

Non-linear genetic regulation of the blood plasma proteome

Authors

Arnor I. Sigurdsson^{1,2,3,*}, Justus F. Gräf^{1,2,3,*}, Zhiyu Yang⁴, 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 Niu¹, Matthias Mann^{1,7}, FinnGen, Bjarni Vilhjalmsson^{3,8,9}, Benjamin M. Neale^{3,10,11}, Andrea Ganna^{4,10,11,12}, Torben Hansen², 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)

34 **Abstract**

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

52

53 **Introduction**

54

55 Genome-wide association studies (GWAS) have identified thousands of associations between
56 genetic variants and phenotypic traits¹. Despite these discoveries, it remains a significant challenge
57 to understand how these genetic variants contribute to the phenotypic traits. This is mainly because
58 the functional impact of many variants is still unknown. This area of focus in modern genetic
59 research is known as Variant-2-Function (V2F). Addressing the V2F challenge is crucial for
60 identifying how genetic variants influence biological pathways, which can improve our
61 understanding of disease mechanisms and allow for more precise drug development².

62

63 In response to this challenge, multi-omics approaches have been applied to population-based
64 cohorts, typically including transcriptomics, proteomics, metabolomics, or microbiomics³⁻⁸. Blood
65 plasma, in particular, serves as an easily accessible and minimally invasive sample for diagnostics
66 and biomarker discovery. Large-scale blood metabolomics are now available from resources like
67 the UK Biobank, further enriching our understanding of the genomic basis of complex traits⁷.
68 Moreover, proteomics-based analyses using either aptamer-based (SomaScan), antibody-based
69 (Olink) or mass spectrometry (MS)-based assays of individual-level biobank samples have
70 revealed thousands of protein quantitative trait loci (pQTLs)^{5,6,9-14}. These can bridge the gap
71 between genetic variants and phenotypes and allow for deeper functional understanding of
72 diseases, improving drug target discovery and contributing to our understanding of genetic effects
73 on disease¹⁵⁻¹⁷. Currently, GWAS has been the most widely applied methodology for discovery of
74 pQTLs. However, GWAS often assumes an additive model and might not fully recapitulate
75 complex, non-additive effects among the variants and relevant covariates. Despite its robustness,
76 it has been shown that deviations from the additive model exist in a number of human loci, for
77 example in the form of dominance effects¹⁸. Additionally, interactions of two or more variants can
78 result in a larger effect on a phenotype than the effect of each single variant, a concept known as
79 epistasis, which can also contribute to non-linear genetic architecture of complex traits¹⁹⁻²¹.

80

81 DL models can capture non-linear effects, which has motivated recent work in applying DL and
82 other non-linear models for both genetic prediction and variant-phenotype association, providing
83 new insights into the genetic architecture of complex traits²²⁻²⁶. For example, in previous studies,

84 we developed and applied DL frameworks for disease prediction in the UK Biobank²⁷, and found
85 potential dominance and epistatic effects, specifically for immunological diseases such as type 1
86 diabetes (T1D), involving the insulin gene and *HLA-DQB1*²⁸⁻³⁰. These complex effects also
87 transfer to molecular quantitative trait loci, as indicated by our previous analysis of 34 common
88 biomarkers in the UK Biobank³¹. Additionally, targeted discovery of epistatic effects between
89 genetic variants uncovered the presence of interactions between the *ABO* blood group and the
90 *FUT2* secretor status that influence blood plasma abundance of gastrointestinal (GI) proteins⁶.
91 Unbiased approaches have been used to identify numerous epistatic and dominance effects that
92 influence the plasma levels of lipids and their effects on cardiovascular diseases³². However, by
93 now, there have not been attempts to systematically characterize non-linear effects that influence
94 blood plasma protein levels.

95
96 Here, we present a systematic, DL-based workflow that allows us to identify non-linear effects
97 like non-linear covariate effects, dominance, and epistasis that influence plasma protein levels. To
98 demonstrate the use of our approach, we examined data from 2,922 blood plasma protein levels
99 measured in 48,594 individuals from the UK Biobank Pharma Proteomics Project (UKB-PPP).
100 Our study presents a nuanced view of non-additive effects that influence plasma proteomics in the
101 UK Biobank. We could illustrate the quantitative and qualitative non-linear regulation of the blood
102 plasma proteome and reveal novel non-linear effects that are highly likely to influence plasma
103 protein abundance. Using our approach, we identified 138 proteins, among which 123 were
104 potentially regulated by non-linear covariate effects and 15 by dominance or epistasis effects. This
105 highlights DL as a useful tool to uncover complex effects that influence molecular quantitative
106 traits, which can contribute to our understanding of the genetic architecture of complex traits.

107 **Results**

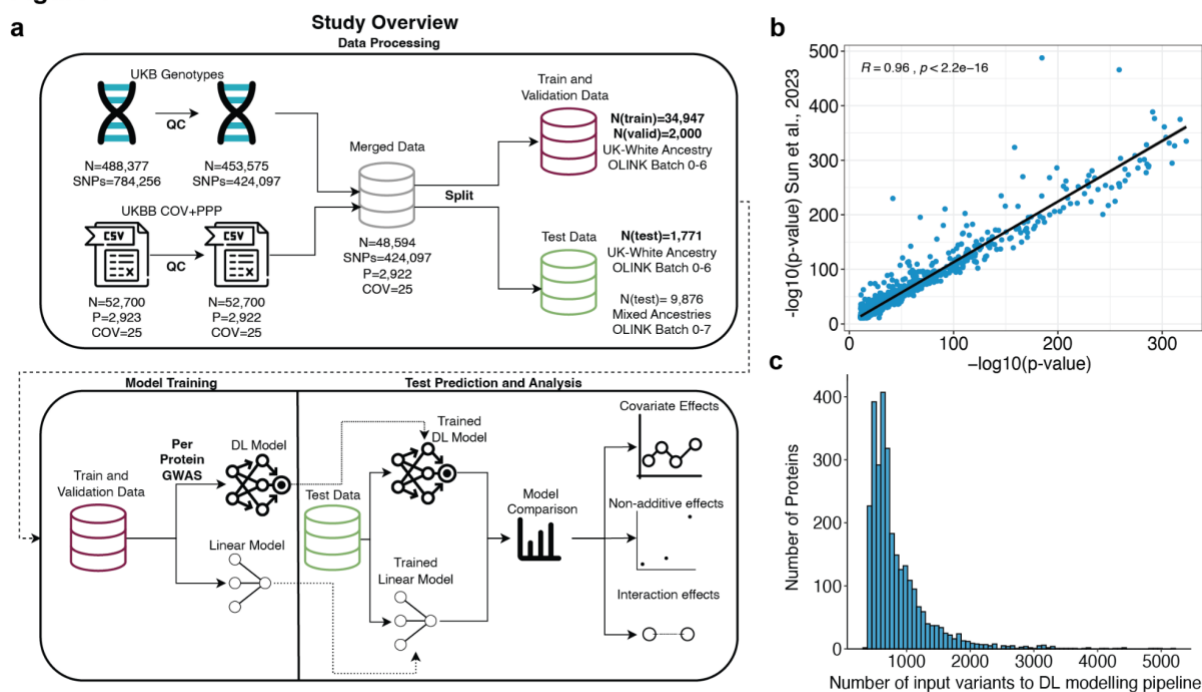
108

109 **Modeling blood plasma protein levels using deep learning**

110 To investigate the scale of non-linearity of genetic control of protein abundances in the blood
111 plasma, we modeled the abundance of 2,922 proteins in the blood plasma proteome in the UKB
112 cohort. Based on our deep learning framework, EIR, and the genome-local net deep learning
113 architecture (GLN)²⁷, we developed an automated framework, EIR-auto-GP, to predict the
114 abundance of a protein from genotypes and covariates (age, sex, UKB center, UKB genetic array,
115 whether an individual was consortium selected and genetic principal components 1-20 (**Figure**
116 **1a**). We used grouped, self-reported ethnicities (**Methods**) that resembled the distribution of the
117 genetic population structure in the UKB to subset individuals of UK-white self-reported ethnicity
118 as the largest group of individuals with similar ancestral background for model training and testing
119 (**Supplementary Figure 1a**). Subsequently, after quality control (QC) of the proteomics data, the
120 remaining 48,594 individuals were split into a train (n=34,947), validation (n=2,000) and test split
121 (n=1,771) (**Figure 1a**). As input to EIR-auto-GP, we used 424,097 measured QC-passed
122 genotypes, which reduced computational complexity compared to the more extensive imputed
123 data. Furthermore, we limited the amount of input variants by using the training dataset to conduct
124 GWAS for each protein and selecting associated variants (**Supplementary Note 1**). When
125 analyzing the results of the per-protein GWAS we found them to be consistent overall with
126 previous work⁶, and overlapping variants showed high correlation (**Figure 1b**). To determine input
127 variants for the DL modeling, we used a less stringent p-value threshold than usually applied to
128 GWAS (P<0.001), resulting in most DL models being trained on $\leq 1,000$ variants (**Figure 1c**).
129 Taken together, the variants identified through our GWAS and subsequently used as inputs for the
130 DL models were likely pQTL candidates.

131

Figure 1



132
 133 **Figure 1. Overview of the study and protein pre-GWAS results. a)** Overview of study design and workflow.
 134 UKB genotypes underwent quality control (QC), resulting in 424,097 QC-passed SNVs. The data were split into
 135 training and validation sets of self-reported UK-white ethnicity, OLINK batch 0-6 (n=34,947 for training; n=2000 for
 136 validation), and test sets stratified by ethnicity and batch: UK white self-reported ethnicity (n=1,771) and mixed
 137 ethnicities (n=9,876), all from OLINK batches 0-6 and 0-7 respectively. The training and validation data were used
 138 to develop DL and linear models, with a per-target GWAS on the training set used to pre-filter input variants for
 139 training the DL model. Finally, predictions and analyses were performed on the test data, and proteins that had
 140 discordant performance between the DL and linear models were investigated for non-linear covariate, non-additive
 141 (e.g., dominance), and interaction (e.g., epistasis) effects. **b)** Correlation of GWAS P-values between the current study
 142 and Sun et al.⁶. Variants with p-values exactly 0, likely due to being below the numerical precision threshold
 143 (underflow), were omitted from the plot. The scatter plot represents the $-\log_{10}(\text{p-values})$ correlation of 1,780
 144 overlapping genetic variants with significant associations ($p < 1.7e-11$) between our analysis and Sun et al.⁶. The strong
 145 correlation ($R=0.96, P < 2e-308$) between p-values demonstrates consistency in identifying significant associations. **c)**
 146 Histogram of the number of input SNVs used for DL model training following per target GWAS pre-filtering, where
 147 only SNVs with p-values < 0.001 (computed on the training set) were considered. For the majority of proteins, fewer
 148 than 1,000 SNVs passed the threshold.

149
 150 **Non-linear effects that influence blood plasma protein abundance**

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

153 an association between proteins with low modeling performance ($R^2 < 0.1$) and the correlation of
154 their measurements with SomaScan measurements in an Icelandic cohort¹⁴ (**Supplementary Note**
155 **2**). To investigate how many blood plasma proteins could be influenced by non-linear covariates
156 and genetic effects, we compared the performance of the DL models (EIR-auto-GP) to a penalized
157 linear model (bigstatsr)³³. We calculated the difference in model performance on the UK-white
158 test set ($n=1,771$) for each protein (**Supplementary Table 1**). For 1,503 of 2,922 proteins (51.4%),
159 the DL model performed better, resulting in a significant difference when modeling plasma protein
160 abundance from genotypes and covariates (paired T-test, two-sided, $t=11.281$, $P=6.4e-29$). To
161 identify specific proteins for which the DL model was significantly better, we bootstrapped the
162 predictions of each protein (**Methods**) and identified 171 proteins (5.8%) with a significant
163 performance increase (non-overlapping 95% confidence intervals) (**Figure 2a, Supplementary**
164 **Figure 2b**). These proteins showed a median increase in R^2 of 0.038 (mean 0.05). To examine
165 whether these results transferred to other metrics, we additionally used Root Mean Squared Error
166 (RMSE) to assess model performance. Among the 171 proteins showing better performance with
167 the DL model as measured by R^2 , the RMSE analysis also found the DL model outperforming on
168 all of these. Specifically, 28 of these proteins also showed significant improvement (non-
169 overlapping confidence intervals). In summary, we replicate that linear models are robust in
170 modeling plasma abundance of measured proteins^{6,18} and that our DL approach can identify
171 candidate proteins with potential non-linear effects that influence their plasma levels.

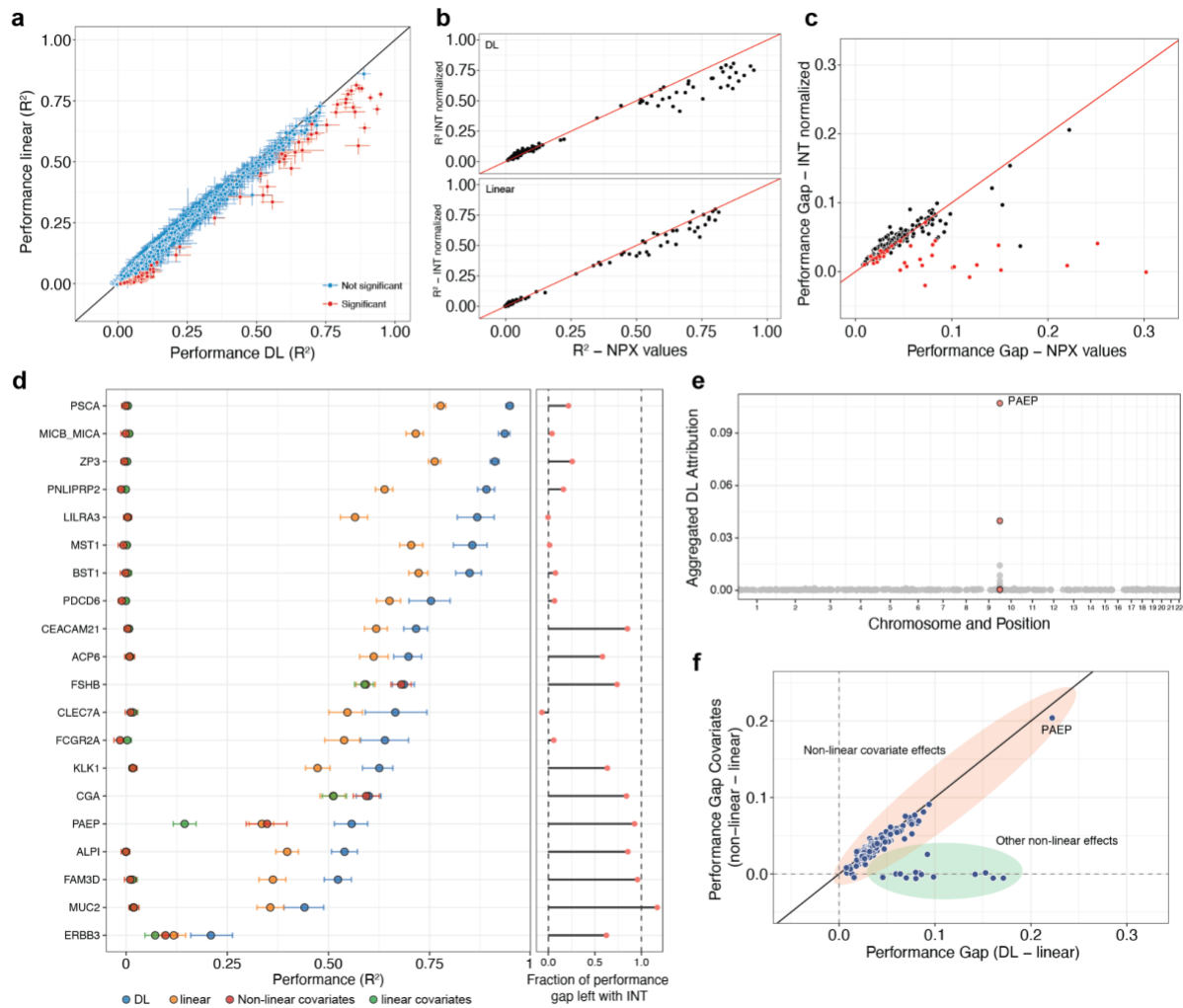
172

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

174 To preserve most of the protein level variance, we modeled the Olink protein expression values
175 (NPX); however, these are non-normal distributed and on a \log_2 -like scale⁶. This could favor the
176 DL model over the linear model without biological non-linear effects because a \log_2 transformation
177 can make a fundamentally linear relationship appear non-linear. Therefore, we re-ran our models
178 for the 171 proteins using Inverse-rank Normal Transformed (INT) protein levels and found
179 reduced R^2 for both DL and linear models, suggesting that information was lost when rank-
180 transforming the protein levels (**Figure 2b, Supplementary Table 2**). Despite this reduction in
181 performance, we found that for 138 (81%) of the 171 proteins, a significant gap in performance
182 between the DL and linear models remained (**Figure 2c**). Conversely, for 33 of the 171 proteins
183 the difference between the DL and linear models was not significant anymore (**Figure 2c**). Where

184 indicated, these 33 proteins were excluded from downstream analyses. The performance gap of
 185 most of the remaining proteins correlated well between NPX and INT normalized values,
 186 indicating that the DL model also identified non-linear effects on INT normalized protein values
 187 (Figure 2c).
 188

Figure 2



189
 190 **Figure 2. Deep learning reveals non-linear genetic and covariate effects.** **a** DL (EIR) and linear (bigstatsr) model
 191 performance (R^2) for all 2,922 proteins (**Supplementary Table 1**). The error bars indicate the 95% confidence
 192 intervals (CI) from 1,000 bootstraps, and proteins with non-overlapping confidence intervals between DL and linear
 193 models are called significant and labeled in red. **b** DL (top) and linear (bottom) model performance of 171 significant
 194 proteins modeled on raw protein expression values (NPX) or INT normalized protein values (**Supplementary Table**
 195 **2**). **c** Performance gap ($R^2 - R^2$) for the 171 significant proteins between DL and linear models (DL-linear) on NPX or
 196 INT normalized protein values. Proteins labeled in red indicate that no significant performance gap (overlapping CIs)

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

208

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

210 We then investigated the proteins with the largest absolute increase in performance by either
211 method (top 20) (**Figure 2d**). For these proteins, the DL model reached an R^2 between 0.21
212 (ERBB3) to 0.95 (PSCA) on the test set (**Figure 2d**). Among these 20 proteins, 11 showed a
213 maintained significant performance gap with the INT normalization. To investigate the
214 contribution of covariates on the performance, we trained and evaluated linear and non-linear
215 (XGBoost³⁴) models using only the covariates (**Supplementary Table 3**). We found that for 8 of
216 these 11 proteins, the genotype data was the main driver of model performance (**Figure 2d**). For
217 example, the DL performance of PSCA and FAM3D was mainly driven by known cis- and trans-
218 pQTL (**Supplementary Figure 2e**), which was expected due to their high association in previous
219 studies⁶. For some proteins, for instance, MICB/A and LILRA3, where genetics primarily
220 contributed to the DL model performance, there was no difference in performance between the DL
221 and linear model when modeling on INT normalized values (**Figure 2d, right panel**). However,
222 other proteins showed sustained performance gaps when modeling on INT normalized values,
223 providing additional confidence that the increased DL performance might be caused by non-linear
224 genetic effects (e.g., CEACAM21, ALPI, FAM3D, or MUC2).

225

226 **Non-linear covariate effects influence protein levels**

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

230 covariates sex and age, related to the age of menopause (**Figure 2e, Supplementary Figure 2f**)⁶.
231 Furthermore, for progesterone-associated endometrial protein (PAEP), we found that a
232 combination of covariates and genetics could account for the increased model performance
233 (**Figure 2d**). Besides age and sex related non-linear effects (**Supplementary Figure 2f**), PAEP
234 levels were influenced by cis-pQTLs which contributed to model performance (**Figure 2e**). When
235 expanding the analysis to all 138 proteins with significant differences, we found that for 123
236 performance gap between DL and linear models could be explained by non-linear covariate effects
237 (**Figure 2f**). These results indicate that we could robustly identify proteins with plasma levels that
238 underlie non-linear covariate effects, which was in line with the non-linear modeling of covariates
239 in other studies²⁵. However, it also revealed that for a substantial fraction of the 138 proteins
240 (10.9%, 15 proteins) effects in the covariates could not account for the increased performance of
241 the DL model to the linear model (**Figure 2f**). This suggests that the improved predictive accuracy
242 obtained with the DL model was not solely due to non-linear covariate effects.

243

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

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

261 their plasma levels. We found that two variants in the *ABO* locus (rs505922, rs8176719, chr 9)
262 showed dominance effects on protein levels for CD209, CLEC4M, FAM3D, ALPI, ABO and
263 MUC2 (**Supplementary Figure 3a**). CD209 is part of the C-type lectin family and is involved in
264 cell adhesion and pathogen recognition³⁷. It is highly similar to CLEC4M in function and
265 sequence. The two genes are located nearby on chr 19^{37,38} and are referred to as *DC-SIGN* and *DC-*
266 *SIGNR*, respectively. Notably, the two variants, rs505922 and rs8176719, were used to impute the
267 blood-types of the *ABO* blood group system in the UKB³⁹⁻⁴², which is known to have co-
268 dominance effects of its A and B alleles. Consistent with this, and the non-additive analyses above
269 (**Figure 3a**), we found dominant blood group effects on the plasma levels of CD209 and CLEC4M
270 (**Figure 3b-c, Supplementary Figure 3b**). We assessed the influence of the dominance effect on
271 model performance by training linear models for CD209 and CLEC4M using genotype and
272 covariate data and one-hot encoded either rs8176719, rs505922 or both. We found that by one-hot
273 encoding these variants the model performance improved by R^2 0.03 (7%), 0.0396 (9.21%) and
274 0.0399 (9.28%) (**Supplementary Figure 3c**). Taken together, using our approach, we could
275 identify varying levels of dominance within loci that regulate plasma protein levels.

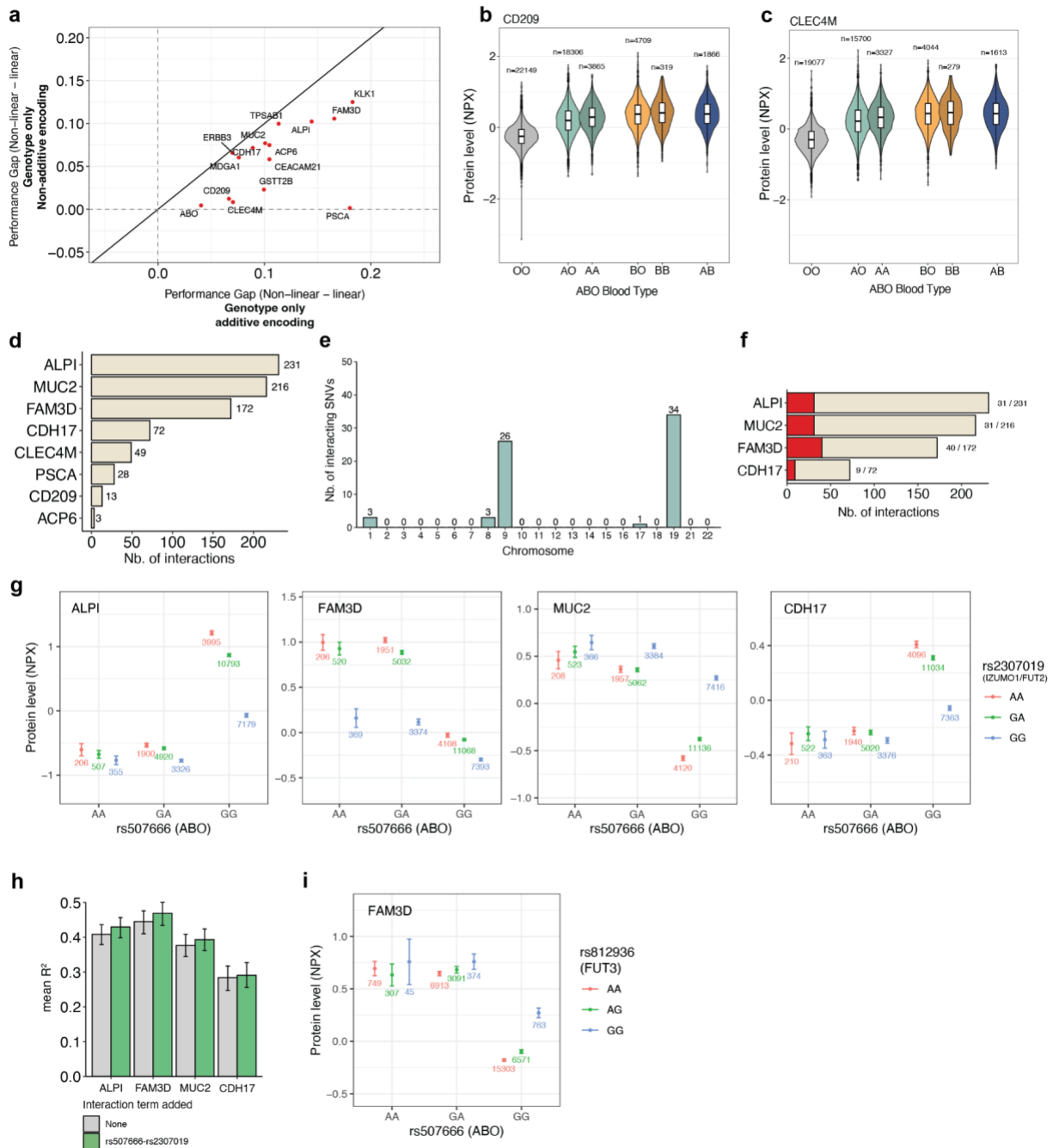
276

277 **Non-linear interactions between genetic variants affect protein levels**

278 Following the previous results, we further investigated proteins where the increased R^2 could not
279 be explained by non-linear covariate effects to identify potential epistatic SNV-SNV interactions.
280 For each of these 15 proteins, we analyzed the 128 SNVs with the highest feature importance in
281 the DL models on the validation set. To achieve this, we applied pairwise Ordinary Least-Squares
282 (OLS) models to the training set (n=34,947) to identify epistatic interactions. Restricting our
283 analysis to SNV pairs on different chromosomes, we identified at least one significant (p-value
284 $<4.46e-08$) interaction for 8 of the 15 proteins and a total of 784 interactions between 67 unique
285 SNVs on 5 chromosomes (**Figure 3d-e, Supplementary Table 5**). The majority of these
286 interactions (753, 96%) were between variants on chr 9 and chr 19 and most of the interacting
287 variants were located near the *ABO* and *FUT2* loci on chr 9 and 19, respectively (**Figure 3e,**
288 **Supplementary Figure 3d-e**). For instance, we identified most interactions for ALPI, MUC2,
289 FAM3D and CDH17 with 230, 216, 172 and 72 interactions, of which 31, 31, 40 and 9 were
290 between variants within the *ABO* and the *FUT2* locus (± 10 kb) (**Figure 3f**). We found an epistatic
291 interaction between the *ABO* variant rs507666 and rs2307019, a variant in the *IZUMO1* gene, 40kb

292 downstream of the *FUT2* locus, that influenced plasma levels of these proteins (**Figure 3g**). The
293 variant rs2307019 was in moderate linkage disequilibrium ($R^2=0.35$) with rs601338 (Trp154Ter),
294 a variant that determines *FUT2* secretor status used in Sun et al., and Snaebjarnarson et al.^{6,32}
295 (**Supplementary Figure 3f**). We thus expected that the interaction between rs507666 and
296 rs2307019 resembled an interaction between the *ABO* and *FUT2* locus. In line with our studies,
297 Sun et al. found that epistatic interactions between the *ABO* and *FUT2* locus have a strong
298 influence on the blood plasma levels of ALPI, MUC2, and FAM3D⁶. To assess if the interaction
299 between rs507666 and rs2307019 influenced modeling performance when predicting plasma
300 proteins levels, we trained linear models using one-hot encoded genotypes and covariates and
301 added an interaction term for rs507666-rs2307019 to predict levels of ALPI, FAM3D, MUC2 and
302 CDH17. We found that, when adding the single interaction term, the linear models improved by
303 R^2 0.021 (5.1%), 0.024 (5.3%), 0.017 (4.5%) and 0.007 (2.3%) respectively, indicating that this
304 epistatic interaction accounts for a substantial fraction of model performance (**Figure 3h**). In
305 summary, these results demonstrate that EIR-auto-GP could identify proteins with epistatic
306 interactions between genetic variants, which we could subsequently validate using targeted OLS
307 models.
308

Figure 3



309
310 **Figure 3. Identification of dominance effects and epistatic interactions between genetic variants on separate**
311 **chromosomes that influence protein levels. a)** Performance gap (R²-R²) between non-linear (XGBoost) and linear
312 models trained and tested on additive or non-additive encoded genotype data for 15 candidate proteins with potential
313 non-linear genetic effects. **b)** NPX protein levels of CD209 (n=51,214) in individuals in the UK Biobank, stratified
314 by their imputed *ABO* blood group (field p23165)^{39–42}. **c)** NPX protein levels CLEC4M (n=44,040) in individuals in
315 the UK Biobank, stratified by their imputed *ABO* blood group. **d)** Number of identified SNV-SNV interactions

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

330

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

332 In addition to the previously reported interactions between *ABO* and *FUT2* above, we identified
333 an interaction between *ABO* (rs507666) and the *FUT3* locus (rs812936) that influenced the blood
334 plasma levels of FAM3D, which is expressed in the gastrointestinal tract (**Figure 3i**). The *FUT3*
335 locus is also known as the Lewis gene, and encodes an alpha(1,3/4)-fucosyltransferase as part of
336 the Lewis antigen system⁴³. rs812936-A, was associated with increased levels of FUT3 plasma
337 levels⁴⁴ and led to decreased FAM3D plasma levels when interacting with rs507666-G in the *ABO*
338 locus (**Figure 3i**). This suggested that the protein level of FAM3D was not only regulated by
339 epistatic interactions between *ABO* and *FUT2*, but also dependent on interactions between *ABO*
340 and *FUT3* variants. As *FUT2* and *FUT3* are located 45 Mbp apart on chr 19 this was likely not
341 caused by LD between the two genes (**Supplementary Figure 3e**). Other proteins influenced by
342 *ABO-FUT2* interactions were enriched for gastrointestinal (GI) expression and may be perturbed
343 in GI disease⁶. We identified *ABO-FUT3* interactions for FAM3D, a GI expressed protein, and
344 thus speculated that *FUT3* could also be involved in regulating plasma abundance of GI expressed
345 proteins. However, when we added this interaction term to a linear model predicting FAM3D
346 levels, the performance did not improve (**Supplementary Figure 3g**). Notably, this interaction
347 was relatively rare in the UKB-PPP with 763 individuals in the training set, 47 in the valid set and
348 37 in the test set that carried this interaction. This indicates that rare interactions might be relevant
349 in regulating plasma protein levels, but that they were difficult for our DL model to detect because

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

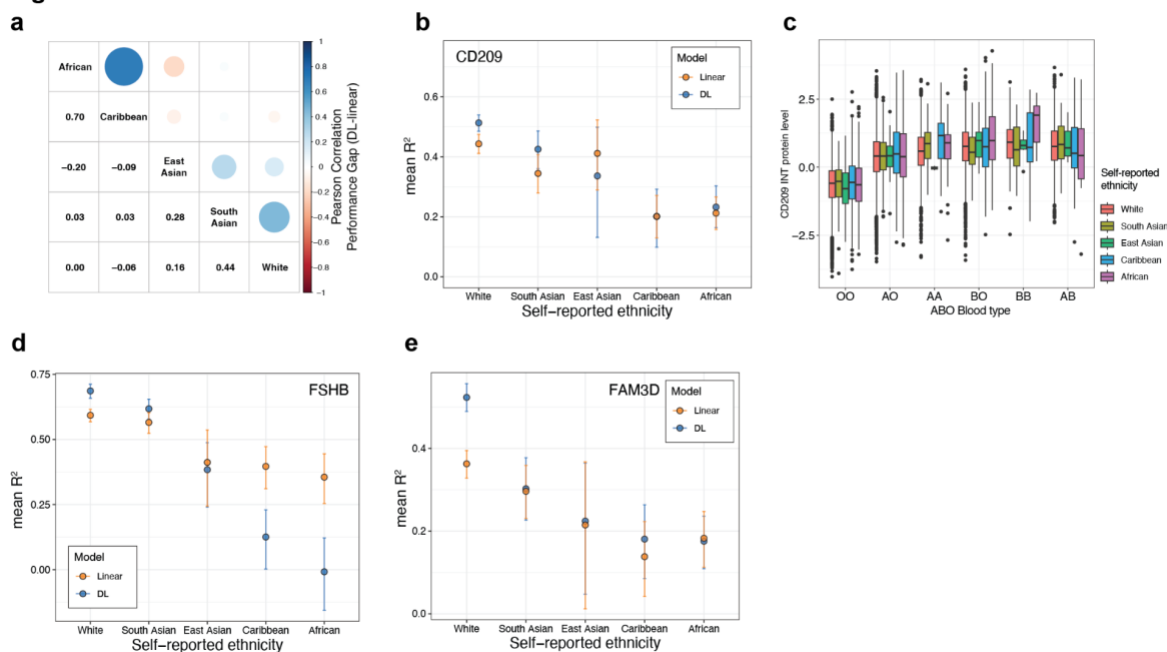
352

353 **Variable non-linear improvements across self-reported ethnicities**

354 Given the importance of understanding model performance across diverse ethnic groups⁴⁵⁻⁴⁹, we
355 investigated how the performance of the linear and DL models trained and evaluated on self-
356 reported UK white ethnic background transferred to other self-reported ethnic backgrounds
357 (“South Asian”, “East Asian”, “African” and “Caribbean”) in the UKB (**Methods**). We generally
358 observed a decline in performance for both DL and linear models across the non-UK white ethnic
359 groups, accompanied by larger confidence intervals for the 138 proteins with non-linear effects
360 (**Supplementary Figure 4a-b, Supplementary Table 6**). Overall, the linear model transferred
361 better to the ‘East Asian’ and ‘South Asian’ test set than the DL models, while the DL models
362 transferred better to ‘African’ and ‘Caribbean’ test sets than the linear models (**Supplementary**
363 **Figure 4a-b**). One contributing factor could be the breakdown and formation of LD patterns across
364 populations, as models trained on the UK white group may select tagging variants that do not
365 replicate in the other ethnic groups^{48,50}. Additionally, the larger confidence intervals could also be
366 partly due to the smaller number of samples available in ethnicity test sets. We observed correlation
367 of performance gaps (mean bootstrapped R^2 DL-linear) between all ethnicity groups, except
368 between ‘East Asian’ and ‘African’ and ‘Caribbean’ (**Figure 4a**). For instance, the DL model
369 outperformed the linear model for CD209 on the ‘South Asian’ and ‘African’ test set, while it
370 showed similar or worse performance on ‘East Asian’ and ‘Caribbean’ test sets (**Figure 4b**).
371 Despite this, CD209 levels showed similar trends when stratified by *ABO* blood type between the
372 different ethnicity groups in the whole UKB-PPP (n=52,700) (**Figure 4c**). Notably, the distribution
373 of *ABO* blood types was different between the ethnic groups in the UKB-PPP, which could
374 influence the performance on the different test sets (**Supplementary Figure 4c**). Above, we found
375 that the DL model could identify non-linear relationships between age and sex for FSHB (**Figure**
376 **2d, Supplementary Figure 2f**) and we found that this could be replicated in the ‘South Asian’,
377 but not in the ‘East Asian’, ‘Caribbean’ or ‘African’ test sets (**Figure 4d**). Despite that, we
378 observed non-linear relationships between age and sex for FSHB in the test sets of non-white self-
379 reported ethnicities (**Supplementary Figure 4**). This might be due to the penalized linear model
380 being less affected by the higher genetic diversity found in African and Caribbean populations^{51,52}.

381 Finally, the non-linear effects for FAM3D, likely caused by epistatic interactions (**Figure 3d-e**)
 382 could not be fully replicated across the ethnicity test sets (**Figure 4e**). The DL model only
 383 outperformed the linear models slightly on ‘South Asian’, ‘East Asian’ and ‘Caribbean’ tests set
 384 with a much smaller extent. As above, this could be due to different LD patterns or variants not
 385 replicating across the ethnic test sets. In summary, these results suggest that cross ethnicity training
 386 is likely needed for the non-linear patterns of the DL model to transfer to individuals from diverse
 387 ethnicities that the model was not trained on.
 388

Figure 4



389 **Figure 4. Examples of deep learning and linear model performance across self-reported ethnicities.** **a)** Pearson
 390 correlation between the performance gap (mean bootstrapped R² DL-linear) of models between different self-reported
 391 ethnicities for 138 proteins with potential non-linear effects in the UK Biobank. The models were trained on
 392 individuals of self-reported ‘white’ ethnicity and tested on individuals of ‘White’, ‘South Asian’, ‘East Asian’,
 393 ‘Caribbean’ or ‘African’ self-reported ethnicities. If the R² was negative for both models, the performance gap was
 394 set to 0. **b)** Mean bootstrapped performance of DL and linear models for CD209. The models were trained on
 395 individuals of self-reported ‘white’ ethnicity and tested on individuals of the respective self-reported ethnicities. The
 396 error bars indicate the 95% confidence intervals. **c)** INT normalized CD209 levels stratified by ABO blood type among
 397 the different self-reported ethnicities. AO and AA blood type correspond to A blood group, and BO and BB blood
 398 type correspond to B blood group. **d)** Mean bootstrapped performance of DL and linear models for FSHB. The models
 399 were trained on individuals of self-reported ‘white’ ethnicity and tested on individuals of the respective self-reported
 400 ethnicities. The error bars indicate the 95% confidence intervals. **e)** Mean bootstrapped performance of DL and linear
 401

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

404

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

406 We replicated our findings of non-linear genetic effects in a cohort of 1,757 individuals from the
407 FinnGen project⁸. Olink protein levels were available for 170 of the 171 proteins with significant
408 performance gap in the UKB. We were able to replicate the dominance effect of the *ABO* blood
409 group tagging variants rs505922 and rs8176719 on plasma levels of CD209 and CLEC4M (**Figure**
410 **5a-b, Supplementary Figure 3a**). Furthermore, we investigated protein levels of FAM3D in
411 individuals with different genotype combinations of *ABO* variant rs507666 and *FUT3* variant
412 rs812936 and found higher levels in individuals with the GG-GG combination (**Figure 5c**). This
413 replicated the discovery of this rare interaction in the UKB (**Figure 3i**), despite the much lower
414 sample size and interaction allele counts (AC=43) in FinnGen. For ALPI, MUC2 and FAM3D we
415 replicated the epistatic effect of rs507666 and the *IZUMO1* variant rs2307019, resembling *ABO*
416 and *FUT2* secretor status interaction, on protein levels similar to the UKB (**Supplementary**
417 **Figure 5a**). Next, we sought to replicate the ability of our EIR-auto-GP workflow to identify non-
418 linear effects using the FinnGen cohort⁸. Using 1,231 and 263 individuals for training and test,
419 respectively, we could replicate the discovery of potential non-linear covariate effects for FSHB
420 and PAEP and potential non-linear genetic effects for MUC2, FAM3D and CD209
421 (**Supplementary Figure 5b**). We noticed that 94 of the 170 proteins had a higher DL performance
422 in FinnGen compared to the UKB (**Supplementary Figure 5c**). We speculate that this was due to
423 the different age distribution of the FinnGen cohort compared to the UKB (Median age FinnGen:
424 53 years, UKB: 58 years) (**Supplementary Figure 5d**). These results demonstrated that our DL
425 model can predict protein levels from genotype and covariate data across cohorts. In summary, we
426 were able to both directly replicate the dominance and interaction effects we discovered in UKB
427 and, despite the significantly lower sample size, replicate the EIR-auto-GP workflow by re-
428 discovery of non-linear effects of several proteins in FinnGen.

429

430

431

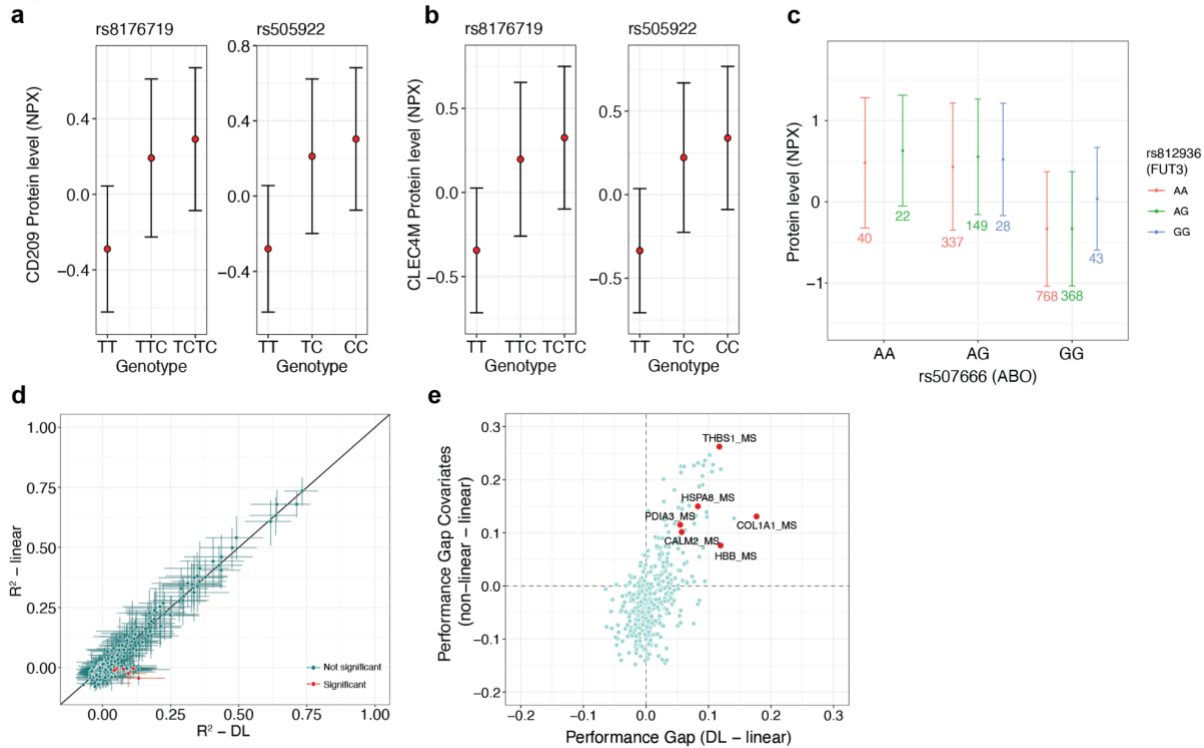
432

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

434 Finally, we aimed to replicate our analyses by training models using data generated by a different
435 proteomics technology. We therefore, as above, retrained models using a Danish cohort of obese
436 children measured using mass spectrometry based (MS) proteomics (The HOLBAEK Study)¹⁰.
437 The cohort consisted of 1,924 children and adolescents between the age of 5–20 years (**Methods**),
438 from which our group previously described the genetic regulation of its plasma proteome¹⁰. Similar
439 to the approach in the UKB and FinnGen, we trained models for 411 MS protein levels in 1,533
440 individuals and tested their performance in 190 individuals (**Supplementary Table 8**). Despite
441 the significantly lower sample size, we observed a similar trend in performance gaps, where the
442 DL model outperformed the linear model on the majority of proteins (246 of 411; 69%)
443 (**Supplementary Figure 5e**). Additionally, with a similar stringent cutoff as in the UKB, we
444 identified 6 (1.6%) proteins that showed a significant (non-overlapping confidence intervals)
445 increase when using the DL model (**Figure 5d**). When modeling protein levels only from
446 covariates using non-linear (XGBoost) and linear models, we found that the DL performance of
447 the significant proteins was likely driven by non-linear covariate effects (**Figure 5e**). Consistently,
448 we found that COL1A1 levels were influenced by non-linear effects between age and sex
449 (**Supplementary Figure 5f**). Taken together, these results demonstrate that our approach can be
450 applied to proteomics data acquired by different assays including both mass spectrometry-based
451 and affinity-based approaches.

452

Figure 5



453
 454 **Figure 5. Replication of non-linear effects across cohorts and platforms.** **a)** Protein levels (NPX) of CD209 in
 455 individuals with different genotypes of rs505922 and rs8176719 in the FinnGen project. Error bars indicate standard
 456 deviation. **b)** Protein levels (NPX) of CLEC4M in individuals with different genotypes of rs505922 and rs8176719 in
 457 the FinnGen project. Error bars indicate standard deviation. **c)** Protein levels (NPX) of FAM3D in individuals with
 458 different genotype combinations of rs507666 and rs812936. The numbers indicate the number of individuals with
 459 respective combinations. Error bars indicate standard deviation. Data points with < 5 individuals were removed. **d)**
 460 DL (EIR) and linear (bigstatsr) model performance (R^2) for all 411 proteins measured by MS in The HOLBAEK
 461 Study. The error bars indicate the 95% confidence interval from 1000 bootstraps, and proteins with non-overlapping
 462 confidence intervals between DL and linear models are called significant and labeled in red. **e)** Performance gap
 463 between DL and linear models on genotype and covariates against the performance gap between non-linear (XGBoost)
 464 and linear models on covariates only. Results for all 411 proteins are shown, and proteins with significant performance
 465 gaps between DL and linear model are labeled in red.

466
 467 **Discussion**

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

472 that many non-linear effects are present, illustrated by the increased performance of the deep
473 learning models compared to linear models. Of the 2,922 measured proteins, we identified 138
474 proteins potentially regulated by non-linear effects like non-linear covariate effects (n=123), or
475 genetic dominance and epistatic interactions between variants (n=15). Our modeling of non-linear
476 relationships between covariates that influence protein levels were in line with previous reports^{6,25}.
477 Many associations have previously been established between plasma proteomics and demographic
478 factors such as age, sex and BMI, and health indications such as liver function^{6,10}. We show that
479 complex relationships between non-genetic factors are widespread, and we speculate that, if we
480 included additional covariates in our analysis (i.e., BMI) we could uncover biologically relevant
481 non-linear relationships. This was outside the scope of the current study but will likely be the
482 subject of future research. We demonstrate that genetic dominance within loci that affect protein
483 levels is rare. This is consistent with previous studies that highlight the robustness of additive
484 models when modeling human traits¹⁸. However, we identified a small group of proteins that are
485 likely influenced by dominance effects in the *ABO* locus. Additionally, we could replicate epistatic
486 interactions between the *ABO* locus and the *FUT2* secretor status that regulate plasma levels of
487 intestinal proteins, as demonstrated before⁶. This shows that our approach can identify epistatic
488 interactions in an unbiased fashion. We also uncovered novel interactions between the *ABO* and
489 *FUT3* loci that influence plasma levels of the intestinally expressed protein FAM3D.

490
491 We uncovered complex effects that improve our understanding of biological pathways, which is a
492 major focus in the V2F challenge. For example, we identified dominance effects of variants in the
493 *ABO* locus on protein levels of CD209, CLEC4M and other proteins. The relationship between the
494 *ABO* locus, specifically the *ABO* blood group system, and plasma abundance of proteins involved
495 in the immune response, could advance our understanding of varying susceptibility to infectious
496 diseases among individuals with different blood types⁵³⁻⁵⁵. Specifically, CD209 and CLEC4M act
497 as attachment receptors for HIV-1 & 2, Ebola virus and other viral and bacterial pathogens^{56,57}.
498 Further, CD209 has been suggested to enhance ACE2-mediated SARS-CoV-2 infection⁵⁸.
499 Previous studies have also linked variants in the *ABO* locus with plasma levels of CD209^{6,59}. Here,
500 we demonstrate that individuals with A, B or AB blood-type have higher plasma levels of CD209
501 and CLEC4M. This suggests a mechanism where the *ABO* blood group system modulates
502 pathogen recognition of dendritic cells through CD209 and CLEC4M, which could be an important

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

506
507 To conduct our analyses, we developed the EIR-auto-GP software toolkit, designed to enable other
508 researchers to apply our DL approach to their studies. The toolkit consists of a fully automated
509 pipeline, allowing for the integration of genetic and covariate data for modeling on quantitative
510 and binary traits. We emphasize that the built-in variant pre-filtering approaches in EIR-auto-GP
511 allow for training on large-scale genetic data directly on CPUs—demonstrated by our ability to
512 train DL models across ten cross-validation (CV) runs in three hours for a single protein on a 16-
513 core computer on DNAnexus. This feature lowers the barrier for entry to research teams without
514 access to high-cost hardware accelerators typically associated with DL.

515
516 Given that some proteins showed a gap in performance between EIR-auto-GP and the linear
517 benchmark models, we wanted to examine what exact factors, e.g., complex effects in the covariate
518 data, dominance effects and interaction effects between SNVs were driving the performance gaps.
519 Therefore, we integrated the ability to fit both linear and non-linear models on different
520 transformations of the input data into EIR-auto-GP. This included the use of covariate only data,
521 additive and one-hot encoded genotype data, allowing us to specifically analyze the impact of these
522 factors. For a more detailed analysis into the exact genetic components that might be driving the
523 performance gaps, we used ordinary least squares (OLS) models for examining SNV genotypes
524 separately as well as SNV-SNV interaction effects. We hope that EIR-auto-GP will help advance
525 genetic research by providing an accessible DL toolkit to model complex genetic effects that
526 influence molecular and disease traits, thereby addressing important aspects of the V2F challenge.

527
528 We initially trained and tested our models on individuals of self-reported white ethnicity, as this
529 is the largest group of individuals with very similar ancestral backgrounds in the UK Biobank and
530 the UKB-PPP^{3,6}. When testing the models on sets of different self-reported ethnicity groups, we
531 observed reduced performance for linear and DL models, which was potentially due to differences
532 in population structure^{46,60}. This study serves as a proof-of-concept of our approach to capture non-

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

535
536 We replicated the effect of the identified epistatic *ABO-FUT3* interaction on plasma levels of
537 FAM3D, MUC2 and ALPI, as well as the dominance effect of the *ABO* locus on plasma levels of
538 CD209 and CLEC4M in the FinnGen cohort. This demonstrates that our findings are transferable
539 beyond the UK-white population of the UK Biobank to other populations. However, given the 20-
540 fold smaller sample size, the DL models might not effectively detect these effects (e.g., due to
541 overfitting) adequately for it to be reflected in significantly better test set performance.
542 Additionally, the increased uncertainty, as indicated by larger confidence intervals, when
543 evaluating the models on a much smaller test set makes it challenging to identify significant results,
544 despite better performance metrics. However, the increasing availability of larger proteomics
545 cohorts will enable the identification of non-linear genetic and covariate effects on protein
546 abundance in an unbiased, large-scale manner.

547

548 **Limitations**

549 A large fraction of the blood plasma proteins could not be accurately modeled from genotypes and
550 the chosen covariates using either the linear or DL model. This may be due to missing causal
551 SNVs, or unaccounted environmental factors, such as BMI and health and disease conditions,
552 which are known to affect protein levels⁶. Interestingly, we found that many proteins that we could
553 not model had a low correlation with SomaScan measurements¹⁴. The comparability between
554 Olink and other protein quantification methods is highly debated in the field^{14,61}. Our findings
555 might indicate that plasma levels for some proteins measured by Olink may not be entirely
556 accurate, which can affect the performance of our models in the UKB. Advances in MS-based
557 proteomics could allow for higher specificity and quantitative accuracy of plasma proteomics in
558 large sample sizes comparable to those of the UKB in the future^{10,62}.

559
560 Regarding modeling, we used a threshold on the pre-GWAS analysis to limit the number of
561 variants used as input to the DL model. The choice is not guaranteed to be optimal for modeling
562 purposes of all proteins, and this step might filter variants with purely complex effects not detected
563 in GWAS. Furthermore, UKB array data were used to conduct the analysis, which might affect the

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

573
574 While our approach of using differences in performance gaps can identify protein levels modulated
575 by interaction effects, it likely does not identify rare interactions. For example, using the OLS
576 models we identified an interaction between variants in the *ABO* and *FUT3* loci which has a low
577 frequency in the present cohort (~2%). These rare effects are unlikely to be learned during model
578 training, and even if captured, may not significantly impact test set performance. This indicates
579 that even the ~50,000 samples of the UKB-PPP might be too small to discover rare variant
580 interactions using our approach.

581
582 **Conclusions and future directions**

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

593
594

595 **Methods**

596

597 **Experimental setup and processing of UK Biobank data**

598 In the genomic quality control (QC) process, we utilized PLINK v1.90b7⁶³ for data analysis and
599 filtering. Our dataset initially consisted of 784,256 autosomal variants and 488,377 individuals.
600 We removed individuals with a relatedness factor ≥ 0.0884 (second degree relatedness), resulting
601 in 453,581 individuals kept for the analysis. We applied the following QC filters: individuals with
602 more than 10% missing genotype data (--mind 0.10) were excluded, resulting in the removal of 6
603 individuals, with 453,575 remaining. Variant level QC involved removing variants with more than
604 1% missing data (--geno 0.01), leading to the exclusion of 143,713 variants. Additionally, variants
605 failing the Hardy-Weinberg Equilibrium test at a threshold of 0.000001 (--hwe 0.000001) were
606 removed, accounting for 153,825 variants. We also applied a minor allele frequency (MAF)
607 threshold of 0.005 (--maf), resulting in the exclusion of 62,621 variants. After the application of
608 these QC steps, the final dataset comprised 424,097 variants and 453,575 individuals. We divided
609 the individuals into train, validation, and test sets for the modeling. In the training dataset, we
610 included exclusively individuals with self-reported ‘UK-white’ ethnicity from Olink batches 0-6.
611 Batch 7 contains consortium selected individuals and individuals from the COVID-19 imaging
612 study and do not follow UKB baseline characteristics⁶. To this end, individuals from batch 7 and
613 all individuals with non-UK-white ethnicity were excluded from the training dataset. Additionally,
614 individuals from batch 7 were excluded in the UK-white test set (n=1,771) used throughout the
615 study. Access to the UK Biobank data was obtained through application 1251 “The metabolically
616 healthy obese and metabolically obese normal-weight in the UK Biobank: Prevalence, genes and
617 lifestyle contributors, disease risk and mortality”.

618

619 **Deep learning model training**

620 The main deep learning models on the UKB were trained with the EIR-auto-GP toolkit
621 (<https://github.com/arnor-sigurdsson/EIR-auto-GP>, commit fb41457). The ordinary least squares
622 (OLS) estimation for allele effects and interaction effects was also done with the toolkit, as well
623 as the direct estimation of protein levels as a function of genotype combinations (commit
624 0d5d762). Besides the genotype input data, categorical covariate inputs were sex, UKB center,
625 UKB genetic array and whether individuals were consortium selected. Continuous inputs were age

626 and genetic principal components 1-20 (UKB data field 22009). Each protein level run consisted
627 of 10 holdout cross validation (CV) runs, using a pre-defined validation set for consistency across
628 runs. Despite repeated use of the same training-validation split, the models differed in each run
629 due to several factors: (a) random initialization of the models; (b) the order of data during mini-
630 batching during training was shuffled independently each run; and (c) while the first 3 CV runs
631 shared a common set of SNVs, the subsequent 7 runs used a different set of SNVs as determined
632 by a Bayesian optimization process (see below). For each protein level run, a GWAS pre-selection
633 with a p-value of 0.001 was applied to the training set and used to reduce the number of variants
634 input to the DL models. The first 3 CV runs used the full set of variants that passed the GWAS
635 pre-selection step, and DL attributions were computed with integrated gradients⁶⁴ on the validation
636 set. After the first 3 CV runs, a Bayesian optimization (BO) loop was applied to optimize for the
637 top variants to include in the following 7 CV runs. The BO was implemented with scikit-optimize
638 (v.0.9.0)⁶⁵, with the objective of optimizing the fraction of top SNVs regarding validation set
639 performance. The top variants were defined by averaging the absolute DL attributions computed
640 on the validation set across the first 3 CV runs. After training the DL models for all 10 CV runs, a
641 final ensemble prediction was applied to the test set.

642

643 **Linear model benchmarking**

644 The training of linear benchmark models was done with bigsnpr (1.12.2)³³ and bigstatsr (1.5.12)³³.
645 All 424,097 variants were used as input for the model, as well as the covariates sex, age, UKB
646 center, UKB genetic array, whether individuals were consortium selected and genomic principal
647 components 1-20. The modeling was conducted with a 10-fold cross-validation (CV) employing a
648 grid search for the α mixing parameter in the elastic net, exploring values [0.0001, 0.001, 0.01,
649 0.1, 1]. Additionally, the approach involves testing multiple values for the λ penalization parameter
650 (default 200). Following this, an ensemble-like process across the CV runs was executed to
651 generate the final model, which was subsequently assessed using the test set.

652

653 **Model performance**

654 The test set predictions of the trained DL and linear models were bootstrapped (n=1,000) and R²
655 and RMSE calculated for each bootstrap generation using sklearn.metrics.r2_score and
656 sklearn.metrics.mean_squared_error. From the resulting distribution, the 95% confidence intervals

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

660

661 **Self reported ethnic grouping**

662 Individuals were stratified according to self-reported ethnic background in the UKB (data field
663 21000). The individual groups were consolidated into 5 main groups for the purpose of the study.
664 The groups were defined as “White”, “South Asian”, “East Asian”, “African”, and “Caribbean”.
665 The “White” group included British, Irish, and other individuals with white backgrounds. The
666 “South Asian” group comprised individuals of Indian, Pakistani, Bangladeshi. Those of Chinese
667 self-reported ethnicities were categorized as “East Asian.” “African” and “Caribbean” groups were
668 kept as indicated in data field 21000.

669

670 **Model Complexity Analysis**

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

684

685 **Genetic non-additivity analysis**

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

688 where the models were fit on each genotype separately ($\text{Target} = \beta_0 + \beta_1 \times \text{SNV}_1 + \beta_2 \times \text{SNV}_2 + \beta_3$
689 $\times \text{SNV}_3 + \text{covariates} + \epsilon$). By examining the p-values and effect sizes coefficients assigned to each
690 genotype, we could explore for each SNV whether it deviated from an additive relationship
691 towards a protein level.

692

693 **Genetic interaction analysis**

694 In addition to investigating performance disparities between the linear and non-linear XGBoost
695 models, we explored potential pairwise interactions among the top 128 SNVs. This entailed fitting
696 Ordinary Least Squares (OLS) models on all possible pairwise combinations of SNVs, where each
697 model utilized each SNV (one-hot encoded) as inputs along with the product interaction term
698 between them ($\text{Target} = \beta_0 + \beta_{11} \times \text{SNV}_{11} + \beta_{12} \times \text{SNV}_{12} + \dots + \beta_{21} \times \text{SNV}_{21} + \beta_{22} \times \text{SNV}_{22} + \dots$
699 $+ \beta_3 \times (\text{SNV}_1 \times \text{SNV}_2) + \text{covariates} + \epsilon$). Across all traits, we tested a total of 1,389,888 pairs, and
700 as such applied a p-value threshold of $0.05 / 1,389,888 = 3.6\text{e-}08$. This approach allowed us to
701 identify which SNV pairs might contribute most to any remaining performance gap between the
702 linear model using one-hot encoded genotype data with tabular data and the XGBoost model
703 trained on the same data.

704

705 **Replication in MS-based proteomics data from The HOLBAEK Study**

706 The HOLBAEK Study consisted of 2,147 children and adolescents (55% girls) between the age of
707 5 and 20, recruited from the Children's Obesity Clinic, accredited Centre for Obesity Management,
708 Copenhagen University Hospital Holbæk, Denmark⁶⁶, and a population-based cohort recruited from
709 schools in 11 municipalities across Zealand, Denmark⁶⁷. Besides age, an eligibility criteria of the
710 Obesity Clinic was BMI above the 90th percentile (BMI SDS ≥ 1.28) according to Danish
711 reference values. The study protocol for The HOLBAEK Study was approved by the ethics
712 committee for the Region Zealand (protocol no. SJ-104) and is registered at the Danish Data
713 Protection Agency (REG-043-2013). The HOLBAEK Study including the obesity clinic cohort
714 and the population-based cohort are also registered at ClinicalTrials.gov (NCT00928473). The MS
715 based proteomics data consisted of 411 protein levels measured across 2,130 of the 2,147 samples,
716 with genotype data available for 1,924 individuals featuring 5,242,958 variants after quality
717 control and filtering¹⁰. Due to the imbalance in the number of features compared to the number of
718 samples, we used PLINK to perform LD pruning (`--indep-pairwise 50 5 0.8`), reducing the variant

719 count to 998,505. We then matched and retained only those samples for which both genotype and
720 phenotype data were available, identifying 1,893 samples with complete data sets. Following the
721 EIR-auto-GP data processing pipeline, data splits were defined as 1,533 training, 170 validation,
722 and 190 test samples. Besides the genotype data, covariates included were sex, BMI, age, time to
723 analysis, MS batch information. Due to the smaller sample size compared to the UKB, we used
724 20-fold cross validation (CV) instead of the 10-fold applied in the UKB. EIR-auto-GP (commit
725 2934974) was used for DL model training and additionally, we found that the default feature
726 selection approach in EIR-auto-GP (i.e., a fixed GWAS threshold and DL attribution based
727 Bayesian optimization (BO) of included SNVs) was susceptible to overfitting in this dataset, based
728 on training and validation set performance. To address this, we devised an alternative, simpler
729 approach focusing on dynamic SNV inclusion based on GWAS p-value rankings. The optimization
730 process began with seeding the algorithm with manual fractions, reflecting SNV subsets from the
731 most significant (p-value threshold of $1e-8$) to the least (up to a p-value of $1e-4$). After this, the
732 BO process to find the optimal fraction of SNVs was allowed to proceed. We found that this
733 approach guided more efficiently towards using fewer SNVs, which resulted in better validation
734 performance.

735

736 **Replication in FinnGen Olink data**

737 The FinnGen quality controlled Olink data consisted of 2,925 measured protein levels across 1,990
738 samples, with genotype data available for 520,210 individuals and 21,331,644 variants initially.
739 The variants were filtered to match those used in the UKB experiments, resulting in a final set of
740 416,802 variants. Retaining only samples where genotype and phenotype data were available, our
741 final set consisted of 1,757 samples. Data splits were defined as 1,231 training, 263 validation and
742 263 test samples. Besides the genotype data, covariates included blood sampling age, sex, genetic
743 testing chip and batch, top 20 genetic PCs and protein examination batch. The DL model training
744 was performed with EIR-auto-GP (commit c141b5a) and the training procedure was the same as
745 described above for the data set from The HOLBAEK Study.

746

747 **FinnGen Ethics statement**

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

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

756 The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers:
757 THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018,
758 THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and
759 THL/1524/5.05.00/2020), Digital and population data service agency (permit numbers:
760 VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution
761 (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA
762 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA
763 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020,
764 THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020,
765 THL/5894/14.06.00/2020, THL/6619/14.06.00/2020, THL/209/14.06.00/2021,
766 THL/688/14.06.00/2021, THL/1284/14.06.00/2021, THL/1965/14.06.00/2021,
767 THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021, Statistics Finland
768 (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20)
769 TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases
770 permission/extract from the meeting minutes on 4th July 2019.

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

781 2021_5015, 2021_5015 Amendment, 2021_5015 Amendment_2, 2021_5023, 2021_5023
782 Amendment, 2021_5023 Amendment_2, 2021_5017, 2021_5017 Amendment, 2022_6001,
783 2022_6001 Amendment, 2022_6006 Amendment, 2022_6006 Amendment, 2022_6006
784 Amendment_2, BB22-0067, 2022_0262, 2022_0262 Amendment, Biobank of Eastern Finland
785 1186/2018 and amendment 22§/2020, 53§/2021, 13§/2022, 14§/2022, 15§/2022, 27§/2022,
786 28§/2022, 29§/2022, 33§/2022, 35§/2022, 36§/2022, 37§/2022, 39§/2022, 7§/2023, 32§/2023,
787 33§/2023, 34§/2023, 35§/2023, 36§/2023, 37§/2023, 38§/2023, 39§/2023, 40§/2023, 41§/2023,
788 Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020),
789 BB2021-0140 8§/2021, 9§/2021, §9/2022, §10/2022, §12/2022, 13§/2022, §20/2022, §21/2022,
790 §22/2022, §23/2022, 28§/2022, 29§/2022, 30§/2022, 31§/2022, 32§/2022, 38§/2022, 40§/2022,
791 42§/2022, 1§/2023, Central Finland Biobank 1-2017, BB_2021-0161, BB_2021-0169, BB_2021-
792 0179, BB_2021-0170, BB_2022-0256, BB_2022-0262, BB22-0067, Decision allowing to
793 continue data processing until 31st Aug 2024 for projects: BB_2021-0179, BB22-0067, BB_2022-
794 0262, BB_2021-0170, BB_2021-0164, BB_2021-0161, and BB_2021-0169, and Terveystalo
795 Biobank STB 2018001 and amendment 25th Aug 2020, Finnish Hematological Registry and
796 Clinical Biobank decision 18th June 2021, Arctic biobank P0844: ARC_2021_1001.

797

798 **Acknowledgments**

799 S.R. and M.M. were supported by the Novo Nordisk Foundation (NNF14CC0001). S.R., J.F.G.,
800 R.L. and T.H. were supported by the Novo Nordisk Foundation (NNF23SA0084103). S.R., R.L.
801 and B.M.N. were supported by the Novo Nordisk Foundation (NNF21SA0072102). J.F.G. was
802 supported by a research grant from the Danish Cardiovascular Academy, which is funded by the
803 Novo Nordisk Foundation, grant number NNF20SA0067242 and The Danish Heart Foundation.
804 This research has been conducted using the UK Biobank Resource under Application Number
805 1251. Our gratitude goes to all participants and their families from the UK Biobank and The
806 HOLBAEK Study. We want to acknowledge the participants and investigators of the FinnGen
807 study. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016
808 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd,
809 Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl),
810 Genentech Inc., Merck Sharp & Dohme LCC, Pfizer Inc., GlaxoSmithKline Intellectual Property
811 Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis

812 AG, and Boehringer Ingelheim International GmbH. Following biobanks are acknowledged for
813 delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank
814 (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of
815 Northern Finland ([https://www.ppsHP.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-](https://www.ppsHP.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx)
816 [briefly-in-English.aspx](https://www.ppsHP.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx)), Finnish Clinical Biobank Tampere ([www.tays.fi/en-](http://www.tays.fi/en-US/Research%20and%20development/Finnish%20Clinical%20Biobank%20Tampere)
817 [US/Research and development/Finnish Clinical Biobank Tampere](http://www.tays.fi/en-US/Research%20and%20development/Finnish%20Clinical%20Biobank%20Tampere)), Biobank of Eastern
818 Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank ([www.ksshP.fi/fi-](http://www.ksshP.fi/fi-FI/Potilaille/Biopankki)
819 [FI/Potilaille/Biopankki](http://www.ksshP.fi/fi-FI/Potilaille/Biopankki)), Finnish Red Cross Blood Service Biobank
820 (www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank
821 (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/) and Arctic Biobank
822 ([https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-](https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-cohorts-and-arctic-biobank)
823 [cohorts-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
824 (<https://www.bbmri-eric.eu/national-nodes/finland/>). Finnish Biobank Cooperative -FINBB
825 (<https://finbb.fi/>) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank
826 data can be accessed through the Fingenious[®] services (<https://site.fingenious.fi/en/>) managed by
827 FINBB.

828

829 **Author contributions**

830 Conceptualization: S.R. Formal Analysis: A.I.S., J.F.G., Z.Y. Investigation: J.F.G., A.I.S.
831 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.
832 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.,
833 L.N., M.M. Visualization: J.F.G., A.I.S. Writing – original draft: A.I.S., J.F.G, S.R. Writing –
834 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.

835

836 **Competing interest**

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

838

839 **Code availability**

840 EIR is available at <https://github.com/arnor-sigurdsson/EIR>. EIR-auto-GP is available at
841 <https://github.com/arnor-sigurdsson/EIR-auto-GP>.

842

843 **Data availability**

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

857

858 **References**

- 859 1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189
860 (2020).
- 861 2. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics.
862 *Science* **373**, 1464–1468 (2021).
- 863 3. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
864 *Nature* **562**, 203–209 (2018).
- 865 4. Shilo, S. *et al.* 10 K: a large-scale prospective longitudinal study in Israel. *Eur. J.*
866 *Epidemiol.* **36**, 1187–1194 (2021).
- 867 5. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and
868 disease. *Nat. Genet.* **53**, 1712–1721 (2021).
- 869 6. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank.
870 *Nature* **622**, 329–338 (2023).
- 871 7. Julkunen, H. *et al.* Atlas of plasma NMR biomarkers for health and disease in 118,461
872 individuals from the UK Biobank. *Nat. Commun.* **14**, 604 (2023).
- 873 8. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated
874 population. *Nature* **613**, 508–518 (2023).
- 875 9. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in
876 30,931 individuals. *Nature Metabolism* **2**, 1135–1148 (2020).
- 877 10. Niu, L. *et al.* Plasma proteome variation and its genetic determinants in children and
878 adolescents. *bioRxiv* (2023) doi:10.1101/2023.03.31.23287853.
- 879 11. Niu, L. *et al.* Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat. Med.*
880 **28**, 1277–1287 (2022).

- 881 12. Gudjonsson, A. *et al.* A genome-wide association study of serum proteins reveals shared
882 loci with common diseases. *Nat. Commun.* **13**, 480 (2022).
- 883 13. Dhindsa, R. S. *et al.* Rare variant associations with plasma protein levels in the UK
884 Biobank. *Nature* **622**, 339–347 (2023).
- 885 14. Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and
886 disease associations. *Nature* **622**, 348–358 (2023).
- 887 15. Lourdasamy, A. *et al.* Identification of cis-regulatory variation influencing protein
888 abundance levels in human plasma. *Hum. Mol. Genet.* **21**, 3719–3726 (2012).
- 889 16. Genetic control of the human brain proteome. *Am. J. Hum. Genet.* **108**, 400–410 (2021).
- 890 17. Ruffieux, H. *et al.* A fully joint Bayesian quantitative trait locus mapping of human protein
891 abundance in plasma. *PLoS Comput. Biol.* **16**, e1007882 (2020).
- 892 18. Palmer, D. S. *et al.* Analysis of genetic dominance in the UK Biobank. *Science* **379**, 1341–
893 1348 (2023).
- 894 19. Epistasis in sporadic Alzheimer’s disease. *Neurobiol. Aging* **30**, 1333–1349 (2009).
- 895 20. Robson, K. J. H. *et al.* Synergy between the C2 allele of transferrin and the C282Y allele of
896 the haemochromatosis gene (HFE) as risk factors for developing Alzheimer’s disease. *J.*
897 *Med. Genet.* **41**, 261–265 (2004).
- 898 21. Williams, S. M. *et al.* Combinations of Variations in Multiple Genes Are Associated With
899 Hypertension. *Hypertension* (2000) doi:10.1161/01.HYP.36.1.2.
- 900 22. Elgart, M. *et al.* Non-linear machine learning models incorporating SNPs and PRS improve
901 polygenic prediction in diverse human populations. *Commun Biol* **5**, 856 (2022).
- 902 23. Albiñana, C. *et al.* Multi-PGS enhances polygenic prediction by combining 937 polygenic
903 scores. *Nat. Commun.* **14**, 1–11 (2023).

- 904 24. Machine learning optimized polygenic scores for blood cell traits identify sex-specific
905 trajectories and genetic correlations with disease. *Cell Genomics* **2**, 100086 (2022).
- 906 25. McCaw, Z. R. *et al.* DeepNull models non-linear covariate effects to improve phenotypic
907 prediction and association power. *Nat. Commun.* **13**, 241 (2022).
- 908 26. Badré, A., Zhang, L., Muchero, W., Reynolds, J. C. & Pan, C. Deep neural network
909 improves the estimation of polygenic risk scores for breast cancer. *J. Hum. Genet.* **66**, 359–
910 369 (2021).
- 911 27. Sigurdsson, A. I. *et al.* Deep integrative models for large-scale human genomics. *Nucleic*
912 *Acids Res.* (2023) doi:10.1093/nar/gkad373.
- 913 28. Wu, Z. *et al.* Two-stage joint selection method to identify candidate markers from genome-
914 wide association studies. *BMC Proc.* **3 Suppl 7**, S29 (2009).
- 915 29. Piriyaongsa, J. *et al.* iLOCi: a SNP interaction prioritization technique for detecting
916 epistasis in genome-wide association studies. *BMC Genomics* **13 Suppl 7**, S2 (2012).
- 917 30. Motzo, C. *et al.* Heterogeneity in the magnitude of the insulin gene effect on HLA risk in
918 type 1 diabetes. *Diabetes* **53**, 3286–3291 (2004).
- 919 31. Sigurdsson, A. I. *et al.* Improved prediction of blood biomarkers using deep learning.
920 *medRxiv* (2022).
- 921 32. Snaebjarnarson, A. S. *et al.* Complex effects of sequence variants on lipid levels and
922 coronary artery disease. *Cell* **186**, 4085–4099.e15 (2023).
- 923 33. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale
924 genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–
925 2787 (2018).
- 926 34. Sharma, N. *XGBoost. The Extreme Gradient Boosting for Mining Applications.* (GRIN

- 927 Verlag, 2018).
- 928 35. Tabular data: Deep learning is not all you need. *Inf. Fusion* **81**, 84–90 (2022).
- 929 36. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform
930 deep learning on tabular data? (2022).
- 931 37. Geijtenbeek, T. B. H. & Gringhuis, S. I. Signalling through C-type lectin receptors: shaping
932 immune responses. *Nat. Rev. Immunol.* **9**, 465–479 (2009).
- 933 38. Guo, Y. *et al.* Structural basis for distinct ligand-binding and targeting properties of the
934 receptors DC-SIGN and DC-SIGNR. *Nat. Struct. Mol. Biol.* **11**, 591–598 (2004).
- 935 39. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci
936 (pQTLs). *PLoS Genet.* **4**, e1000072 (2008).
- 937 40. Paré, G. *et al.* Novel association of ABO histo-blood group antigen with soluble ICAM-1:
938 results of a genome-wide association study of 6,578 women. *PLoS Genet.* **4**, e1000118
939 (2008).
- 940 41. Wolpin, B. M. *et al.* Pancreatic cancer risk and ABO blood group alleles: results from the
941 pancreatic cancer cohort consortium. *Cancer Res.* **70**, 1015–1023 (2010).
- 942 42. Groot, H. E. *et al.* Genetically Determined ABO Blood Group and its Associations With
943 Health and Disease. *Arterioscler. Thromb. Vasc. Biol.* **40**, 830–838 (2020).
- 944 43. Marcus, D. M. The ABO and Lewis blood-group system. Immunochemistry, genetics and
945 relation to human disease. *N. Engl. J. Med.* **280**, 994–1006 (1969).
- 946 44. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
- 947 45. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health
948 disparities. *Nat. Genet.* **51**, 584–591 (2019).
- 949 46. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human

- 950 populations. *Nat. Commun.* **10**, 3328 (2019).
- 951 47. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores
952 in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
- 953 48. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and
954 applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 373 (2022).
- 955 49. Wang, Y., Tsuo, K., Kanai, M., Neale, B. M. & Martin, A. R. Challenges and Opportunities
956 for Developing More Generalizable Polygenic Risk Scores. *Annu Rev Biomed Data Sci* **5**,
957 293–320 (2022).
- 958 50. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across
959 Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
- 960 51. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans.
961 *Science* **324**, 1035–1044 (2009).
- 962 52. Moreno-Estrada, A. *et al.* Reconstructing the Population Genetic History of the Caribbean.
963 *PLoS Genet.* **9**, e1003925 (2013).
- 964 53. Cserti, C. M. & Dzik, W. H. The ABO blood group system and *Plasmodium falciparum*
965 malaria. *Blood* **110**, 2250–2258 (2007).
- 966 54. Wu, S.-C. *et al.* Blood group A enhances SARS-CoV-2 infection. *Blood* **142**, 742–747
967 (2023).
- 968 55. Nordgren, J. & Svensson, L. Genetic Susceptibility to Human Norovirus Infection: An
969 Update. *Viruses* **11**, (2019).
- 970 56. Geijtenbeek, T. B. H. *et al.* Identification of different binding sites in the dendritic cell-
971 specific receptor DC-SIGN for intercellular adhesion molecule 3 and HIV-1. *J. Biol. Chem.*
972 **277**, 11314–11320 (2002).

- 973 57. Lin, G. *et al.* Differential N-linked glycosylation of human immunodeficiency virus and
974 Ebola virus envelope glycoproteins modulates interactions with DC-SIGN and DC-SIGNR.
975 *J. Virol.* **77**, 1337–1346 (2003).
- 976 58. Lempp, F. A. *et al.* Lectins enhance SARS-CoV-2 infection and influence neutralizing
977 antibodies. *Nature* **598**, 342–347 (2021).
- 978 59. Anisul, M. *et al.* A proteome-wide genetic investigation identifies several SARS-CoV-2-
979 exploited host targets of clinical relevance. *Elife* **10**, (2021).
- 980 60. Gomez, F., Hirbo, J. & Tishkoff, S. A. Genetic variation and adaptation in Africa:
981 implications for human evolution and disease. *Cold Spring Harb. Perspect. Biol.* **6**,
982 a008524 (2014).
- 983 61. Katz, D. H. *et al.* Proteomic profiling platforms head to head: Leveraging genetics and
984 clinical traits to compare aptamer- and antibody-based methods. *Sci Adv* **8**, eabm5164
985 (2022).
- 986 62. Bader, J. M., Albrecht, V. & Mann, M. MS-Based Proteomics of Body Fluids: The End of
987 the Beginning. *Mol. Cell. Proteomics* **22**, 100577 (2023).
- 988 63. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
989 datasets. *Gigascience* **4**, 7 (2015).
- 990 64. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. (2017).
- 991 65. scikit-optimize/scikit-optimize. doi:10.5281/zenodo.5565057.
- 992 66. Holm, J.-C. *et al.* Chronic care treatment of obese children and adolescents. *Int. J. Pediatr.*
993 *Obes.* **6**, 188–196 (2011).
- 994 67. Vissing Landgrebe, A. *et al.* Population-based pediatric reference values for serum
995 parathyroid hormone, vitamin D, calcium, and phosphate in Danish/North-European white

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

997