoders from different architectures (Figure 2). We evaluated their accuracy by examining how well they explained SNPs, using a minimum squared Pearson correlation of 0.3 on test data, as done in references[20,21]). For this purpose, we utilized the UK-Biobank comprising ≈500,000 individuals and 780,000 SNPs and to avoid expanding the feature representation too much, each SNP was coded as continuous variables using 0 for reference genome, 1 for heterozygous, and 2 for homozygous in alternative genome.

Since a deep fully connected AE would require a excessive number of parameters, we addressed this issue by employing specialized forms of locally connected first hidden layers. These layers link SNPs based on their proximity in chromosomal location while also identifying novel data-driven associations beyond traditional linkage disequilibrium. They account for a significant portion of SNP correlations, as shown in Figure S1 of the Supplementary Information. We initially performed an optimization strategy on each chromosome as one model, which supported that a sliding window of 500 SNPs to be locally connected followed by double-layered deep neural network (DNN), which supported that a chromosome could be compressed about 25-fold with good accuracy (Figure 2). These parameters were subsequently fixed, and we tested alterations using four different AEs with somewhat different architectures. First, three types of cis-chromosomal AEs were created, which could be trained for each chromosome (excluding the sex chromosomes) in parallel. These were a linear principal component-based AE (Figure 2a, hereafter referred to as PCA AE), a non-linear shallow AE (Figure 2b) and, a cis-chromosomal deep AE (Figure 2c). Moreover, we trained a deep trans-chromosomal deep AE (Figure 2d) that connected the different chromosomes together in the second layer, which allows for inter-chromosomal connections.

Comparing the accuracy of these four AEs showed that each of the non-linear AEs consistently outperformed the linear AE (i.e. PCA AE) for each chromosome respectively (Figure 3a). Interestingly, chromosome six showed the highest fraction of well-explained variants (Figure 3a). This finding can be attributed to the high variability resulting from the highly polymorphic nature of the HLA genes located at this chromosome. However, on average the difference in accuracy between cis- and trans-DAEs were rather small, which could be expected as most of this is well-described by heritage through LD blocks. Even so, the trans-chromosomal deep AE yielded 53.2% well-explained SNPs, which was higher than the cis-chromosomal shallow AE (52.2%) and the cis- chromosomal deep AE (51.1%), were each of these contained a set of about 5% state-variables compared to all the SNPs. Throughout the subsequent analyses we aimed to explore what new biological knowledge this unbiased data-driven representation could reveal and as the trans-chromosomal deep AE showed the least bias and seemed most promising, and we continued our analysis exploring this throughout the paper.

SNP co-localization in the auto-encoder associated network modular genes from different chromosomes

Next, we aimed to dissect the meaning of the co-association between SNPs with respect to being associated to the same latent node. For this purpose, we used node light-up starting from each node and thereby generating a ranked list of SNPs for each node. We then used six different ranking cut-offs, selecting the top 20, 40, 80, 160, 320, or 640 associated SNPs for each node. Subsequently, we analyzed the corresponding sets of mapped genes by assigning each SNP to its closest gene (see Methods). We further mapped those resulting sets on the protein-protein interaction network (using STRING database) and calculated the corresponding average path length between all the pair of genes in each set (Figure 3b). Interestingly, we found a significant lower average path length for each of the first three cut-offs than expected by chance (permutation $P < 10^{-16}$ for 80 SNPs or less). This suggests that those genes co-localise in the interactome into so called network modules. In order to see that this was not confounded by the construction of our first latent node layer that associated SNPs of the same chromosome to the same node, we repeated the prior association study by considering only genes at different chromosomes. Strikingly, we found minimal quantitative differences between the analysis (Figure 3b shows both types of interactions with almost identical results).Therefore, the observed co-localization extends beyond cis-interactions from SNPs within the same LD block and is a result of trans-genetic regulations.

Lastly, since our layer allows for inter-chromosomal associations, we aimed to determine the relative strength of these interactions. For this purpose, we calculated an interaction score by counting the co-appearance of SNP from a different chromosome as a top SNP at the output-layer while doing light-up of any of the SNPs as input layer from the given chromosome. This is summarised in Figure 3c.

Trans-DAE and DisGeNET SNPs identified new non-LD SNPs of ten diseases which were highly associated with disease using alternative GWAS data

In order to explore the potential disease etiology learned from the trans-DAE, we set up a test set for six diseases with the most samples (at least 18,000 UK Biobank patients), namely Atrial Fibrillation (AFIb), Asthma, Coronary Artery Disease (CAD), Hypertension (HTN), Myocardial Infraction (MI) and Type 2 Diabetes (T2D). This choice was motivated by our interest to classify diseases in our downstream analysis (see next section). For each disease, we used the seed SNPs collected from DisGeNET[22] and performed a light-up procedure starting from either the first or second layer of the trans-DAE to associate 200 alternative SNPs of the disease (see Methods). The main difference between these lies in that the first layered approach can propagate SNPs over long distances through a fully connected layer, while the second layered approach generates local associations.

To compare our results, we used LD-associated SNPs as a baseline model and tested the three new association lists for enrichment using separate genome-wide association studies (GWAS) with all genome-wide associated SNPs ($P < 5 \times 10^{-8}$; Figure 4a). First, comparing the enrichment in each disease, we found that in only two of the eight cases the LD approach scored higher than the 2nd layer approach (binomial test $P = 0.04$) and three cases for the 1st layered approach (non-significant) respectively. Moreover, we found non-significant differences between the different light-up layers, which often performed similarly. However, most of these differences were no significant yielding only three significant differences, namely for asthma, MS, and RA where both trans-chromosomal deepAE approaches outperformed the LD approach. The significance of first layer modules in case of Asthma, MS and RA are 1.19×10^{-3}, 1.42×10^{-52} and 2.10×10^{-26}. Similarly second layer shows for Asthma: 2.41×10^{-6}, MS: 4.48×10^{-50} and RA: 2.23×10^{-21}.

To determine the similarity between the identified SNPs of the three tested approaches we pairwise intersected them for each disease and computed the Jaccard index. This analysis showed that our both trans-chromosomal deepAE based approaches differed substantially from the LD approach in all cases. Interestingly, this was true also in the cases of CAD and HTN (Figure 4b) where the LD approach were mostly enriched for disease genes. As all approaches showed significant results, we proceeded our analysis of the predictive power of the trans-chromosomal deep AE using the union of these three methods, which led to the identification of 12-137 SNPs (mean 65.875) per disease (median = 48). Furthermore, our method exhibited similar predictive abilities for disease-associated SNPs in other cis-type autoencoders with non-linear activation, as demonstrated in the trans-chromosomal autoencoder (Supplementary Figure S2).

Combining latent representations with SNP-SNP relations increases predictive power

Lastly, we aimed to test whether the learned representation also could enhance disease classification accuracy using the previously analysed six diseases and in addition also two independent autoimmune case-control cohorts of RA and MS cohorts with about 9,000 indivduals each. To achieve this, we associated the SNPs obtained from three different approaches: (1) LD-based approach, (2) selecting SNPs using the trans-DAE, and (3) a consensus approach that combined scores from trans-DAE related methods. We then utilized a machine learning model (LASSO regression) to predict the phenotypic labels using subsets of these SNPs as well as age and sex as features. We performed 10-fold cross-validation and evaluated the results using the area under the precision recall curve (AUC) on the unseen 10% of the data. Comparing the two SNP-based approaches and the two hidden node-based approaches, we found that the differences were often non-significant, whereas in fact, the consensus approach scored higher than each individual method in all eight cases (binomial $P = 2.5 \times 10^{-6}$) (Figure 5).

To further analyze the significance, we performed paired t-tests on the scores of the ten folds in the cross-validation. We found that the DAE approach was significantly better than the LD based method in all cases except RA. Moreover, the consensus method demonstrated superior performance compared to both the LD-based and DAE-based methods. Notably, highly significant results were observed for the well-powered UK-Biobank diseases (Table 2).

In summary, these results suggest that the DAE and LD approaches identify complementary SNPs, and their predictive power can synergize with each other. This finding opens up new possibilities for enhanced predictive performance in GWASs.

Discussion

Although vast genomics data repositories exist, genetic association studies mainly use linear models and focus on blocks of cis-associated SNPs. This reliance on simpler methods stems from the computational challenges and the complications of managing more parameters when exploring associations outside cis regions to pinpoint disease associated genes. In this paper, we introduce an innovative customized sparse layer designed for the training of trans-chromosomal deep-AEs. Utilizing the UK-Biobank data, we juxtaposed this with simpler cis-AEs, observing a reduced reconstruction error. Notably, we detected a co-localisation pattern in the protein-protein interaction network between SNPs associated with trans-genetic SNPs and the same hidden node. This pattern emerged when illuminating from the initial layer and through our specialized decoder layer. Though it might be tempting to equate this pattern with functional relevance, it's crucial to note that association alone doesn't confirm this. Nonetheless, our constraints led us to a representation that, when applied, revealed new disease-associated SNPs. Validation using alternative GWASs for the same diseases indicated that our trans-chromosomal deep-AE predictions recognized SNPs that other research studies corroborated. Furthermore, these predictions showed greater enrichment than traditional LD-based techniques in most diseases. Additionally, we delved into the potential of the extracted representations for transfer learning, aiming to predict patient outcomes with enhanced precision. The results demonstrated that integrating established disease-associated SNPs with our derived representation consistently increased the cross-validation classification accuracy. However, the specific criteria for determining "higher accuracy" need more explicit definition. In essence, our comprehensive genetic representation offers a data-driven, functional, and discernible method, invaluable in transfer learning. This approach not only paves the way for uncovering new genetic links but also fortifies supervised learning endeavors.

Customized layers play a crucial role in incorporating biological knowledge into the learning process, providing a means to constrain the parameter space. This constraint increases the likelihood of model identifiability and improves generalization capabilities. In the context of linkage disequilibrium blocks and inheritance insights, the first layers of a model are naturally structured as a locally connected graph, while subsequent layers interconnect these entities. For instance, van Hilten et al.[15] employed this approach to aggregate gene-level statistics, which were then combined in subsequent layers based on pathway knowledge, ultimately enabling phenotype prediction. Mocanu et al. used topological sparsity in supervised DNN without compromising predictive performance[23]. However, this approach necessitated significant memory allocation, limiting its application to genomic data due to the memory requirements being proportional to the product of nodes in consecutive layers. Our proposed trans-chromosomal deep AE drastically reduces the number of parameters stored in RAM. It efficiently leverages LD blocks as layers, which are subsequently combined in subsequent layers while storing only the necessary parameters in internal memory. However, introducing multiple intermediate fully connected layers could enhance the capture of nonlinearity, albeit at the cost of making it more challenging to interpret the resulting deep autoencoder. Despite this complexity, we have successfully demonstrated the ability to associate new SNPs with their corresponding seeds.

We observed that genes proximal to our identified SNPs often co-localized within the protein-protein interaction network, suggesting a functional gene relationship. This is consistent with earlier genetic studies on gene networks related to complex diseases and aligns with our prior work on transcriptomic deep autoencoders, where a similar co-localization with protein-protein interaction networks was noted[19].

While deep learning often faces interpretability challenges, our trans-chromosomal deep AE provides insights. The hidden layer nodes directly relate to the genomic coordinates of the variants, elucidating cis- and trans-chromosomal relationships among genomic regions linked to variant blocks. While adding multiple intermediate layers captures more nonlinearity, it complicates the deep AE's interpretability. Nevertheless, we effectively linked new SNPs with their corresponding seeds. Our results emphasize the improved interpretability of our model and suggest its applicability to other omic-trained autoencoders[24].

Complex diseases are influenced by a web of interactions between genetic factors and lifestyle/environmental risk factors, presenting challenges in diagnosis and treatment outcomes. This intricate nature limits the predictive accuracy of disease modeling. It's vital to incorporate significant covariates like BMI and smoking to discern genuine disease associations[25]. It's worth noting, as a limitation of our study, that the disease variants identified through our methodology are not influenced by exposure factors related to the phenotype label. Integrating our approach with these risk factors could yield more comprehensive results, improving disease gene mapping. For disease risk prediction, however, we rely on label information. Consequently, age and gender were incorporated as input features in our LASSO penalized logistic regression.

Phenotypic labels in complex diseases can often be unclear due to overlapping diagnoses that can vary over time, unlike genetics, which remains consistent. In addition, the genetics feature vectors often overlook important counterparts, for instance environmental factors making significant differences in similar magnitude to genetics. Nevertheless, most DNN approaches are trained directly supervised to predict these labels, which indeed should significantly affect their generalizability[26]. In contrast, unsupervised machine learning, though less explored, has the potential to generate more robust features. They reflect a compromise by predicting multiple labels simultaneously, thereby putatively reflecting the genetic state of the organism. Our presented approach highlights various aspects of the unsupervised deep learning landscape, for instance in navigating genetic imputation methods including both cis- and trans-chromosomal variants relationships. This could be advantageous over the existing tools that solely utilize the cis relationship[27].

Our proposed transfer learning method offers a benchmark for genetic disease risk prediction. This model efficiently compresses independent data, prioritizes disease variants, and facilitates diverse machine learning strategies, such as semi-supervised learning across varied human populations. We've validated its potency using lasso and consensus score-based techniques, highlighting enhanced predictive accuracy. Traditional statistical algorithms, such as those detailed in[28], often remain ethnicity-specific. These typically use additive models that focus solely on the contributions of independent risk alleles, neglecting non-linear interactions and prevalent alleles[29–32]. Our methodology streamlines genome-wide data, making machine learning applications more accessible. By transforming genome-wide SNP counts into compact representations via the trans-chromosomal deep AE, we achieve more efficient and effective genetic analyses.

High-throughput molecular measurements characterise multi-view cellular states, offering the opportunity to integrate diverse data into a unified framework. This integration can boost our understanding of how genetic components contribute to functional variations and causing disease. However, implementing such holistic approaches can be computationally complex, potentially more even than our current study. Another challenge lies in the lack of labels, both technically and biologically, associated with the data. By leveraging the capabilities of trans-chromosomal deep AE on large-scale data, we can harness the potential of unlabeled data. This allows us to learn the specific data manifold with intrinsic dimensions, enabling wider application of semi-supervised learning to improve prediction performance. Additionally, the trans-chromosomal deep AE manifests cis and trans DNA variant relationships throughout the genome, reducing variables and enabling simplified models to incorporate the nonlinear impact on gene expression or chromatin accessibility in different human cell types.[33]. Interpretation methods based on deep learning models can further uncover the functional implications of the DNA sites and their related variants[34].

Although multi-omic data is inherently heterogeneous, incorporating prior knowledge can help overcome incompatibilities by linking related features[35]. This approach can greatly enhance the relevance of such data to our methodology. Through training these models, we can capture personalized molecular-level physiological information and represent it in a compressed form[36]. By compressing large feature spaces into latent spaces, these models can better accommodate multi-tasking and decode shared representations for regression, stratification, or classification purposes.

Furthermore, as the availability of large genotype data and related environmental factors continues to advance, the future holds the potential for extensive patient care data paired with genotype profiles. This will significantly enhance our understanding of physiological processes, development, diseases, and therapeutic approaches, particularly through the utilization of single-cell data[37]. By combining patient-specific clinical measurements, our approach holds the potential to increase the relevance of artificial intelligence in healthcare applications, paving the way for direct implementation in patient care[38].

Methods

Human Genotype data

We accessed UKB Application 43117 data in plink binary format, there were 488,377 individual's profile consisting of a total 784,256 autosomal chromosome variants. Each variant's value for an individual could be 0, 1, or 2, meaning homozygosity for the reference genome, heterozygosity, or homozygosity for the alternative genome, respectively. UK Biobank contains many ethnicity's but as our study aimed at fining a unifying latent representation we ignored such differences and allowed there putative effects being potentially included in the latent representation. The genetic profile matrix, denoted as G with m variants and n individuals, can be represented as $[g_{ij}]_{m\times n}$, where g_{ij} can take values from the set 0, 1, 2. Since the majority of entries in the matrix are zero values (71.63%), we replaced the missing values of G with 0.

Similarly, we have two national Swedish case-cohort data for MS and RA, comprising 5,566 cases and 5,615 controls for MS, and 4,071 cases and 2,952 controls for RA. To ensure compatibility with the UKB data, we performed imputation on the Swedish cohort data, resulting in an increase from 100,556 (MS) and 237,856 (RA, pre-imputation) variants to a total of 692,963 imputed variants. We utilized 1000 genome project data as a reference panel to impute the data. For MS and RA, we chose a sliding-window approach with window lengths of 20 centimorgans (cM) and 15 cM, respectively, and a shifting size of 2 cM. The imputation was carried out using the beagle 5.4 package (version: 22Jul22.46e). We used plink and bcftools software to handle file format conversion and merge the data accordingly.

Building and training autoencoders

The initial approach involved constructing an autoencoder for each autosomal chromosome, which was designed to fit within the available RAM sizes of our computational resources. This type of autoencoder is referred to as a cis-chromosomal autoencoder. Three variations of cis-chromosomal autoencoders were constructed: a PCA-based shallow autoencoder, a nonlinear shallow autoencoder, and a nonlinear deep autoencoder. The construction methods for each type are described below (Figure 2):

PCA-based shallow Autoencoder. For each autosomal chromosome, we computed covariance matrices of the genetic data after subtracting the mean across the training samples. In order to make it comparable with the corresponding non-linear autoencoder, we selected a fixed number of eigen vectors, denoted as p, which is the same as the number of hidden nodes. These eigen vectors $[v_1, v_2, v_3, \dots, v_p]$ corresponded to the top p leading eigenvalues of the covariance matrix using Rspectra[39]. These eigenvectors were arranged as columns in a matrix E, which serves as the encoder part of the PCA-based AE (Figure 1c). The encoder and decoder parts are defined as follows:

$$Y = EG - E\langle G\rangle \tag{1}$$

The decoder part,

$$Y = YE' + \langle G\rangle \tag{2}$$

Here, E' denotes the transpose of matrix E.

Nonlinear cis-Shallow Autoencoder. This type of autoencoder consists of a single hidden layer with sigmoid activation functions in all layers for each chromosome (Figure 1d) .

Nonlinear cis-Deep Autoencoder. In this approach, we constructed a deep autoencoder using a locally connected layer followed by three standard dense layers from the Keras library. All nodes in the deep autoencoder have sigmoid activation functions. The kernel size and shifting parameters in the locally connected layers are 500 and 10 respectively. First layer learns the nearby variants LD correlations on the genome. These autoencoders also capture only the variants relations withing the chromosomes. To account for both cis and trans chromosomal variant relations, we created a customised decoder layer for trans-chromosomal deep AE and explained as follows.

trans-chromosomal deep AE. There are cis and trans variant relations within and across the chromosomes. To analyze this, we combined autosomal chromosome data from chr 1 to chr 22. The autoencoder trained on the combined chromosomal data would be a preferred approach, as it allows for the representation of personalized genetic content in a highly reduced dimensional space. To make the training feasible within our available computational resources, we created a customised decoder layer with sparse bi-regular graph structure. Using this decoder layer, we built a trans-chromosomal deep AE (Figure 2e). The trans-chromosomal DAE consists of two standard Keras layers: the first layer being locally connected, followed by a dense layer, and the last one being our customized locally connected layer. In the standard locally connected layer, we selected kernel size and shift parameters of 500 and 25, respectively. The connecting method employed in the last hidden and output layer ensures that each node covers pq nodes in the output layer. To cover all the nodes in the output layer ($n_{L3} = 784256$), we selected the number hidden nodes, n_{L2} to be 49016, such that $n_{L3} = n_{L2} \times p$, for our case $p = 16$. The parameter $q = 100$ determines the counts of connections in each node of the output layer. In this set up each node in L_2 covers a window of $p \times q$ nodes in L_3 and the next hidden node slides down-side in the range of p nodes at L_3 layer. To keep the boundary symmetry, we consider a circular way where before the top node, the most down nodes appear in the L_3 layer.

In other words, suppose $W_{L3} = [w_{ij}]$ be the sparse weight matrix of the last layer, where the zero elements are not involved in the training process. The trainable w_{ij} exists if the i^{th} node of the last hidden layer is connected to the j^{th} node in the output layer. The connectivity of the i^{th} node with the j^{th} node is determined as follows:

$$(p-1)k < j \leq pk \tag{3}$$

$$p = \bigcup_{d=0}^{q-1} \{x : x = \mod(H - q + d + i, H)\} \text{ for i = 1,2,3,...H} \tag{4}$$

Interpreting the trans-chromosomal deep AE

Decoding trans-chromosomal deep AE with PPI

To interpret the genome-wide SNP-SNP relations captured in the trained model, we calculated the influence of each node in the first hidden layer on the output layer. Let's denote the weight matrices and bias term in the k^{th} layer as w_{ij}^{k} and b^{k}, respectively. The score for a i^{th} node can be computed using the following equation:

$$u_i^1 = S(w_{ij}^3(S(w_{ij}^2 I_1 + b^2) + b^3)) \tag{5}$$

Where $S(x) = \frac{1}{1+e^{-x}}$. Here, I_1 represents an identity matrix with the same dimension as the number of hidden nodes in the first layer. The weight matrices are denoted as w_{ij}^{k} and bias terms as b^k in k^{th} layer.

Next, we proceeded to compute the average distance in protein-protein interaction (PPI) of the genes closest to the SNPs. These SNPs were ranked using light-up method[19] from the first hidden layer of the trans-chromosomal deep AE. The closest genes to the top-ranked SNPs, which exhibit biases higher than the 88th percentile and lower than the 99th percentile at the output layer, form the PPI module.

Decoding chromosomal relations of trans-chromosomal deep AE. We made some modifications to the trans-chromosomal deepAE. Firstly, we replaced all sigmoid activations with linear activations. Additionally, we substituted the bias terms in each layer with zero vectors of same dimensions. In order to map chromosomal relations, we activated each variant with 1 , while keeping rest input nodes as zeros, for each chromosome at the input layer and then counted the top 50 related chromosomes at the output layer. It leads to a chromosomal relation matrix. The counts less than 10 were replaced with zeros. Since we need only triangular part to visualise, therefore we replaced lower triangular part of chromosomal relation matrix with the upper triangular part of the matrix.

To establish chromosomal relations, we implemented a specific activation scheme. At the input layer, we activated each variant by setting it to 1, while keeping the remaining input nodes as zeros for each chromosome. At the output layer, we counted the top 50 related chromosomes. This process generated a chromosomal relation matrix.

To enhance visualization, we focused on the triangular part of the chromosomal relation matrix. Since we only needed to visualize this portion, we replaced the lower triangular part of the matrix with the upper triangular part. This adjustment simplified the representation while retaining the essential information.

Associating disease variants

We define disease score of a variant for a disease using the trans-chromosomal deep AE and corresponding the curated variants from DisGeNET (Enrichment analysis):

LD based method for enrichment. The disease genetic variants form haplotype blocks and the linked variants in such blocks are likely to be disease associated[40,41]. We computed LD measure r^2 for pairwise SNPs on the UKB data using PLINK software by selecting a block of 20000 variants: it approximate count of SNPs available in shortest autosomal chromosome. The disease score of a variant is its mean value of r^2 over all the seed DisGeNET variants.

Autoencoder based method for enrichment. First, we define a matrix of a h^{th} layer associated for an i^{th} variants at the output layer, v^h

$$v^h = |w_{ij}^{h}||w_{ij}^{h+1}|..w_{ij}^{L} \tag{6}$$

w_{ij}^{h} is a weight matrix of h^{th} layer, L is the total number of layers. The disease score for a k^{th} SNP,

$$s_k = \frac{1}{p}\sum_{i=1}^{p} D_c(v_k^h, v_i^h) \tag{7}$$

Where $D_c(v_k^h, v_i^h)$ is cosine distance between v_k^h and v_i^h, which are the k^{th} UK-biobank variant and i^{th} DisGeNET variant of the matrix v^h for a disease respectively. It measures the sites varying with disease SNPs having a certain level similarity with other unknown disease SNPs, since only the weight vectors from the related weight matrix leads with their variation while prediction the patients SNPs from their compressed representation. Here, we fixed $p = 5$, for localised SNPs giving top score for each k^{th} SNP. $p \neq 1$ since that manifest the contribution of single disease SNP which is likely to be affected with noise. Also, we avoided the mean over large number of the disease SNPs since such unknown disease SNPs are likely to have the contribution in their score of many localised disease locations of the genome, though it should be from one location, which may make it noisy.

Predicting disease related individuals

Our classification method consists of two steps (A) feature construction methods and (B) supervised classification using lasso. We used nominally significant GWAS SNPs (P<0.05) with their p-values as seed SNPs. This cut-off was used to search for enriched latent variables where the seed signals were enriched which subsequently were aggregated for corroborative effects. For each SNP we defined the following scores At We defined following methods at SNP level: **Trans-chromosomal deepAE based SNPs feature.** We computed two types for disease modules: first (h = 1) and second (h = 2) hidden layers of trans-chromosomal autoencoders. The revised the above disease SNPs scoring equation including p-values, $P_i^{\prime}s$ contributing in a same direction of the the disease module score, as : The disease score for a k^{th} SNP,

$$s_k = -\left(\frac{1}{log(0.05)} + \frac{1}{log(P_i)}\right) \frac{1}{p} \sum_{i=1}^{p} \left(0.5 + D_c(v_k^h, v_i^h)\right) \tag{8}$$

Here, we fixed $p = 5$, for those seed SNPs giving top score for k^{th} SNP which account take the mean effect for only localised disease SNPs along the genome. The constant terms in the both brackets balance the contribution of the variable terms to each other from when any or both the variable term(s) tend towards zero. We selected top 12000 SNPs for each uk-biobank disease that is optimal to performance in the classification for available samples. The presented results are robust to two-fold variations in these proposed hyperparameters.

LD based SNPs feature. We define LD score based variants to select the features as our baseline classification model. The disease score of a variant is its mean value of $-r_{LD}^2(log(0.05) + log(P_i))$ over all the significant GWAS SNPs ($P < 0.05$), the constant term $log(0.05)$ is added to balancing the continuation for r_{LD}^2 with p-values as we took only the significant GWAS SNPs.

LD based disease classification. We prioritised the variants as above LD based SNPs feature score. This method we call as our baseline method.

Disease module for classification. To select the features for classifying the patients, we selected the union of 12000 SNPs from each of the two types of above trans-chromosomal deepAE based SNPs feature, and 12000 from the LD based module. We keep same number of features in the base line method as disease module.

Compressed features using trans-chromosomal deep AE. The two further types of classification methods are based on its (c) first and (d) second layer compressed representations. Filtered compressed features using module: The classification accuracy is dependent on feature sample ratio. In the case that we do not have sufficient disease samples (independent MS and RA data sets), we prioritize the compressed features using the disease module for each disease. The score for classification for hidden node is the

$$s^h = |[w_{ij}]|^h |[w_{ij}]|^{h+1} w_{ij}^L d \tag{9}$$

Where d is the vector of dimension as the number of SNPs is UK-biobank with 1 , if SNPs is disease module otherwise 0.

Supervised classification of uk-biobank disease samples. We selected a binary classification approach, so all the UK-biobank samples were divided into two classes, case samples and the rest of all the samples except case samples, for each disease. We presented the results for diseases having about 20K samples in UK-biobank data. Each of the four types of methods define our feature selection criteria, as described above, for corresponding four classification methods namely, LD based baseline, our proposed method: module, first layer and second layer compressed representations of trans-chromosomal deep AE as described above in feature construction method (a), (b), (c) and (d). AUC is computed under 10-fold cross-validation after including age and gender as two more features, for the optimal lambda parameter of lasso in the sklearn-python library (version 0.23.2). Our consensus model based method took the mean score of the four methods on a test data set for each individual sample.

Supervised Classification of MS and RA independent samples from Swedish cohort data. We computed two hidden layers compressed representations and the same way defined modules trans-chromosomal deep AE and tested the application of transfer-learning to classify diseases and re-use of training independently on MS and RA in Swedish cohort data. We also included the two additional features age and gender as input to the LASSO-model after imputing the missing values of age with mean value. In order to increase the prediction performance we included the training part UK-biobank case and control samples were included in only each of the 10-fold training Swedish data sets. In order to have a similar age distribution of the UK-biobank samples, we subtract their mean for the both case and control samples and add their related mean from Swedish cohort data.

References

1. Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
2. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature* **562**, 210–216 (2018).

3. Hao, Z., AghaKouchak, A., Nakhjiri, N. & Farahmand, A. Global integrated drought monitoring and prediction system (GIDMaPS) data sets. *figshare* http://dx.doi.org/10.6084/m9.figshare.853801 (2014).

4. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. genetics* **50**, 906–908 (2018).

5. Naito, T. *et al.* A deep learning method for hla imputation and trans-ethnic mhc fine-mapping of type 1 diabetes. *Nat. communications* **12**, 1–14 (2021).

6. Zhao, B. *et al.* Common genetic variation influencing human white matter microstructure. *Science* **372**, eabf3736 (2021).

7. AlQuraishi, M. & Sorger, P. K. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nat. methods* **18**, 1169–1180 (2021).

8. Badré, A., Zhang, L., Muchero, W., Reynolds, J. C. & Pan, C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J. Hum. Genet.* **66**, 359–369 (2021).

9. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).

10. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing. *ieee Comput. intelligenCe magazine* **13**, 55–75 (2018).

11. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

12. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genet. medicine* **17**, 405–423 (2015).

13. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

14. Detlefsen, N. S., Hauberg, S. & Boomsma, W. Learning meaningful representations of protein sequences. *Nat. communications* **13**, 1–12 (2022).

15. van Hilten, A. *et al.* Gennet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun. biology* **4**, 1–9 (2021).

16. Sigurdsson, A. I. *et al.* Deep integrative models for large-scale human genomics. *bioRxiv* (2021).

17. Cheung, V. G. *et al.* Polymorphic cis-and trans-regulation of human gene expression. *PLoS biology* **8**, e1000480 (2010).

18. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034 (2019).

19. Dwivedi, S. K., Tjärnberg, A., Tegnér, J. & Gustafsson, M. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nat. communications* **11**, 1–10 (2020).

20. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. epidemiology* **34**, 816–834 (2010).

21. Kim, Y. J., Lee, J., Kim, B.-J., Consortium, T.-G. & Park, T. A new strategy for enhancing imputation quality of rare variants from next-generation sequencing data via combining snp and exome chip data. *BMC genomics* **16**, 1–11 (2015).

22. Piñero, J. *et al.* The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research* **48**, D845–D855 (2020).

23. Mocanu, D. C. *et al.* Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nat. communications* **9**, 1–12 (2018).

24. Kulminski, A. M. Complex phenotypes and phenomenon of genome-wide inter-chromosomal linkage disequilibrium in the human genome. *Exp. Gerontol.* **46**, 979–986 (2011).

25. McCaw, Z. R. *et al.* Deepnull models non-linear covariate effects to improve phenotypic prediction and association power. *Nat. communications* **13**, 1–10 (2022).

26. Bengio, Y. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, 17–36 (JMLR Workshop and Conference Proceedings, 2012).

27. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *The Am. J. Hum. Genet.* **108**, 1880–1890 (2021).

28. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. genetics* **50**, 1219–1224 (2018).

29. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).

30. Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).

31. Dias, R. & Torkamani, A. Artificial intelligence in clinical and genomic diagnostics. *Genome medicine* **11**, 1–12 (2019).

32. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. genetics* **54**, 573–580 (2022).

33. Gazal, S. *et al.* Combining snp-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).

34. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153 (PMLR, 2017).

35. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).

36. Wong, A. K., Sealfon, R. S., Theesfeld, C. L. & Troyanskaya, O. G. Decoding disease: from genomes to networks to phenotypes. *Nat. Rev. Genet.* **22**, 774–790 (2021).

37. Argelaguet, R., Cuomo, A. S., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. biotechnology* **39**, 1202–1215 (2021).

38. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. medicine* **25**, 44–56 (2019).

39. Qiu, Y., Mei, J., Guennebaud, G. & Niesen, J. Rspectra: solvers for large scale eigenvalue and svd problems. *R package version 0.16-0* (2019).

40. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. heredity* **56**, 18–31 (2003).

41. Slatkin, M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).

Acknowledgements (not compulsory)

Computational resources were granted by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), including the National Supercomputer Centre (NSC, Berzelius-2021-26 and Berzelius-2022-156) and the High Performance Computing Center North (HPC2N, SNIC 2021/5-131 and SNIC 2021/22-199).

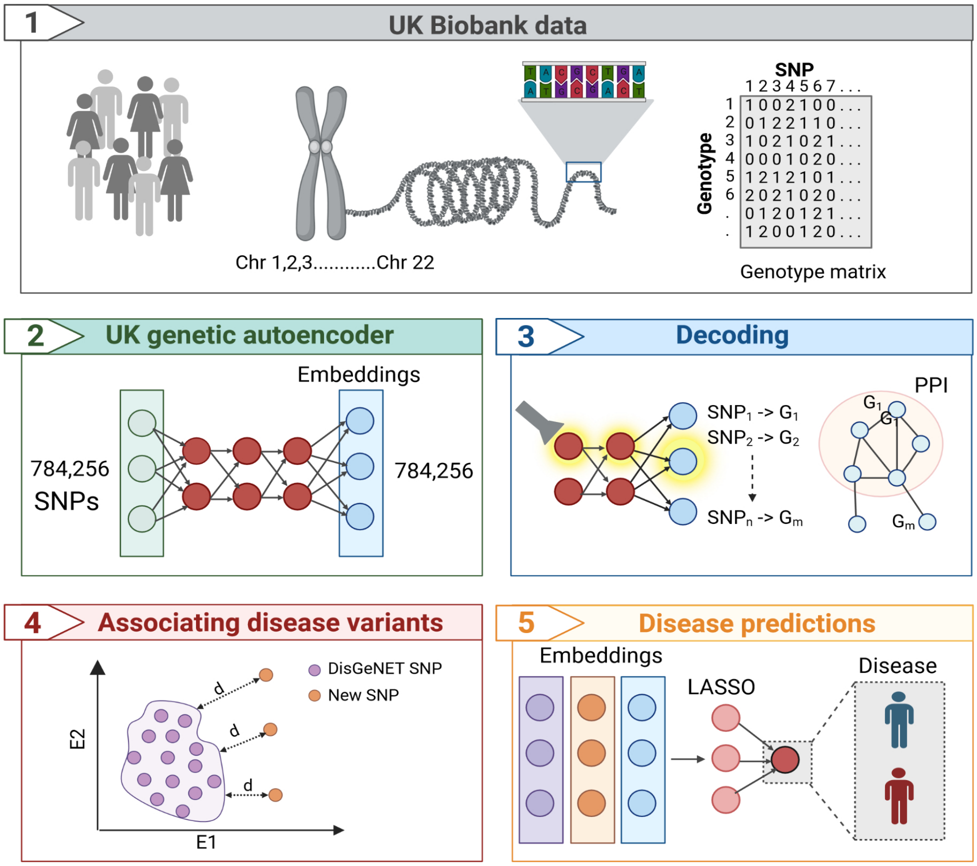
This work is supported by the Swedish Research Council (Grant 2019-04193) and the Wallenberg AI, Autonomous Systems and Software Program (WASP) and SciLifeLab and Wallenberg National Program for Data-Driven Life Science (DDLS) (WASPDDLS21-040/KAW 2020.0239).

Author contributions statement

M.G. and S.K.D. conceived the study. S.K.D. performed deep learning training and analysis which were supervised by MG with inputs from S.H., L.P. and R.J. All authors contributed in writing and approval of the final draft for publication.

Disease	P-value for module method w.r.t. LD method	P-value for All method w.r.t. LD method	P-value for All method w.r.t. module method
CAD	0.01251*	0.002797**	5.188e-05****
Asthma	0.001664**	1.285e-09****	4.604e-07****
T2D	5.119e-06****	1.231e-07****	1.675e-10****
ATR	0.0009817***	3.241e-06****	1.097e-05****
HTN	0.03231*	7.283e-13****	2.846e-12****
MYO	0.0004017***	3.387e-06****	0.0002558***
MS	0.0004034***	0.0001068****	0.009082**
RA	0.6926	0.007895**	0.02509*

Table 1. Significance scores in AUC performance of various diseases for all the possible pair using t-test among LD, module and combined score based method.



Genotype \ SNP	1	2	3	4	5	6	7	...
1	1	0	0	2	1	0	0	...
2	0	1	2	2	1	1	0	...
3	1	0	2	1	0	2	1	...
4	0	0	0	1	0	2	0	...
5	1	2	1	2	1	0	1	...
6	2	0	2	1	0	2	0	...
.	0	1	2	0	1	2	1	...
.	1	2	0	0	1	2	0	...

Figure 1. Schematic diagram of interpreting an genetic autoencoder, defining the disease variants and classifying patients. 1, Genotype data from the UK Biobank, comprising 500,000 individuals, was used to test and train autoencoders with different architectures. **2,** Genetic autoencoder . **3 ,** Replicating the co-association of localization of genes in protein protein interaction with associated variants at the hidden layer of genetic autoencoder . **4,** describe the disease SNPs defining method using DisGeNET SNPs using cosine similarity based distance metric. **5,** LD, trans-chromosomal deep AE generated various embedding are used as input to enhance the patient classification using LASSO method

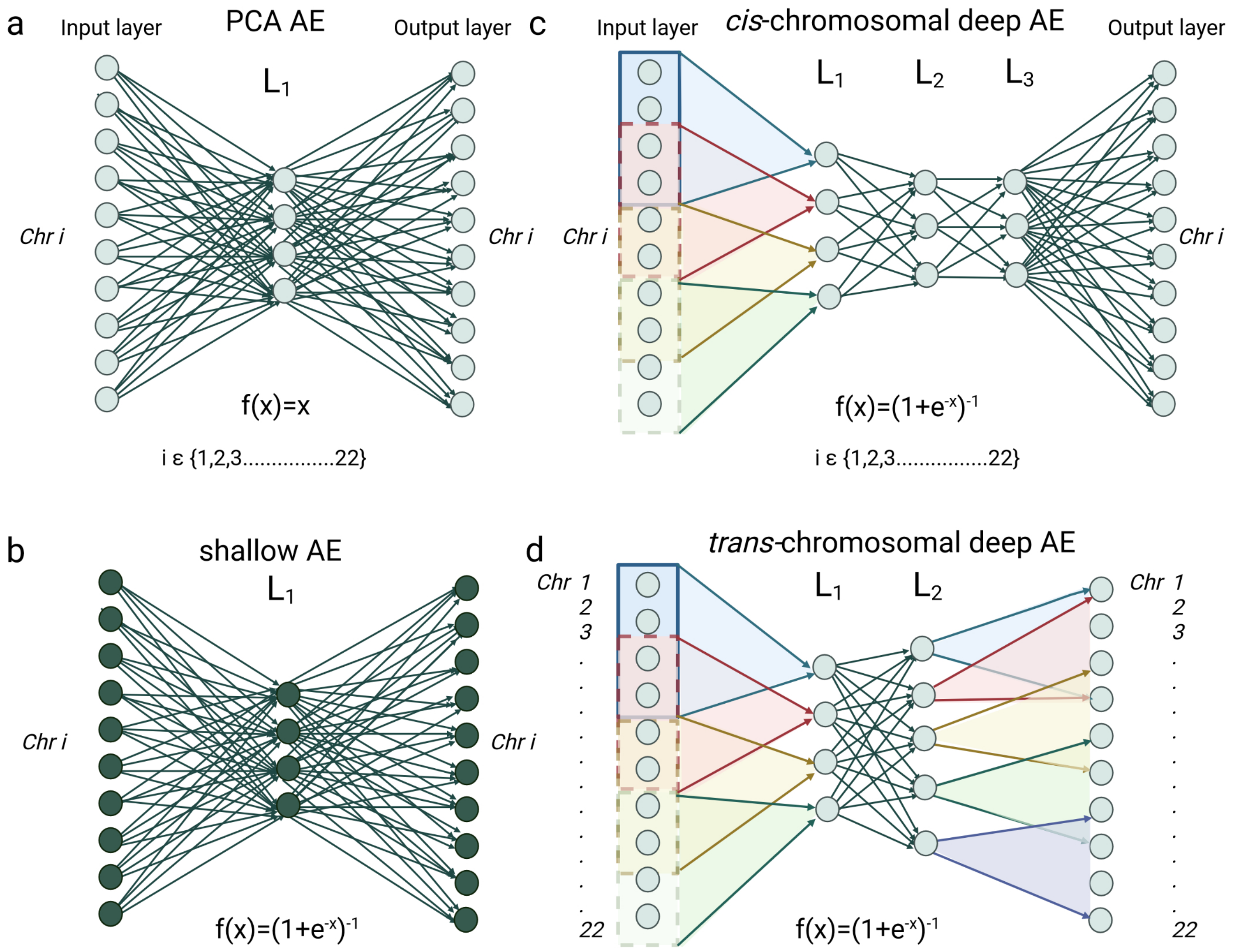


Figure 2. **Schematic representation of autoencoders. a and b,** Linear (PCA) and non-linear shallow autoencoders with one hidden layer and one densely connected layer trained for each autosomal chromosome separately. **c,** Deep autoencoder with three hidden layers; first hidden layer is locally connected with a connecting strategy of having a kernel size of 4 and shifting parameter 2 with the rest of the layers being densely connected. This autoencoder was trained on each chromosome separately, here termed cis-chromosomal deep AE. **d,** Trans-chromosomal deep autoencoder with a first locally connected layer and a dense layer followed by a customized locally connected layer that was trained on all autosomal chromosomes combined. Figure was created using BioRender.com

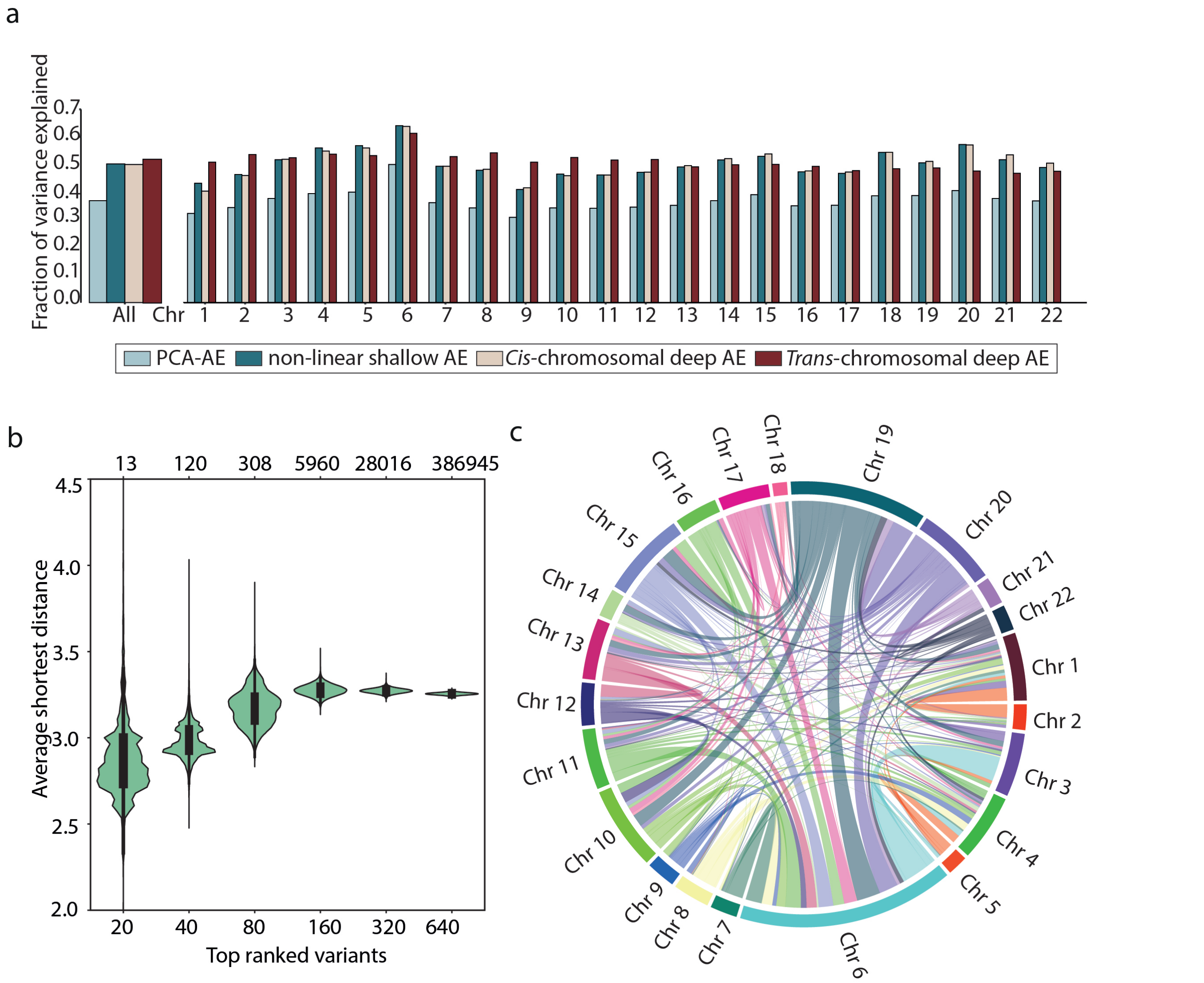


Figure 3. Prediction performance of various autoencoders in uk-biobank data and unfolding the biological insights. a, The fraction of explained SNPs in UK biobank test data set by PCA AE (light blue), non-linear shallow AE (blue) and cis-chromosomal deep AE (beige) trained separately for each autosomal chromosome and by trans-chromosomal deep AE (darkred). **b,** Shows the inter and intra chromosomal SNPs relations using light-up from first hidden layer of trans-chromosomal deep AE. The total number of connections among the genes are shown at the top x-label. **c,** trans-chromosomal deep AE interpreted as pairwise chromosomal relations using the light-up.

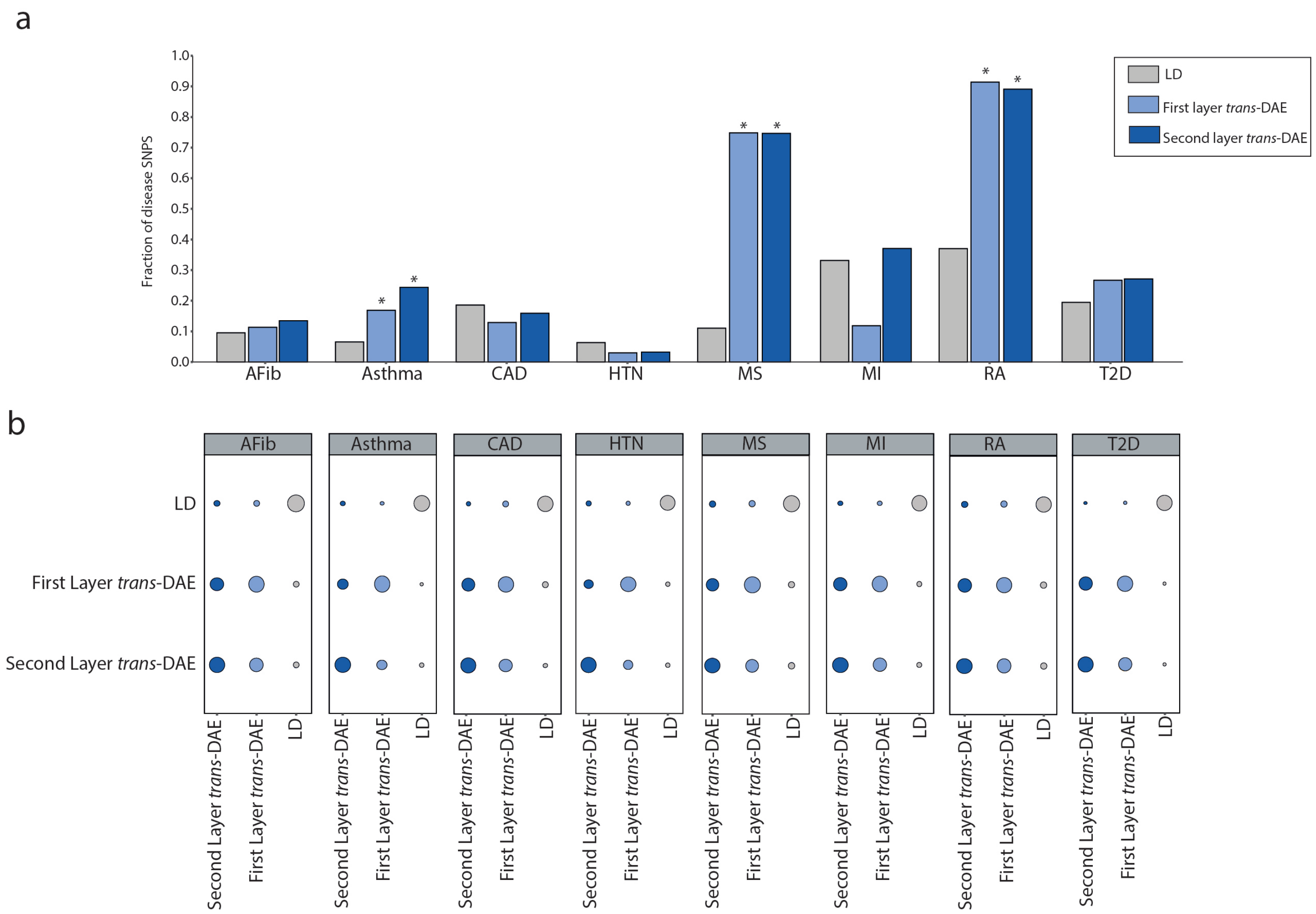


Figure 4. Disease enrichment of the variant modules. a, Pairwise Jaccard index of the methods among the top 200 prioritised variants. **b,** The autosomal autoencoder based SNP module shows accuracy in predicting the GWAS significant variant ($p - value < 5 \times 10^{-8}$) from independent study.

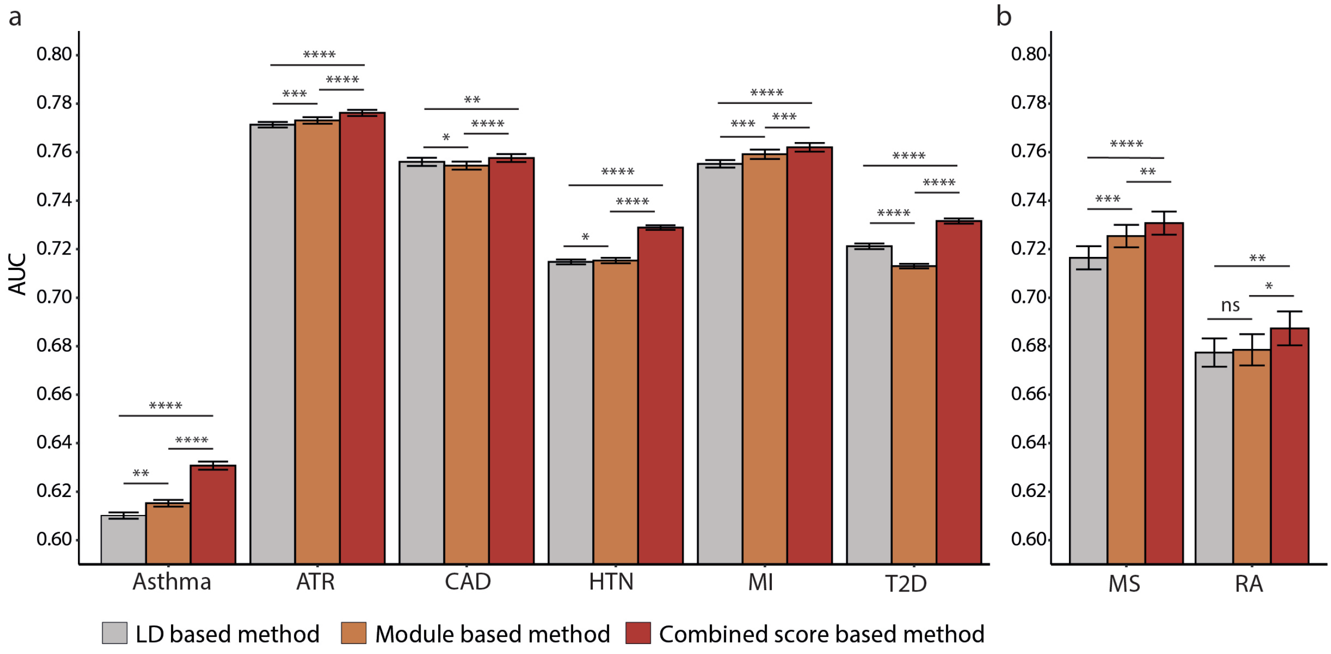


Figure 5. Increase in disease prediction performance after adding the autoencoder derived features using lasso. a, exhibits AUC of uk-biobank disease prediction under 10-fold cross-validation data-sets using LD (left bar), trans-DAE derived module (middle bar) and combined scores generated with these two methods other two methods (right bar) that use the input features as first and second hidden layer compressed representations. Similarly **b,** shows AUCs of Rheumatoid arthritis (RA) and Multiple sclerosis (MS) on independent data-set via integrating uk-biobank samples in training only. Combine score based method wins 10 times out of 10-folds data sets for Asthma, Atrial-fibrosis , Hypertension, Myocardial-infraction, and Type-2 diabetes, also it is 9 times for MS and 8 times for CAD and RA over LD and modules based methods. The age and gender were included as additional two features in the lasso-model.

Supplementary Files

- SupplementalInformation.docx