medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Non-linear genetic regulation of the blood plasma proteome

Authors

Arnor I. Sigurdsson[1,2,3,*], Justus F. Gräf[1,2,3,*], Zhiyu Yang[4], Kirstine Ravn[1,2], Jonas Meisner[1,2,5], Roman Thielemann[1,2,3], Henry Webel[1,2], Roelof A. J. Smit[2,3,6], Lili Niu[1], Matthias Mann[1,7], FinnGen, Bjarni Vilhjalmsson[3,8,9], Benjamin M. Neale[3,10,11], Andrea Ganna[4,10,11,12], Torben Hansen[2], Ruth J. F. Loos[2,3,6], Simon Rasmussen[1,2,3]

Affiliations

1. Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen N, Denmark
2. Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
3. Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
4. Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland
5. Mental Health Centre Copenhagen, Copenhagen University Hospital, Denmark
6. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
7. Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany
8. National Centre for Register-based Research, Aarhus BSS, Aarhus University, Aarhus, Denmark
9. Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark.
10. Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA
11. Analytic and Translational Genetics Unit, ATGU, Massachusetts General Hospital, Boston, MA, USA
12. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

Author List Footnotes

*These authors contributed equally

Correspondence: Simon Rasmussen (srasmuss@sund.ku.dk)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

Although thousands of genetic variants are linked to human traits and diseases, the underlying mechanisms influencing these traits remain largely unexplored. One important aspect is to understand how proteins are regulated by the genome by identifying protein quantitative trait loci (pQTLs). Beyond this, there is a need to understand the role of complex genetics effects such as dominance and epistasis that regulate plasma proteins and protein biomarkers. Therefore, we developed EIR-auto-GP, a deep learning-based approach, to identify such effects. Our results complement the additive genetic regulation identified in previous pQTLs screens by adding a nuanced view of the complex genetic regulation of plasma proteins. Applying this method to the UK Biobank proteomics cohort of 48,594 individuals, we identified 138 proteins that were regulated by non-linear effects, including non-linear covariates (123) as well as genetic dominance and epistasis (15). We uncovered a novel epistatic interaction between the *ABO* and *FUT3* loci, and demonstrated dominance effects of the *ABO* locus on plasma levels of pathogen recognition receptors CD209 and CLEC4M. Furthermore, we replicated these findings and the methodology across Olink and mass spectrometry-based cohorts and concluded that large sample sizes are needed to discover more complex genetic effects. Our approach presents a systematic, large-scale attempt to identify complex effects of plasma protein levels and can be applied to study other tissues or molecular QTLs.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Genome-wide association studies (GWAS) have identified thousands of associations between genetic variants and phenotypic traits[1]. Despite these discoveries, it remains a significant challenge to understand how these genetic variants contribute to the phenotypic traits. This is mainly because the functional impact of many variants is still unknown. This area of focus in modern genetic research is known as Variant-2-Function (V2F). Addressing the V2F challenge is crucial for identifying how genetic variants influence biological pathways, which can improve our understanding of disease mechanisms and allow for more precise drug development[2].

In response to this challenge, multi-omics approaches have been applied to population-based cohorts, typically including transcriptomics, proteomics, metabolomics, or microbiomics[3–8]. Blood plasma, in particular, serves as an easily accessible and minimally invasive sample for diagnostics and biomarker discovery. Large-scale blood metabolomics are now available from resources like the UK Biobank, further enriching our understanding of the genomic basis of complex traits[7]. Moreover, proteomics-based analyses using either aptamer-based (SomaScan), antibody-based (Olink) or mass spectrometry (MS)-based assays of individual-level biobank samples have revealed thousands of protein quantitative trait loci (pQTLs)[5,6,9–14]. These can bridge the gap between genetic variants and phenotypes and allow for deeper functional understanding of diseases, improving drug target discovery and contributing to our understanding of genetic effects on disease[15–17]. Currently, GWAS has been the most widely applied methodology for discovery of pQTLs. However, GWAS often assumes an additive model and might not fully recapitulate complex, non-additive effects among the variants and relevant covariates. Despite its robustness, it has been shown that deviations from the additive model exist in a number of human loci, for example in the form of dominance effects[18]. Additionally, interactions of two or more variants can result in a larger effect on a phenotype than the effect of each single variant, a concept known as epistasis, which can also contribute to non-linear genetic architecture of complex traits[19–21].

DL models can capture non-linear effects, which has motivated recent work in applying DL and other non-linear models for both genetic prediction and variant-phenotype association, providing new insights into the genetic architecture of complex traits[22–26]. For example, in previous studies,

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

we developed and applied DL frameworks for disease prediction in the UK Biobank[27], and found potential dominance and epistatic effects, specifically for immunological diseases such as type 1 diabetes (T1D), involving the insulin gene and *HLA-DQB1*[28–30]. These complex effects also transfer to molecular quantitative trait loci, as indicated by our previous analysis of 34 common biomarkers in the UK Biobank[31]. Additionally, targeted discovery of epistatic effects between genetic variants uncovered the presence of interactions between the *ABO* blood group and the *FUT2* secretor status that influence blood plasma abundance of gastrointestinal (GI) proteins[6]. Unbiased approaches have been used to identify numerous epistatic and dominance effects that influence the plasma levels of lipids and their effects on cardiovascular diseases[32]. However, by now, there have not been attempts to systematically characterize non-linear effects that influence blood plasma protein levels.

Here, we present a systematic, DL-based workflow that allows us to identify non-linear effects like non-linear covariate effects, dominance, and epistasis that influence plasma protein levels. To demonstrate the use of our approach, we examined data from 2,922 blood plasma protein levels measured in 48,594 individuals from the UK Biobank Pharma Proteomics Project (UKB-PPP). Our study presents a nuanced view of non-additive effects that influence plasma proteomics in the UK Biobank. We could illustrate the quantitative and qualitative non-linear regulation of the blood plasma proteome and reveal novel non-linear effects that are highly likely to influence plasma protein abundance. Using our approach, we identified 138 proteins, among which 123 were potentially regulated by non-linear covariate effects and 15 by dominance or epistasis effects. This highlights DL as a useful tool to uncover complex effects that influence molecular quantitative traits, which can contribute to our understanding of the genetic architecture of complex traits.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Results

Modeling blood plasma protein levels using deep learning

To investigate the scale of non-linearity of genetic control of protein abundances in the blood plasma, we modeled the abundance of 2,922 proteins in the blood plasma proteome in the UKB cohort. Based on our deep learning framework, EIR, and the genome-local net deep learning architecture (GLN)[27], we developed an automated framework, EIR-auto-GP, to predict the abundance of a protein from genotypes and covariates (age, sex, UKB center, UKB genetic array, whether an individual was consortium selected and genetic principal components 1-20 (**Figure 1a**). We used grouped, self-reported ethnicities (**Methods**) that resembled the distribution of the genetic population structure in the UKB to subset individuals of UK-white self-reported ethnicity as the largest group of individuals with similar ancestral background for model training and testing (**Supplementary Figure 1a**). Subsequently, after quality control (QC) of the proteomics data, the remaining 48,594 individuals were split into a train (n=34,947), validation (n=2,000) and test split (n=1,771) (**Figure 1a**). As input to EIR-auto-GP, we used 424,097 measured QC-passed genotypes, which reduced computational complexity compared to the more extensive imputed data. Furthermore, we limited the amount of input variants by using the training dataset to conduct GWAS for each protein and selecting associated variants (**Supplementary Note 1**). When analyzing the results of the per-protein GWAS we found them to be consistent overall with previous work[6], and overlapping variants showed high correlation (**Figure 1b**). To determine input variants for the DL modeling, we used a less stringent p-value threshold than usually applied to GWAS ($P<0.001$), resulting in most DL models being trained on ≤1,000 variants（**Figure 1c**). Taken together, the variants identified through our GWAS and subsequently used as inputs for the DL models were likely pQTL candidates.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

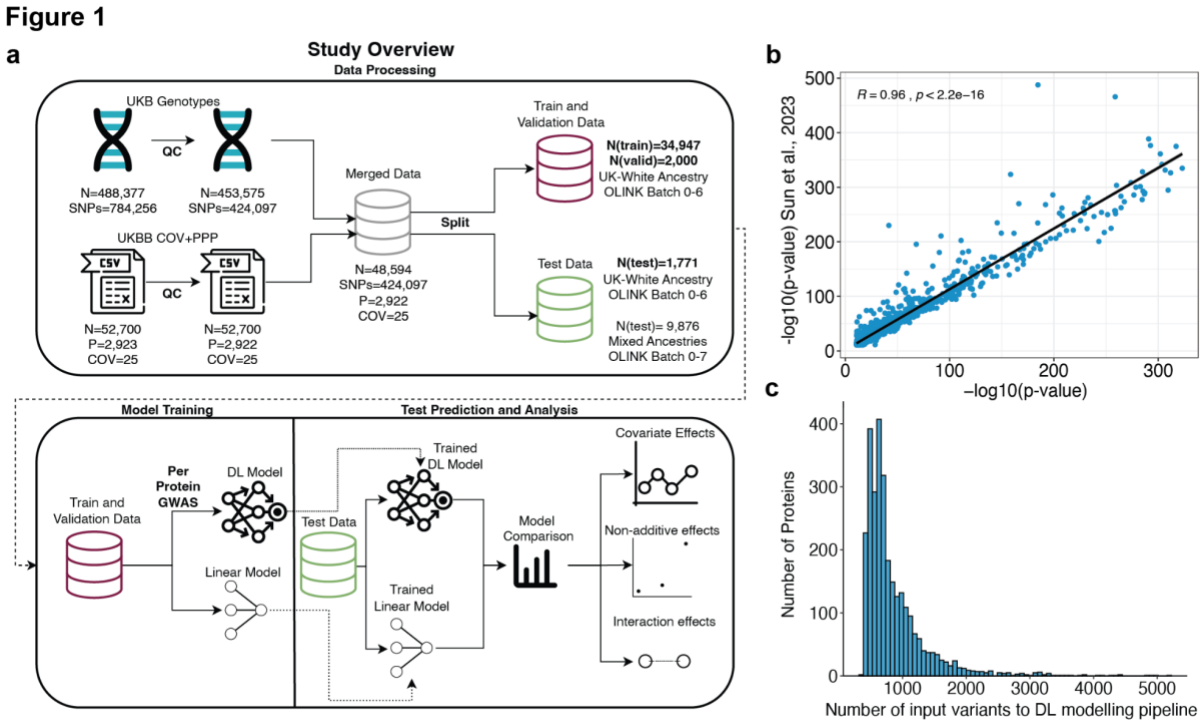

Figure 1. Overview of the study and protein pre-GWAS results. a) Overview of study design and workflow. UKB genotypes underwent quality control (QC), resulting in 424,097 QC-passed SNVs. The data were split into training and validation sets of self-reported UK-white ethnicity, OLINK batch 0-6 (n=34,947 for training; n=2000 for validation), and test sets stratified by ethnicity and batch: UK white self-reported ethnicity (n=1,771) and mixed ethnicities (n=9,876), all from OLINK batches 0-6 and 0-7 respectively. The training and validation data were used to develop DL and linear models, with a per-target GWAS on the training set used to pre-filter input variants for training the DL model. Finally, predictions and analyses were performed on the test data, and proteins that had discordant performance between the DL and linear models were investigated for non-linear covariate, non-additive (e.g., dominance), and interaction (e.g., epistasis) effects. **b)** Correlation of GWAS P-values between the current study and Sun et al.[6]. Variants with p-values exactly 0, likely due to being below the numerical precision threshold (underflow), were omitted from the plot. The scatter plot represents the $-\log_{10}$(p-values) correlation of 1,780 overlapping genetic variants with significant associations (p<1.7e-11) between our analysis and Sun et al.[6]. The strong correlation (R=0.96, P<2e-308) between p-values demonstrates consistency in identifying significant associations. **c)** Histogram of the number of input SNVs used for DL model training following per target GWAS pre-filtering, where only SNVs with p-values < 0.001 (computed on the training set) were considered. For the majority of proteins, fewer than 1,000 SNVs passed the threshold.

Non-linear effects that influence blood plasma protein abundance

The performance (R^2) of the DL and linear models reached up to 0.95 and 0.86 with a median performance of 0.04 and 0.03, respectively (**Figure 2a, Supplementary Figure 2a-b**). We found

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

an association between proteins with low modeling performance (R2<0.1) and the correlation of their measurements with SomaScan measurements in an Icelandic cohort[14] (**Supplementary Note 2**). To investigate how many blood plasma proteins could be influenced by non-linear covariates and genetic effects, we compared the performance of the DL models (EIR-auto-GP) to a penalized linear model (bigstatsr)[33]. We calculated the difference in model performance on the UK-white test set (n=1,771) for each protein (**Supplementary Table 1**). For 1,503 of 2,922 proteins (51.4%), the DL model performed better, resulting in a significant difference when modeling plasma protein abundance from genotypes and covariates (paired T-test, two-sided, t=11.281, P=6.4e-29). To identify specific proteins for which the DL model was significantly better, we bootstrapped the predictions of each protein (**Methods**) and identified 171 proteins (5.8%) with a significant performance increase (non-overlapping 95% confidence intervals) (**Figure 2a**, **Supplementary Figure 2b)**. These proteins showed a median increase in R^2 of 0.038 (mean 0.05). To examine whether these results transferred to other metrics, we additionally used Root Mean Squared Error (RMSE) to assess model performance. Among the 171 proteins showing better performance with the DL model as measured by R^2, the RMSE analysis also found the DL model outperforming on all of these. Specifically, 28 of these proteins also showed significant improvement (non-overlapping confidence intervals). In summary, we replicate that linear models are robust in modeling plasma abundance of measured proteins[6,18] and that our DL approach can identify candidate proteins with potential non-linear effects that influence their plasma levels.

EIR-auto-GP can identify non-linear effects using NPX and INT protein abundance data

To preserve most of the protein level variance, we modeled the Olink protein expression values (NPX); however, these are non-normal distributed and on a $\log_2$-like scale[6]. This could favor the DL model over the linear model without biological non-linear effects because a $\log_2$ transformation can make a fundamentally linear relationship appear non-linear. Therefore, we re-ran our models for the 171 proteins using Inverse-rank Normal Transformed (INT) protein levels and found reduced R^2 for both DL and linear models, suggesting that information was lost when rank-transforming the protein levels (**Figure 2b, Supplementary Table 2**). Despite this reduction in performance, we found that for 138 (81%) of the 171 proteins, a significant gap in performance between the DL and linear models remained (**Figure 2c**). Conversely, for 33 of the 171 proteins the difference between the DL and linear models was not significant anymore (**Figure 2c**). Where

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

indicated, these 33 proteins were excluded from downstream analyses. The performance gap of most of the remaining proteins correlated well between NPX and INT normalized values, indicating that the DL model also identified non-linear effects on INT normalized protein values (**Figure 2c**).

Figure 2

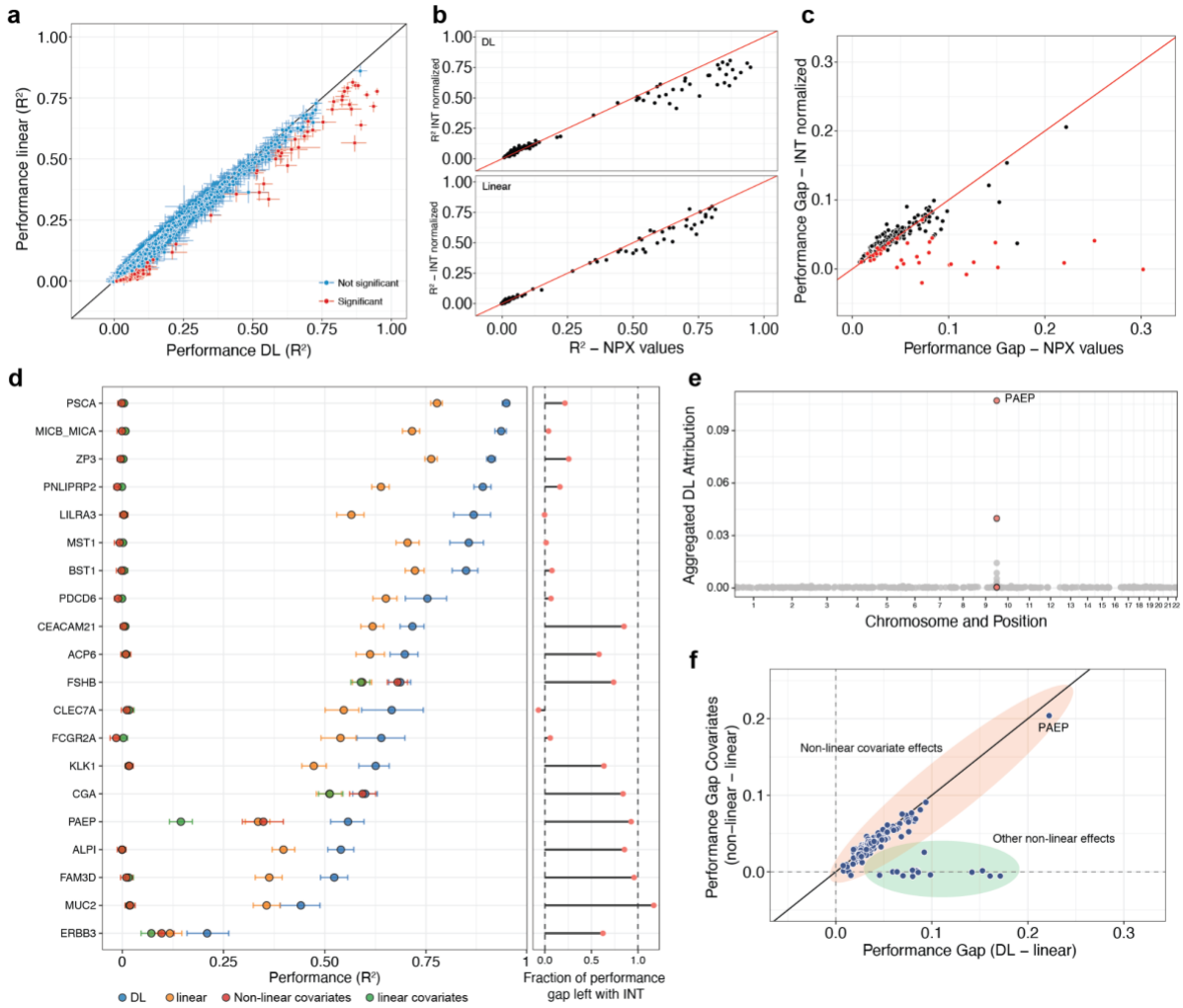

Figure 2. Deep learning reveals non-linear genetic and covariate effects. a) DL (EIR) and linear (bigstatsr) model performance (R^2) for all 2,922 proteins (**Supplementary Table 1**). The error bars indicate the 95% confidence intervals (CI) from 1,000 bootstraps, and proteins with non-overlapping confidence intervalsbetween DL and linear models are called significant and labeled in red. **b)** DL (top) and linear (bottom) model performance of 171 significant proteins modeled on raw protein expression values (NPX) or INT normalized protein values (**Supplementary Table 2**). **c)** Performance gap (R^2-R^2) for the 171 significant proteins between DL and linear models (DL-linear) on NPX or INT normalized protein values. Proteins labeled in red indicate that no significant performance gap (overlapping CIs)

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

was found when modeling on INT normalized protein values **d)** DL and linear model performance for the top 20 significant proteins with the largest absolute performance gap in R^2 between the DL and linear models are shown. Additionally, performance of linear and non-linear (XGBoost) models trained only on covariates are shown (**Supplementary Table 3**). The covariates include demographic information (age and sex), the genetic array, genetic principal components (GPC1-GPC20), whether individuals were consortium selected, and the research center location for participant measurement. On the right, the fraction of the performance gap that remains when modeling on INT values instead of NPX protein levels is shown. **e)** Aggregated DL attribution of 487 SNVs across the genome that was used as input to model PAEP protein levels. Variants located within the *PAEP* gene are labeled in red. **f**) Performance gap (R2-R2) between DL and linear models on genotype and covariates against the performance gap between non-linear (XGBoost) and linear models on covariates only. Orange and green areas indicate if protein levels underlie non-linear covariate effects or other non-linear effects in the input data.

Genetics was the main driver of model performance for a subset of proteins

We then investigated the proteins with the largest absolute increase in performance by either method (top 20) (**Figure 2d**). For these proteins, the DL model reached an R^2 between 0.21 (ERBB3) to 0.95 (PSCA) on the test set (**Figure 2d**). Among these 20 proteins, 11 showed a maintained significant performance gap with the INT normalization. To investigate the contribution of covariates on the performance, we trained and evaluated linear and non-linear (XGBoost[34]) models using only the covariates (**Supplementary Table 3**). We found that for 8 of these 11 proteins, the genotype data was the main driver of model performance (**Figure 2d**). For example, the DL performance of PSCA and FAM3D was mainly driven by known cis- and trans-pQTL (**Supplementary Figure 2e**), which was expected due to their high association in previous studies[6]. For some proteins, for instance, MICB/A and LILRA3, where genetics primarily contributed to the DL model performance, there was no difference in performance between the DL and linear model when modeling on INT normalized values (**Figure 2d, right panel**). However, other proteins showed sustained performance gaps when modeling on INT normalized values, providing additional confidence that the increased DL performance might be caused by non-linear genetic effects (e.g., CEACAM21, ALPI, FAM3D, or MUC2).

Non-linear covariate effects influence protein levels

For 3 of the 11 significant proteins, we found that non-linearities in the covariates could account for the entire gain in performance (FSHB, CGA & PAEP) (**Figure 2e**). For instance, the gain in R^2 for follicle stimulating hormone subunit beta (FSHB) could be entirely explained by the

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

covariates sex and age, related to the age of menopause (**Figure 2e, Supplementary Figure 2f**)[6]. Furthermore, for progestogen-associated endometrial protein (PAEP), we found that a combination of covariates and genetics could account for the increased model performance (**Figure 2d**). Besides age and sex related non-linear effects **(Supplementary Figure 2f**), PAEP levels were influenced by cis-pQTLs which contributed to model performance (**Figure 2e**). When expanding the analysis to all 138 proteins with significant differences, we found that for 123 the performance gap between DL and linear models could be explained by non-linear covariate effects (**Figure 2f**). These results indicate that we could robustly identify proteins with plasma levels that underlie non-linear covariate effects, which was in line with the non-linear modeling of covariates in other studies[25]. However, it also revealed that for a substantial fraction of the 138 proteins (10.9%, 15 proteins) effects in the covariates could not account for the increased performance of the DL model to the linear model (**Figure 2f**). This suggests that the improved predictive accuracy obtained with the DL model was not solely due to non-linear covariate effects.

Dominance in the *ABO* locus influences plasma levels of CD209 and CLEC4M

Next, we investigated the contribution of dominant genetic effects in modeling plasma protein levels. Because dominance effects can be modeled by a linear model when using non-additive encoded genotypes, such as one-hot encoding, we compared non-linear (XGBoost) and linear models on genotype data using additive and non-additive encoding (**Supplementary Table 4**). To focus on key genetic variants, we utilized the DL model feature importance computed on the validation set to select the top 128 SNVs (**Methods**). This reduced set allowed us to use XGBoost, which is known for its robust performance on structured data[35,36] but might not scale as efficiently to the high-dimensional datasets. Comparing the non-linear XGBoost with linear models served as an additional verification of non-linear effects beyond the original models trained on the full set of features. We found that for a group of proteins (CD209, CLEC4M, ABO, PSCA) using a linear model with non-additive encoding of genotypes improved the performance of the linear model to be almost equal to the non-linear model (**Figure 3a**). This indicated that the non-linear effects underlying their plasma levels were likely due to genetic dominance. Furthermore, we identified multiple proteins where the non-additive model could partly improve the performance of the linear model (e.g., KLK1, FAM3D, MUC2, ALPI, CEACAM21 and more) (**Figure 3a**). This indicated that for these proteins, both dominance effects but also other non-linear genetic effects influenced

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

their plasma levels. We found that two variants in the *ABO* locus (rs505922, rs8176719, chr 9) showed dominance effects on protein levels for CD209, CLEC4M, FAM3D, ALPI, ABO and MUC2 (**Supplementary Figure 3a**). CD209 is part of the C-type lectin family and is involved in cell adhesion and pathogen recognition[37]. It is highly similar to CLEC4M in function and sequence. The two genes are located nearby on chr 19[37,38] and are referred to as *DC-SIGN* and *DC-SIGNR*, respectively. Notably, the two variants, rs505922 and rs8176719, were used to impute the blood-types of the *ABO* blood group system in the UKB[39–42], which is known to have co-dominance effects of its A and B alleles. Consistent with this, and the non-additive analyses above (**Figure 3a)**, we found dominant blood group effects on the plasma levels of CD209 and CLEC4M (**Figure 3b-c, Supplementary Figure 3b**). We assessed the influence of the dominance effect on model performance by training linear models for CD209 and CLEC4M using genotype and covariate data and one-hot encoded either rs8176719, rs505922 or both. We found that by one-hot encoding these variants the model performance improved by R^2 0.03 (7%), 0.0396 (9.21%) and 0.0399 (9.28%) (**Supplementary Figure 3c**). Taken together, using our approach, we could identify varying levels of dominance within loci that regulate plasma protein levels.

Non-linear interactions between genetic variants affect protein levels

Following the previous results, we further investigated proteins where the increased R^2 could not be explained by non-linear covariate effects to identify potential epistatic SNV-SNV interactions. For each of these 15 proteins, we analyzed the 128 SNVs with the highest feature importance in the DL models on the validation set. To achieve this, we applied pairwise Ordinary Least-Squares (OLS) models to the training set (n=34,947) to identify epistatic interactions. Restricting our analysis to SNV pairs on different chromosomes, we identified at least one significant (p-value <4.46e-08) interaction for 8 of the 15 proteins and a total of 784 interactions between 67 unique SNVs on 5 chromosomes (**Figure 3d-e, Supplementary Table 5**). The majority of these interactions (753, 96%) were between variants on chr 9 and chr 19 and most of the interacting variants were located near the *ABO* and *FUT2* loci on chr 9 and 19, respectively (**Figure 3e, Supplementary Figure 3d-e**). For instance, we identified most interactions for ALPI, MUC2, FAM3D and CDH17 with 230, 216, 172 and 72 interactions, of which 31, 31, 40 and 9 were between variants within the *ABO* and the *FUT2* locus (+/- 10kb) (**Figure 3f**). We found an epistatic interaction between the *ABO* variant rs507666 and rs2307019, a variant in the *IZUMO1* gene, 40kb

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

downstream of the *FUT2* locus, that influenced plasma levels of these proteins (**Figure 3g**). The variant rs2307019 was in moderate linkage disequilibrium (R^2=0.35) with rs601338 (Trp154Ter), a variant that determines *FUT2* secretor status used in Sun et al., and Snaebjarnarson et al.[6,32] (**Supplementary Figure 3f**). We thus expected that the interaction between rs507666 and rs2307019 resembled an interaction between the *ABO* and *FUT2* locus. In line with our studies, Sun et al. found that epistatic interactions between the *ABO* and *FUT2* locus have a strong influence on the blood plasma levels of ALPI, MUC2, and FAM3D[6]. To assess if the interaction between rs507666 and rs2307019 influenced modeling performance when predicting plasma proteins levels, we trained linear models using one-hot encoded genotypes and covariates and added an interaction term for rs507666-rs2307019 to predict levels of ALPI, FAM3D, MUC2 and CDH17. We found that, when adding the single interaction term, the linear models improved by R2 0.021 (5.1%), 0.024 (5.3%), 0.017 (4.5%) and 0.007 (2.3%) respectively, indicating that this epistatic interaction accounts for a substantial fraction of model performance (**Figure 3h**). In summary, these results demonstrate that EIR-auto-GP could identify proteins with epistatic interactions between genetic variants, which we could subsequently validate using targeted OLS models.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

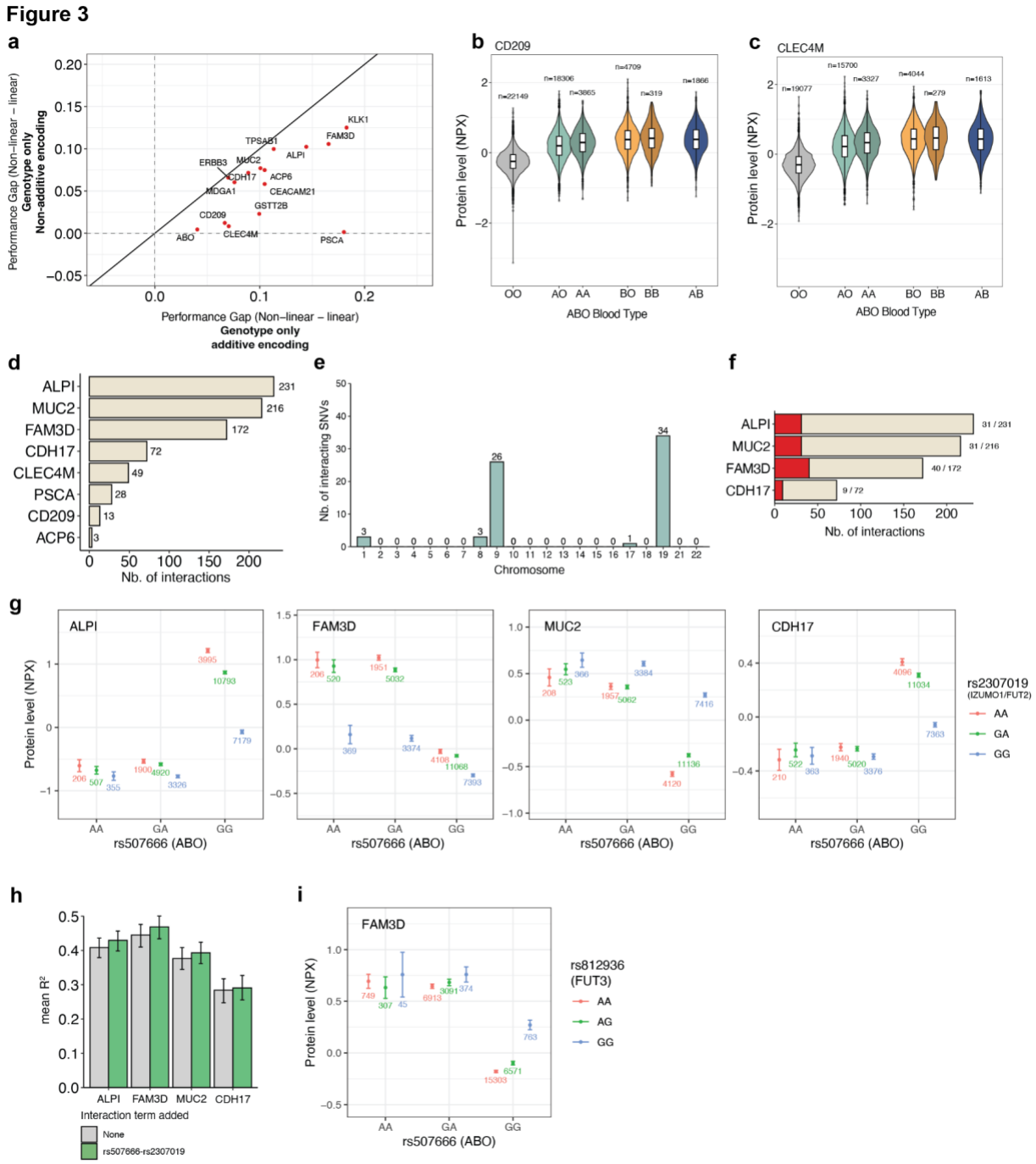

Figure 3. Identification of dominance effects and epistatic interactions between genetic variants on separate chromosomes that influence protein levels. a) Performance gap (R2-R2) between non-linear (XGBoost) and linear models trained and tested on additive or non-additive encoded genotype data for 15 candidate proteins with potential non-linear genetic effects. **b**) NPX protein levels of CD209 (n=51,214) in individuals in the UK Biobank, stratified by their imputed *ABO* blood group (field p23165)[39–42]. **c)** NPX protein levels CLEC4M (n=44,040) in individuals in the UK Biobank, stratified by their imputed *ABO* blood group. **d)** Number of identified SNV-SNV interactions

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(p<3.6e-08) per protein. 171 proteins with significant gap in performance between DL and linear model were tested in an Ordinary least-squares (OLS) model, of which 14 had at least one significant interaction and of which 6 were excluded due to potential false positive non-linear effect (**Figure 2b, c**). Interactions were limited to interactions between SNVs on two different chromosomes. **e)** Number of unique interacting SNVs per chromosome. **f)** Number of interactions between SNVs on *ABO* and *FUT2* loci (+/- 10kb) as a fraction of the total number of interactions for each protein. **g)** Protein expression levels (NPX) of ALPI, MUC2, FAM3D and CDH17 for individuals of the training dataset (n=34,947) with all combinations of genotypes of the interacting variants rs507666 (*ABO*) and rs2307019 (*IZUMO1*/*FUT2*). Error bars indicate the 95% confidence interval and the number of individuals with the respective interaction are shown below each data point. **h)** Linear model performance to predict ALPI, FAM3D, MUC2 and CDH17 plasma levels trained on one-hot encoded genotypes and covariates. Interaction between rs507666 and rs2307019 was added as a single term to assess performance improvement. Error bars indicate 95% confidence interval of 1000 bootstraps. **i)** Protein levels of FAM3D for all genotype combinations of the interacting variants rs507666 (*ABO*) and rs812936 (*FUT3*) for individuals from the training dataset (n=34,947). Error bars indicate the 95% confidence interval and the number of individuals with the respective interaction are shown below each data point.

FAM3D protein levels depend on interactions between *ABO* and *FUT3* loci

In addition to the previously reported interactions between *ABO* and *FUT2* above, we identified an interaction between *ABO* (rs507666) and the *FUT3* locus (rs812936) that influenced the blood plasma levels of FAM3D, which is expressed in the gastrointestinal tract (**Figure 3i**). The *FUT3* locus is also known as the Lewis gene, and encodes an alpha(1,3/4)-fucosyltransferase as part of the Lewis antigen system[43]. rs812936-A, was associated with increased levels of FUT3 plasma levels[44] and led to decreased FAM3D plasma levels when interacting with rs507666-G in the *ABO* locus (**Figure 3i**). This suggested that the protein level of FAM3D was not only regulated by epistatic interactions between *ABO* and *FUT2*, but also dependent on interactions between *ABO* and *FUT3* variants. As *FUT2* and *FUT3* are located 45 Mbp apart on chr 19 this was likely not caused by LD between the two genes (**Supplementary Figure 3e**). Other proteins influenced by *ABO-FUT2* interactions were enriched for gastrointestinal (GI) expression and may be perturbed in GI disease[6]. We identified *ABO-FUT3* interactions for FAM3D, a GI expressed protein, and thus speculated that *FUT3* could also be involved in regulating plasma abundance of GI expressed proteins. However, when we added this interaction term to a linear model predicting FAM3D levels, the performance did not improve (**Supplementary Figure 3g**). Notably, this interaction was relatively rare in the UKB-PPP with 763 individuals in the training set, 47 in the valid set and 37 in the test set that carried this interaction. This indicates that rare interactions might be relevant in regulating plasma protein levels, but that they were difficult for our DL model to detect because

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

improvements for only a few individuals in the validation set are not likely to be prioritized by the model.

Variable non-linear improvements across self-reported ethnicities

Given the importance of understanding model performance across diverse ethnic groups[45–49], we investigated how the performance of the linear and DL models trained and evaluated on self-reported UK white ethnic background transferred to other self-reported ethnic backgrounds ("South Asian", "East Asian", "African" and "Caribbean") in the UKB (**Methods**). We generally observed a decline in performance for both DL and linear models across the non-UK white ethnic groups, accompanied by larger confidence intervals for the 138 proteins with non-linear effects (**Supplementary Figure 4a-b, Supplementary Table 6)**. Overall, the linear model transferred better to the 'East Asian' and 'South Asian' test set than the DL models, while the DL models transferred better to 'African' and 'Caribbean' test sets than the linear models (**Supplementary Figure 4a-b**). One contributing factor could be the breakdown and formation of LD patterns across populations, as models trained on the UK white group may select tagging variants that do not replicate in the other ethnic groups[48,50]. Additionally, the larger confidence intervals could also be partly due to the smaller number of samples available in ethnicity test sets. We observed correlation of performance gaps (mean bootstrapped R^2 DL-linear) between all ethnicity groups, except between 'East Asian' and 'African' and 'Caribbean' (**Figure 4a**). For instance, the DL model outperformed the linear model for CD209 on the 'South Asian' and 'African' test set, while it showed similar or worse performance on 'East Asian' and 'Caribbean' test sets (**Figure 4b**). Despite this, CD209 levels showed similar trends when stratified by *ABO* blood type between the different ethnicity groups in the whole UKB-PPP (n=52,700) (**Figure 4c**). Notably, the distribution of *ABO* blood types was different between the ethnic groups in the UKB-PPP, which could influence the performance on the different test sets (**Supplementary Figure 4c**). Above, we found that the DL model could identify non-linear relationships between age and sex for FSHB (**Figure 2d, Supplementary Figure 2f**) and we found that this could be replicated in the 'South Asian', but not in the 'East Asian', 'Caribbean' or 'African' test sets (**Figure 4d**). Despite that, we observed non-linear relationships between age and sex for FSHB in the test sets of non-white self-reported ethnicities (**Supplementary Figure 4**). This might be due to the penalized linear model being less affected by the higher genetic diversity found in African and Caribbean populations[51,52].

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Finally, the non-linear effects for FAM3D, likely caused by epistatic interactions (**Figure 3d-e**) could not be fully replicated across the ethnicity test sets (**Figure 4e**). The DL model only outperformed the linear models slightly on 'South Asian', 'East Asian' and 'Caribbean' tests set with a much smaller extent. As above, this could be due to different LD patterns or variants not replicating across the ethnic test sets. In summary, these results suggest that cross ethnicity training is likely needed for the non-linear patterns of the DL model to transfer to individuals from diverse ethnicities that the model was not trained on.

Figure 4

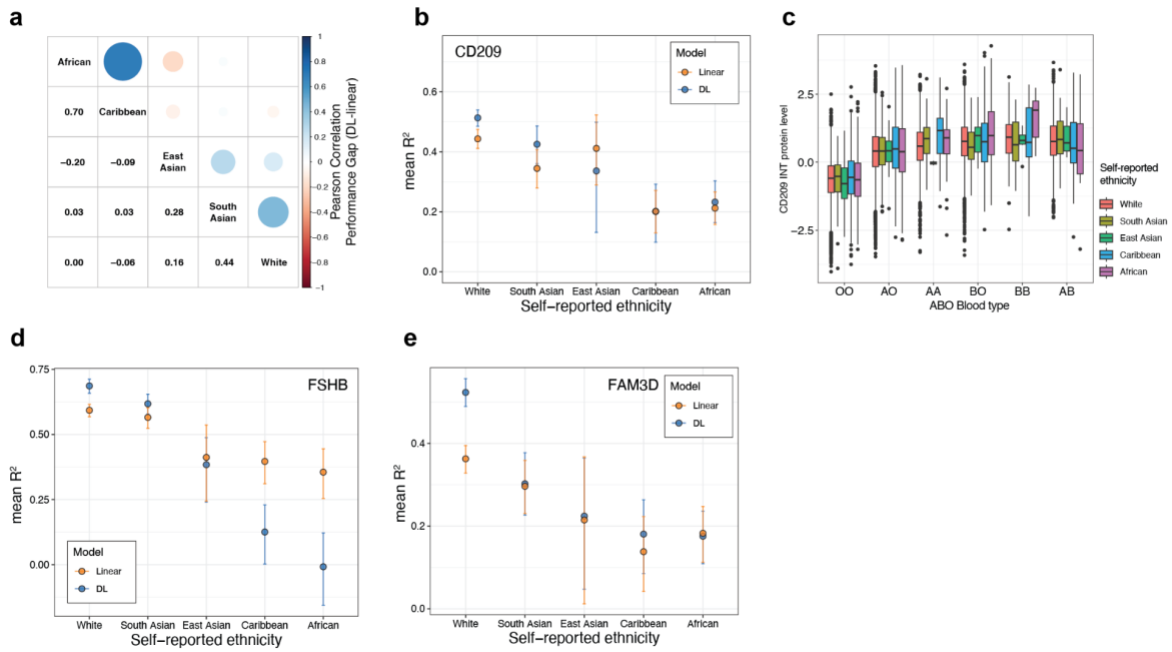

Figure 4. Examples of deep learning and linear model performance across self-reported ethnicities. a) Pearson correlation between the performance gap (mean bootstrapped R2 DL-linear) of models between different self-reported ethnicities for 138 proteins with potential non-linear effects in the UK Biobank. The models were trained on individuals of self-reported 'white' ethnicity and tested on individuals of 'White', 'South Asian', 'East Asian', 'Caribbean' or 'African' self-reported ethnicities. If the R2 was negative for both models, the performance gap was set to 0. **b)** Mean bootstrapped performance of DL and linear models for CD209. The models were trained on individuals of self-reported 'white' ethnicity and tested on individuals of the respective self-reported ethnicities. The error bars indicate the 95% confidence intervals. **c)** INT normalized CD209 levels stratified by *ABO* blood type among the different self-reported ethnicities. AO and AA blood type correspond to A blood group, and BO and BB blood type correspond to B blood group. **d)** Mean bootstrapped performance of DL and linear models for FSHB. The models were trained on individuals of self-reported 'white' ethnicity and tested on individuals of the respective self-reported ethnicities. The error bars indicate the 95% confidence intervals. **e)** Mean bootstrapped performance of DL and linear

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

models for FAM3D. The models were trained on individuals of self-reported 'white' ethnicity and tested on individuals of the respective self-reported ethnicities. The error bars indicate the 95% confidence intervals.

Replication of non-linear effects and validation of the EIR-auto-GP workflow in FinnGen

We replicated our findings of non-linear genetic effects in a cohort of 1,757 individuals from the FinnGen project[8]. Olink protein levels were available for 170 of the 171 proteins with significant performance gap in the UKB. We were able to replicate the dominance effect of the *ABO* blood group tagging variants rs505922 and rs8176719 on plasma levels of CD209 and CLEC4M (**Figure 5a-b**, **Supplementary Figure 3a**). Furthermore, we investigated protein levels of FAM3D in individuals with different genotype combinations of *ABO* variant rs507666 and *FUT3* variant rs812936 and found higher levels in individuals with the GG-GG combination (**Figure 5c**). This replicated the discovery of this rare interaction in the UKB (**Figure 3i**), despite the much lower sample size and interaction allele counts (AC=43) in FinnGen. For ALPI, MUC2 and FAM3D we replicated the epistatic effect of rs507666 and the *IZUMO1* variant rs2307019, resembling *ABO* and *FUT2* secretor status interaction, on protein levels similar to the UKB (**Supplementary Figure 5a**). Next, we sought to replicate the ability of our EIR-auto-GP workflow to identify non-linear effects using the FinnGen cohort[8]. Using 1,231 and 263 individuals for training and test, respectively, we could replicate the discovery of potential non-linear covariate effects for FSHB and PAEP and potential non-linear genetic effects for MUC2, FAM3D and CD209 (**Supplementary Figure 5b**). We noticed that 94 of the 170 proteins had a higher DL performance in FinnGen compared to the UKB (**Supplementary Figure 5c**). We speculate that this was due to the different age distribution of the FinnGen cohort compared to the UKB (Median age FinnGen: 53 years, UKB: 58 years) (**Supplementary Figure 5d**). These results demonstrated that our DL model can predict protein levels from genotype and covariate data across cohorts. In summary, we were able to both directly replicate the dominance and interaction effects we discovered in UKB and, despite the significantly lower sample size, replicate the EIR-auto-GP workflow by re-discovery of non-linear effects of several proteins in FinnGen.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Validating discovery of non-linear effects using mass spectrometry-based proteomics

Finally, we aimed to replicate our analyses by training models using data generated by a different proteomics technology. We therefore, as above, retrained models using a Danish cohort of obese children measured using mass spectrometry based (MS) proteomics (The HOLBAEK Study)[10]. The cohort consisted of 1,924 children and adolescents between the age of 5–20 years (**Methods**), from which our group previously described the genetic regulation of its plasma proteome[10]. Similar to the approach in the UKB and FinnGen, we trained models for 411 MS protein levels in 1,533 individuals and tested their performance in 190 individuals (**Supplementary Table 8**). Despite the significantly lower sample size, we observed a similar trend in performance gaps, where the DL model outperformed the linear model on the majority of proteins (246 of 411; 69%) (**Supplementary Figure 5e**). Additionally, with a similar stringent cutoff as in the UKB, we identified 6 (1.6%) proteins that showed a significant (non-overlapping confidence intervals) increase when using the DL model (**Figure 5d**). When modeling protein levels only from covariates using non-linear (XGBoost) and linear models, we found that the DL performance of the significant proteins was likely driven by non-linear covariate effects (**Figure 5e**). Consistently, we found that COL1A1 levels were influenced by non-linear effects between age and sex (**Supplementary Figure 5f**). Taken together, these results demonstrate that our approach can be applied to proteomics data acquired by different assays including both mass spectrometry-based and affinity-based approaches.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

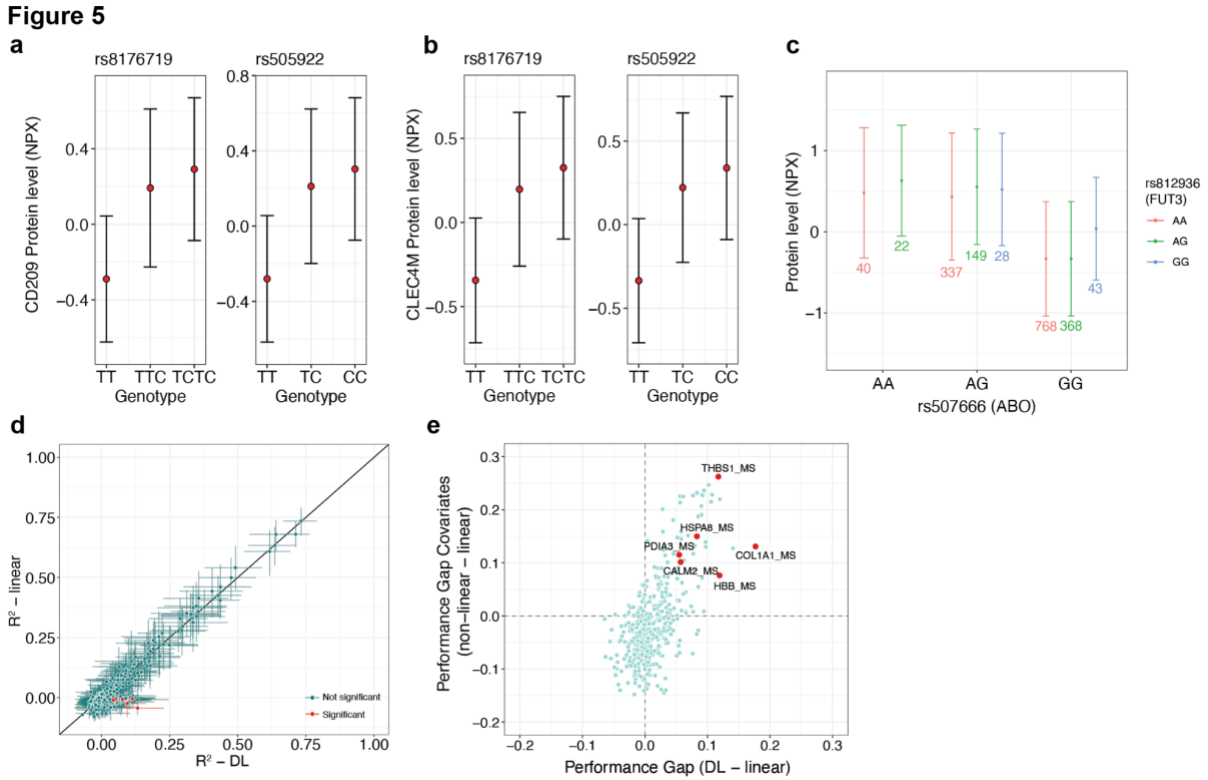

Figure 5. Replication of non-linear effects across cohorts and platforms. a) Protein levels (NPX) of CD209 in individuals with different genotypes of rs505922 and rs8176719 in the FinnGen project. Error bars indicate standard deviation. **b)** Protein levels (NPX) of CLEC4M in individuals with different genotypes of rs505922 and rs8176719 in the FinnGen project. Error bars indicate standard deviation. **c)** Protein levels (NPX) of FAM3D in individuals with different genotype combinations of rs507666 and rs812936. The numbers indicate the number of individuals with respective combinations. Error bars indicate standard deviation. Data points with < 5 individuals were removed. **d)** DL (EIR) and linear (bigstatsr) model performance (R2) for all 411 proteins measured by MS in The HOLBAEK Study. The error bars indicate the 95% confidence interval from 1000 bootstraps, and proteins with non-overlapping confidence intervals between DL and linear models are called significant and labeled in red. **e)** Performance gap between DL and linear models on genotype and covariates against the performance gap between non-linear (XGBoost) and linear models on covariates only. Results for all 411 proteins are shown, and proteins with significant performance gaps between DL and linear model are labeled in red.

Discussion

Here, we present a large-scale, systematic attempt to study non-linear genetic and covariate interactions that affect blood plasma protein levels. We used DL on genetics and plasma proteomics data of 48,594 individuals from the UK Biobank to identify proteins with underlying complex effects. While replicating the effect of many pQTLs from Sun et al., our results indicate

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

that many non-linear effects are present, illustrated by the increased performance of the deep learning models compared to linear models. Of the 2,922 measured proteins, we identified 138 proteins potentially regulated by non-linear effects like non-linear covariate effects (n=123), or genetic dominance and epistatic interactions between variants (n=15). Our modeling of non-linear relationships between covariates that influence protein levels were in line with previous reports[6,25]. Many associations have previously been established between plasma proteomics and demographic factors such as age, sex and BMI, and health indications such as liver function[6,10]. We show that complex relationships between non-genetic factors are widespread, and we speculate that, if we included additional covariates in our analysis (i.e., BMI) we could uncover biologically relevant non-linear relationships. This was outside the scope of the current study but will likely be the subject of future research. We demonstrate that genetic dominance within loci that affect protein levels is rare. This is consistent with previous studies that highlight the robustness of additive models when modeling human traits[18]. However, we identified a small group of proteins that are likely influenced by dominance effects in the *ABO* locus. Additionally, we could replicate epistatic interactions between the *ABO* locus and the *FUT2* secretor status that regulate plasma levels of intestinal proteins, as demonstrated before[6]. This shows that our approach can identify epistatic interactions in an unbiased fashion. We also uncovered novel interactions between the *ABO* and *FUT3* loci that influence plasma levels of the intestinally expressed protein FAM3D.

We uncovered complex effects that improve our understanding of biological pathways, which is a major focus in the V2F challenge. For example, we identified dominance effects of variants in the *ABO* locus on protein levels of CD209, CLEC4M and other proteins. The relationship between the *ABO* locus, specifically the *ABO* blood group system, and plasma abundance of proteins involved in the immune response, could advance our understanding of varying susceptibility to infectious diseases among individuals with different blood types[53–55]. Specifically, CD209 and CLEC4M act as attachment receptors for HIV-1 & 2, Ebola virus and other viral and bacterial pathogens[56,57]. Further, CD209 has been suggested to enhance ACE2-mediated SARS-CoV–2 infection[58]. Previous studies have also linked variants in the *ABO* locus with plasma levels of CD209[6,59]. Here, we demonstrate that individuals with A, B or AB blood-type have higher plasma levels of CD209 and CLEC4M. This suggests a mechanism where the *ABO* blood group system modulates pathogen recognition of dendritic cells through CD209 and CLEC4M, which could be an important

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

link between the *ABO* locus and its impact on susceptibility to infectious diseases. Detailed mechanisms of how either the *ABO* encoded alpha 1-3-Galactosyltransferase or the *ABO* blood group antigens regulate levels of CD209 and CLEC4M remain unclear.

To conduct our analyses, we developed the EIR-auto-GP software toolkit, designed to enable other researchers to apply our DL approach to their studies. The toolkit consists of a fully automated pipeline, allowing for the integration of genetic and covariate data for modeling on quantitative and binary traits. We emphasize that the built-in variant pre-filtering approaches in EIR-auto-GP allow for training on large-scale genetic data directly on CPUs—demonstrated by our ability to train DL models across ten cross-validation (CV) runs in three hours for a single protein on a 16-core computer on DNAnexus. This feature lowers the barrier for entry to research teams without access to high-cost hardware accelerators typically associated with DL.

Given that some proteins showed a gap in performance between EIR-auto-GP and the linear benchmark models, we wanted to examine what exact factors, e.g., complex effects in the covariate data, dominance effects and interaction effects between SNVs were driving the performance gaps. Therefore, we integrated the ability to fit both linear and non-linear models on different transformations of the input data into EIR-auto-GP. This included the use of covariate only data, additive and one-hot encoded genotype data, allowing us to specifically analyze the impact of these factors. For a more detailed analysis into the exact genetic components that might be driving the performance gaps, we used ordinary least squares (OLS) models for examining SNV genotypes separately as well as SNV-SNV interaction effects. We hope that EIR-auto-GP will help advance genetic research by providing an accessible DL toolkit to model complex genetic effects that influence molecular and disease traits, thereby addressing important aspects of the V2F challenge.

We initially trained and tested our models on individuals of self-reported white ethnicity, as this is the largest group of individuals with very similar ancestral backgrounds in the UK Biobank and the UKB-PPP[3,6]. When testing the models on sets of different self-reported ethnicity groups, we observed reduced performance for linear and DL models, which was potentially due to differences in population structure[46,60]. This study serves as a proof-of-concept of our approach to capture non-

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

linear effects that influence protein abundances, and it should be applied to more diverse cohorts in the future.

We replicated the effect of the identified epistatic *ABO-FUT3* interaction on plasma levels of FAM3D, MUC2 and ALPI, as well as the dominance effect of the *ABO* locus on plasma levels of CD209 and CLEC4M in the FinnGen cohort. This demonstrates that our findings are transferable beyond the UK-white population of the UK Biobank to other populations. However, given the 20-fold smaller sample size, the DL models might not effectively detect these effects (e.g., due to overfitting) adequately for it to be reflected in significantly better test set performance. Additionally, the increased uncertainty, as indicated by larger confidence intervals, when evaluating the models on a much smaller test set makes it challenging to identify significant results, despite better performance metrics. However, the increasing availability of larger proteomics cohorts will enable the identification of non-linear genetic and covariate effects on protein abundance in an unbiased, large-scale manner.

Limitations

A large fraction of the blood plasma proteins could not be accurately modeled from genotypes and the chosen covariates using either the linear or DL model. This may be due to missing causal SNVs, or unaccounted environmental factors, such as BMI and health and disease conditions, which are known to affect protein levels[6]. Interestingly, we found that many proteins that we could not model had a low correlation with SomaScan measurements[14]. The comparability between Olink and other protein quantification methods is highly debated in the field[14,61]. Our findings might indicate that plasma levels for some proteins measured by Olink may not be entirely accurate, which can affect the performance of our models in the UKB. Advances in MS-based proteomics could allow for higher specificity and quantitative accuracy of plasma proteomics in large sample sizes comparable to those of the UKB in the future[10,62].

Regarding modeling, we used a threshold on the pre-GWAS analysis to limit the number of variants used as input to the DL model. The choice is not guaranteed to be optimal for modeling purposes of all proteins, and this step might filter variants with purely complex effects not detected in GWAS. Furthermore, UKB array data were used to conduct the analysis, which might affect the

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

completeness of genetic variants analyzed. Additionally, we modeled using both the Olink provided NPX and INT transformed protein abundance values. Modeling using the NPX values was motivated by preserving effect sizes of the protein levels, while modeling on the INT values was done to identify potential false positives. These could be caused by the non-linear nature of the NPX values that could favor the DL models over the linear models, which we found to be the case for 33 of the proteins. However, performing the INT transformation removes the notion of scale in the data, effectively converting the data to ranks, which in itself impacts modeling. For instance, we found that both the DL and linear model had reduced R^2 when modeling on INT compared to NPX values, indicating that there was a general loss of information.

While our approach of using differences in performance gaps can identify protein levels modulated by interaction effects, it likely does not identify rare interactions. For example, using the OLS models we identified an interaction between variants in the *ABO* and *FUT3* loci which has a low frequency in the present cohort (~2%). These rare effects are unlikely to be learned during model training, and even if captured, may not significantly impact test set performance. This indicates that even the ~50,000 samples of the UKB-PPP might be too small to discover rare variant interactions using our approach.

Conclusions and future directions

While the majority of pQTL studies are performed using additive linear models, we demonstrate that non-additive, complex genetic effects can influence plasma protein levels. Modeling complex traits requires models that can learn from complex relationships in the input data. DL makes it possible to do such analysis and is not, as in our case, restricted to modeling plasma proteomics but can additionally be applied to model other molecular traits and environmental effects. Furthermore, such approaches can model covariate and environmental effects without specifying interaction terms a priori and could be used for discovering interaction effects such as ExE and GxE effects. Overall, we conclude that DL has provided additional value in understanding the complex genetic regulation of molecular traits and that discoveries of complex effects will likely scale with larger sample sizes and more diverse cohorts.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Methods

Experimental setup and processing of UK Biobank data

In the genomic quality control (QC) process, we utilized PLINK v1.90b7[63] for data analysis and filtering. Our dataset initially consisted of 784,256 autosomal variants and 488,377 individuals. We removed individuals with a relatedness factor >= 0.0884 (second degree relatedness), resulting in 453,581 individuals kept for the analysis. We applied the following QC filters: individuals with more than 10% missing genotype data (--mind 0.10) were excluded, resulting in the removal of 6 individuals, with 453,575 remaining. Variant level QC involved removing variants with more than 1% missing data (--geno 0.01), leading to the exclusion of 143,713 variants. Additionally, variants failing the Hardy-Weinberg Equilibrium test at a threshold of 0.000001 (--hwe 0.000001) were removed, accounting for 153,825 variants. We also applied a minor allele frequency (MAF) threshold of 0.005 (--maf), resulting in the exclusion of 62,621 variants. After the application of these QC steps, the final dataset comprised 424,097 variants and 453,575 individuals. We divided the individuals into train, validation, and test sets for the modeling. In the training dataset, we included exclusively individuals with self-reported 'UK-white' ethnicity from Olink batches 0-6. Batch 7 contains consortium selected individuals and individuals from the COVID-19 imaging study and do not follow UKB baseline characteristics[6]. To this end, individuals from batch 7 and all individuals with non-UK-white ethnicity were excluded from the training dataset. Additionally, individuals from batch 7 were excluded in the UK-white test set (n=1,771) used throughout the study. Access to the UK Biobank data was obtained through application 1251 "The metabolically healthy obese and metabolically obese normal-weight in the UK Biobank: Prevalence, genes and lifestyle contributors, disease risk and mortality".

Deep learning model training

The main deep learning models on the UKB were trained with the EIR-auto-GP toolkit (https://github.com/arnor-sigurdsson/EIR-auto-GP, commit fb41457). The ordinary least squares (OLS) estimation for allele effects and interaction effects was also done with the toolkit, as well as the direct estimation of protein levels as a function of genotype combinations (commit 0d5d762). Besides the genotype input data, categorical covariate inputs were sex, UKB center, UKB genetic array and whether individuals were consortium selected. Continuous inputs were age

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

and genetic principal components 1-20 (UKB data field 22009). Each protein level run consisted of 10 holdout cross validation (CV) runs, using a pre-defined validation set for consistency across runs. Despite repeated use of the same training-validation split, the models differed in each run due to several factors: (a) random initialization of the models; (b) the order of data during mini-batching during training was shuffled independently each run; and (c) while the first 3 CV runs shared a common set of SNVs, the subsequent 7 runs used a different set of SNVs as determined by a Bayesian optimization process (see below). For each protein level run, a GWAS pre-selection with a p-value of 0.001 was applied to the training set and used to reduce the number of variants input to the DL models. The first 3 CV runs used the full set of variants that passed the GWAS pre-selection step, and DL attributions were computed with integrated gradients[64] on the validation set. After the first 3 CV runs, a Bayesian optimization (BO) loop was applied to optimize for the top variants to include in the following 7 CV runs. The BO was implemented with scikit-optimize (v.0.9.0)[65], with the objective of optimizing the fraction of top SNVs regarding validation set performance. The top variants were defined by averaging the absolute DL attributions computed on the validation set across the first 3 CV runs. After training the DL models for all 10 CV runs, a final ensemble prediction was applied to the test set.

Linear model benchmarking

The training of linear benchmark models was done with bigsnpr (1.12.2)[33] and bigstatsr (1.5.12)[33]. All 424,097 variants were used as input for the model, as well as the covariates sex, age, UKB center, UKB genetic array, whether individuals were consortium selected and genomic principal components 1-20. The modeling was conducted with a 10-fold cross-validation (CV) employing a grid search for the α mixing parameter in the elastic net, exploring values [0.0001, 0.001, 0.01, 0.1, 1]. Additionally, the approach involves testing multiple values for the λ penalization parameter (default 200). Following this, an ensemble-like process across the CV runs was executed to generate the final model, which was subsequently assessed using the test set.

Model performance

The test set predictions of the trained DL and linear models were bootstrapped (n=1,000) and R^2 and RMSE calculated for each bootstrap generation using sklearn.metrics.r2_score and sklearn.metrics.mean_squared_error. From the resulting distribution, the 95% confidence intervals

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

were calculated using the 2.5% and 97.5% percentiles for each protein. Performance gaps were calculated for each protein by subtracting the mean bootstrapped R^2 of the linear models from the mean bootstrapped R^2 of the DL or non-linear models.

Self reported ethnic grouping

Individuals were stratified according to self-reported ethnic background in the UKB (data field 21000). The individual groups were consolidated into 5 main groups for the purpose of the study. The groups were defined as "White", "South Asian", "East Asian", "African", and "Caribbean". The "White" group included British, Irish, and other individuals with white backgrounds. The "South Asian" group comprised individuals of Indian, Pakistani, Bangladeshi. Those of Chinese self-reported ethnicities were categorized as "East Asian." "African" and "Caribbean" groups were kept as indicated in data field 21000.

Model Complexity Analysis

To examine which factors might be contributing to performance differences between the linear and DL models, we systematically explored various data configurations for the covariate and genotype input data. Specifically, we generated 5 different sets of input data configurations: tabular (covariate) data alone, additively encoded genotype data exclusively, one-hot encoded genotype data exclusively, additively encoded genotype with tabular data, and one-hot encoded genotype data with tabular data. The one-hot encoding was used to examine whether a linear model allowed to fit on genotypes separately would close the performance gap (e.g., due to effects in the data resembling dominance). For each of these five data configurations, we trained a linear Elastic Net model as well as a non-linear XGBoost model, resulting in 10 different data-model combinations. To limit the computational complexity, we limited the genotypes to the top 128 SNVs. These were selected base on the absolute DL attribution scores, which were computed on the validation set in the first 3 CV runs in the main experiments, then averaged. The same training, validation and test set splits were used as in the main experiments.

Genetic non-additivity analysis

Beyond examining performance differences between linear models when using an additive or one-hot encoding, we also fit Ordinary Least Squares (OLS) models on each of the top 128 SNVs,

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where the models were fit on each genotype separately (Target = $\beta_0 + \beta_1 \times SNV_1 + \beta_2 \times SNV_2 + \beta_3 \times SNV_3$ + covariates + ϵ). By examining the p-values and effect sizes coefficients assigned to each genotype, we could explore for each SNV whether it deviated from an additive relationship towards a protein level.

Genetic interaction analysis

In addition to investigating performance disparities between the linear and non-linear XGBoost models, we explored potential pairwise interactions among the top 128 SNVs. This entailed fitting Ordinary Least Squares (OLS) models on all possible pairwise combinations of SNVs, where each model utilized each SNV (one-hot encoded) as inputs along with the product interaction term between them (Target = $\beta_0 + \beta_{11} \times SNV1_1 + \beta_{12} \times SNV1_2 + \ldots + \beta_{21} \times SNV2_1 + \beta_{22} \times SNV2_2 + \ldots + \beta_3 \times (SNV1 \times SNV2)$ + covariates + ϵ). Across all traits, we tested a total of 1,389,888 pairs, and as such applied a p-value threshold of 0.05 / 1,389,888 = 3.6e-08. This approach allowed us to identify which SNV pairs might contribute most to any remaining performance gap between the linear model using one-hot encoded genotype data with tabular data and the XGBoost model trained on the same data.

Replication in MS-based proteomics data from The HOLBAEK Study

The HOLBAEK Study consisted of 2,147 children and adolescents (55% girls) between the age of 5 and 20, recruited from the Children's Obesity Clinic, accredited Centre for Obesity Management, Copenhagen University Hospital Holbæk, Denmark[66], and a population-based cohort recruited from schools in 11 municipalities across Zealand, Denmark[67]. Besides age, an eligibility criteria of the Obesity Clinic was BMI above the 90th percentile (BMI SDS >= 1.28) according to Danish reference values. The study protocol for The HOLBAEK Study was approved by the ethics committee for the Region Zealand (protocol no. SJ-104) and is registered at the Danish Data Protection Agency (REG-043-2013). The HOLBAEK Study including the obesity clinic cohort and the population-based cohort are also registered at ClinicalTrials.gov (NCT00928473). The MS based proteomics data consisted of 411 protein levels measured across 2,130 of the 2,147 samples, with genotype data available for 1,924 individuals featuring 5,242,958 variants after quality control and filtering[10]. Due to the imbalance in the number of features compared to the number of samples, we used PLINK to perform LD pruning (--indep-pairwise 50 5 0.8), reducing the variant

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

count to 998,505. We then matched and retained only those samples for which both genotype and phenotype data were available, identifying 1,893 samples with complete data sets. Following the EIR-auto-GP data processing pipeline, data splits were defined as 1,533 training, 170 validation, and 190 test samples. Besides the genotype data, covariates included were sex, BMI, age, time to analysis, MS batch information. Due to the smaller sample size compared to the UKB, we used 20-fold cross validation (CV) instead of the 10-fold applied in the UKB. EIR-auto-GP (commit 2934974) was used for DL model training and additionally, we found that the default feature selection approach in EIR-auto-GP (i.e., a fixed GWAS threshold and DL attribution based Bayesian optimization (BO) of included SNVs) was susceptible to overfitting in this dataset, based on training and validation set performance. To address this, we devised an alternative, simpler approach focusing on dynamic SNV inclusion based on GWAS p-value rankings. The optimization process began with seeding the algorithm with manual fractions, reflecting SNV subsets from the most significant (p-value threshold of 1e-8) to the least (up to a p-value of 1e-4). After this, the BO process to find the optimal fraction of SNVs was allowed to proceed. We found that this approach guided more efficiently towards using fewer SNVs, which resulted in better validation performance.

Replication in FinnGen Olink data

The FinnGen quality controlled Olink data consisted of 2,925 measured protein levels across 1,990 samples, with genotype data available for 520,210 individuals and 21,331,644 variants initially. The variants were filtered to match those used in the UKB experiments, resulting in a final set of 416,802 variants. Retaining only samples where genotype and phenotype data were available, our final set consisted of 1,757 samples. Data splits were defined as 1,231 training, 263 validation and 263 test samples. Besides the genotype data, covariates included blood sampling age, sex, genetic testing chip and batch, top 20 genetic PCs and protein examination batch. The DL model training was performed with EIR-auto-GP (commit c141b5a) and the training procedure was the same as described above for the data set from The HOLBAEK Study.

FinnGen Ethics statement

Study subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017.

The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020, THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020, THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021, THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021, Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20) TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases permission/extract from the meeting minutes on 4th July 2019.

The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 11 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, BB2021_65, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, HUS/430/2021 §28, §29, HUS/150/2022 §12, §13, §14, §15, §16, §17, §18, §23, §58, §59, HUS/128/2023 §18, Auria Biobank AB17-5154 and amendment #1 (August 17 2020) and amendments BB_2021-0140, BB_2021-0156 (August 26 2021, Feb 2 2022), BB_2021-0169, BB_2021-0179, BB_2021-0161, AB20-5926 and amendment #1 (April 23 2020) and it´s modifications (Sep 22 2021), BB_2022-0262, BB_2022-0256, Biobank Borealis of Northern Finland_2017_1013, 2021_5010, 2021_5010 Amendment, 2021_5018, 2021_5018 Amendment,

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2021_5015, 2021_5015 Amendment, 2021_5015 Amendment_2, 2021_5023, 2021_5023 Amendment, 2021_5023 Amendment_2, 2021_5017, 2021_5017 Amendment, 2022_6001, 2022_6001 Amendment, 2022_6006 Amendment, 2022_6006 Amendment, 2022_6006 Amendment_2, BB22-0067, 2022_0262, 2022_0262 Amendment, Biobank of Eastern Finland 1186/2018 and amendment 22§/2020, 53§/2021, 13§/2022, 14§/2022, 15§/2022, 27§/2022, 28§/2022, 29§/2022, 33§/2022, 35§/2022, 36§/2022, 37§/2022, 39§/2022, 7§/2023, 32§/2023, 33§/2023, 34§/2023, 35§/2023, 36§/2023, 37§/2023, 38§/2023, 39§/2023, 40§/2023, 41§/2023, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020), BB2021-0140 8§/2021, 9§/2021, §9/2022, §10/2022, §12/2022, 13§/2022, §20/2022, §21/2022, §22/2022, §23/2022, 28§/2022, 29§/2022, 30§/2022, 31§/2022, 32§/2022, 38§/2022, 40§/2022, 42§/2022, 1§/2023, Central Finland Biobank 1-2017, BB_2021-0161, BB_2021-0169, BB_2021-0179, BB_2021-0170, BB_2022-0256, BB_2022-0262, BB22-0067, Decision allowing to continue data processing until 31st Aug 2024 for projects: BB_2021-0179, BB22-0067,BB_2022-0262, BB_2021-0170, BB_2021-0164, BB_2021-0161, and BB_2021-0169, and Terveystalo Biobank STB 2018001 and amendment 25th Aug 2020, Finnish Hematological Registry and Clinical Biobank decision 18th June 2021, Arctic biobank P0844: ARC_2021_1001.

Acknowledgments

S.R. and M.M. were supported by the Novo Nordisk Foundation (NNF14CC0001). S.R., J.F.G., R.L. and T.H. were supported by the Novo Nordisk Foundation (NNF23SA0084103). S.R., R.L. and B.M.N. were supported by the Novo Nordisk Foundation (NNF21SA0072102). J.F.G. was supported by a research grant from the Danish Cardiovascular Academy, which is funded by the Novo Nordisk Foundation, grant number NNF20SA0067242 and The Danish Heart Foundation. This research has been conducted using the UK Biobank Resource under Application Number 1251. Our gratitude goes to all participants and their families from the UK Biobank and The HOLBAEK Study. We want to acknowledge the participants and investigators of the FinnGen study. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme LCC, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

AG, and Boehringer Ingelheim International GmbH. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/) and Arctic Biobank (https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-cohorts-and-arctic-biobank). All Finnish Biobanks are members of BBMRI.fi infrastructure (https://www.bbmri-eric.eu/national-nodes/finland/). Finnish Biobank Cooperative -FINBB (https://finbb.fi/) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the Fingenious® services (https://site.fingenious.fi/en/) managed by FINBB.

Author contributions

Conceptualization: S.R. Formal Analysis: A.I.S., J.F.G., Z.Y. Investigation: J.F.G., A.I.S. Methodology: A.I.S., J.F.G., J.M., K.R., R.T., H.W., R.A.J.S. Resources: S.R., R.L., T.H. Software: A.I.S. Supervision: S.R., R.J.F.L., T.H., B.V., A.G., B.M.N. Validation: Z.Y., A.G., L.N., M.M. Visualization: J.F.G., A.I.S. Writing – original draft: A.I.S., J.F.G, S.R. Writing – review & editing: A.I.S., J.F.G, S.R., R.A.J.S., R.J.F.L., J.M., K.R., R.T., H.W.

Competing interest

S.R. is the founder and owner of BioAI. The remaining authors declare no competing interests.

Code availability

EIR is available at https://github.com/arnor-sigurdsson/EIR. EIR-auto-GP is available at https://github.com/arnor-sigurdsson/EIR-auto-GP.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Data availability

UK Biobank genotype, proteomics and covariate data is available to approved researchers through the UK Biobank (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access). All analyses using UK Biobank data were performed at the Research Analysis Platform at DNAnexus (https://ukbiobank.dnanexus.com/). Combined summary statistics of the per-protein GWAS performed in this study (Supplementary Data) are available through Zenodo (https://doi.org/10.5281/zenodo.12654966). Individual-level genotypes and register data from FinnGen participants can be accessed by approved researchers via the Fingenious portal (https://site.fingenious.fi/en/) hosted by the Finnish Biobank Cooperative FinBB (https://finbb.fi/en/). Data release to FinBB is timed to the biannual public release of FinnGen summary results, which occurs 12 months after FinnGen consortium members can start working with the data. Data from The HOLBAEK Study is not publicly available due to the need to maintain privacy of study participants but is available on reasonable request. Searchable results are available online at proteomevariation.org.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021).
3. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
4. Shilo, S. *et al.* 10 K: a large-scale prospective longitudinal study in Israel. *Eur. J. Epidemiol.* **36**, 1187–1194 (2021).
5. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721 (2021).
6. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
7. Julkunen, H. *et al.* Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nat. Commun.* **14**, 604 (2023).
8. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
9. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature Metabolism* **2**, 1135–1148 (2020).
10. Niu, L. *et al.* Plasma proteome variation and its genetic determinants in children and adolescents. *bioRxiv* (2023) doi:10.1101/2023.03.31.23287853.
11. Niu, L. *et al.* Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat. Med.* **28**, 1277–1287 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

12. Gudjonsson, A. *et al.* A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nat. Commun.* **13**, 480 (2022).
13. Dhindsa, R. S. *et al.* Rare variant associations with plasma protein levels in the UK Biobank. *Nature* **622**, 339–347 (2023).
14. Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348–358 (2023).
15. Lourdusamy, A. *et al.* Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum. Mol. Genet.* **21**, 3719–3726 (2012).
16. Genetic control of the human brain proteome. *Am. J. Hum. Genet.* **108**, 400–410 (2021).
17. Ruffieux, H. *et al.* A fully joint Bayesian quantitative trait locus mapping of human protein abundance in plasma. *PLoS Comput. Biol.* **16**, e1007882 (2020).
18. Palmer, D. S. *et al.* Analysis of genetic dominance in the UK Biobank. *Science* **379**, 1341–1348 (2023).
19. Epistasis in sporadic Alzheimer's disease. *Neurobiol. Aging* **30**, 1333–1349 (2009).
20. Robson, K. J. H. *et al.* Synergy between the C2 allele of transferrin and the C282Y allele of the haemochromatosis gene (HFE) as risk factors for developing Alzheimer's disease. *J. Med. Genet.* **41**, 261–265 (2004).
21. Williams, S. M. *et al.* Combinations of Variations in Multiple Genes Are Associated With Hypertension. *Hypertension* (2000) doi:10.1161/01.HYP.36.1.2.
22. Elgart, M. *et al.* Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Commun Biol* **5**, 856 (2022).
23. Albiñana, C. *et al.* Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nat. Commun.* **14**, 1–11 (2023).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

24. Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease. *Cell Genomics* **2**, 100086 (2022).

25. McCaw, Z. R. *et al.* DeepNull models non-linear covariate effects to improve phenotypic prediction and association power. *Nat. Commun.* **13**, 241 (2022).

26. 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).

27. Sigurdsson, A. I. *et al.* Deep integrative models for large-scale human genomics. *Nucleic Acids Res.* (2023) doi:10.1093/nar/gkad373.

28. Wu, Z. *et al.* Two-stage joint selection method to identify candidate markers from genome-wide association studies. *BMC Proc.* **3 Suppl 7**, S29 (2009).

29. Piriyapongsa, J. *et al.* iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. *BMC Genomics* **13 Suppl 7**, S2 (2012).

30. Motzo, C. *et al.* Heterogeneity in the magnitude of the insulin gene effect on HLA risk in type 1 diabetes. *Diabetes* **53**, 3286–3291 (2004).

31. Sigurdsson, A. I. *et al.* Improved prediction of blood biomarkers using deep learning. *medRxiv* (2022).

32. Snaebjarnarson, A. S. *et al.* Complex effects of sequence variants on lipid levels and coronary artery disease. *Cell* **186**, 4085–4099.e15 (2023).

33. 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).

34. Sharma, N. *XGBoost. The Extreme Gradient Boosting for Mining Applications*. (GRIN

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Verlag, 2018).

35. Tabular data: Deep learning is not all you need. *Inf. Fusion* **81**, 84–90 (2022).

36. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? (2022).

37. Geijtenbeek, T. B. H. & Gringhuis, S. I. Signalling through C-type lectin receptors: shaping immune responses. *Nat. Rev. Immunol.* **9**, 465–479 (2009).

38. Guo, Y. *et al.* Structural basis for distinct ligand-binding and targeting properties of the receptors DC-SIGN and DC-SIGNR. *Nat. Struct. Mol. Biol.* **11**, 591–598 (2004).

39. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* **4**, e1000072 (2008).

40. Paré, G. *et al.* Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genet.* **4**, e1000118 (2008).

41. Wolpin, B. M. *et al.* Pancreatic cancer risk and ABO blood group alleles: results from the pancreatic cancer cohort consortium. *Cancer Res.* **70**, 1015–1023 (2010).

42. Groot, H. E. *et al.* Genetically Determined ABO Blood Group and its Associations With Health and Disease. *Arterioscler. Thromb. Vasc. Biol.* **40**, 830–838 (2020).

43. Marcus, D. M. The ABO and Lewis blood-group system. Immunochemistry, genetics and relation to human disease. *N. Engl. J. Med.* **280**, 994–1006 (1969).

44. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

45. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

46. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

populations. *Nat. Commun.* **10**, 3328 (2019).

47. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
48. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 373 (2022).
49. Wang, Y., Tsuo, K., Kanai, M., Neale, B. M. & Martin, A. R. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu Rev Biomed Data Sci* **5**, 293–320 (2022).
50. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
51. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
52. Moreno-Estrada, A. *et al.* Reconstructing the Population Genetic History of the Caribbean. *PLoS Genet.* **9**, e1003925 (2013).
53. Cserti, C. M. & Dzik, W. H. The ABO blood group system and Plasmodium falciparum malaria. *Blood* **110**, 2250–2258 (2007).
54. Wu, S.-C. *et al.* Blood group A enhances SARS-CoV-2 infection. *Blood* **142**, 742–747 (2023).
55. Nordgren, J. & Svensson, L. Genetic Susceptibility to Human Norovirus Infection: An Update. *Viruses* **11**, (2019).
56. Geijtenbeek, T. B. H. *et al.* Identification of different binding sites in the dendritic cell-specific receptor DC-SIGN for intercellular adhesion molecule 3 and HIV-1. *J. Biol. Chem.* **277**, 11314–11320 (2002).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

57. Lin, G. *et al.* Differential N-linked glycosylation of human immunodeficiency virus and Ebola virus envelope glycoproteins modulates interactions with DC-SIGN and DC-SIGNR. *J. Virol.* **77**, 1337–1346 (2003).

58. Lempp, F. A. *et al.* Lectins enhance SARS-CoV-2 infection and influence neutralizing antibodies. *Nature* **598**, 342–347 (2021).

59. Anisul, M. *et al.* A proteome-wide genetic investigation identifies several SARS-CoV-2-exploited host targets of clinical relevance. *Elife* **10**, (2021).

60. Gomez, F., Hirbo, J. & Tishkoff, S. A. Genetic variation and adaptation in Africa: implications for human evolution and disease. *Cold Spring Harb. Perspect. Biol.* **6**, a008524 (2014).

61. Katz, D. H. *et al.* Proteomic profiling platforms head to head: Leveraging genetics and clinical traits to compare aptamer- and antibody-based methods. *Sci Adv* **8**, eabm5164 (2022).

62. Bader, J. M., Albrecht, V. & Mann, M. MS-Based Proteomics of Body Fluids: The End of the Beginning. *Mol. Cell. Proteomics* **22**, 100577 (2023).

63. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

64. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. (2017).

65. scikit-optimize/scikit-optimize. doi:10.5281/zenodo.5565057.

66. Holm, J.-C. *et al.* Chronic care treatment of obese children and adolescents. *Int. J. Pediatr. Obes.* **6**, 188–196 (2011).

67. Vissing Landgrebe, A. *et al.* Population-based pediatric reference values for serum parathyroid hormone, vitamin D, calcium, and phosphate in Danish/North-European white

medRxiv preprint doi: https://doi.org/10.1101/2024.07.04.24309942; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

children and adolescents. *Clin. Chim. Acta* **523**, 483–490 (2021).