

Title: Integrative proteomic analyses across common cardiac diseases yield new mechanistic insights and enhanced prediction

Authors:

Art Schuermans^{1,2,3,*}, Ashley B. Pournamdari^{1,4,*}, Jiwoo Lee^{1,2}, Rohan Bhukar^{1,2}, Shriienidhie Ganesh^{1,2}, Nicholas Darosa^{1,2}, Aeron M. Small^{1,5}, Zhi Yu^{1,2}, Whitney Hornsby^{1,2}, Satoshi Koyama^{1,2}, James L. Januzzi, Jr.^{6,7,8}, Michael C. Honigberg^{1,2,7,8,§}, Pradeep Natarajan^{1,2,7,8,§}.

¹ Program in Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Cambridge, MA, USA

² Cardiovascular Research Center and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

³ Faculty of Medicine, KU Leuven, Leuven, Belgium

⁴ Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

⁵ Cardiovascular Medicine Division, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

⁶ Baim Institute for Clinical Research, Boston, MA, USA

⁷ Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA

⁸ Department of Medicine, Harvard Medical School, Boston, MA, USA

*These authors share first authorship.

§These authors co-supervised this work.

Correspondence to:

Michael C. Honigberg, MD, MPP

Massachusetts General Hospital

185 Cambridge St. CPZN 3.187

Pradeep Natarajan, MD, MMSc

Massachusetts General Hospital

185 Cambridge St., CPZN 3.184

Boston, MA 02114

Phone: 617-726-1843

Fax: 617-726-2203

E-mail: mhonigberg@mg.harvard.edu

Twitter: @mchonig

Boston, MA 02114

Phone: 617-726-1843

Fax: 617-726-2203

E-mail: pnatarajan@mg.harvard.edu

Twitter: @pnatarajanmd

Word count: 4,000 / 4,000

Abstract:

Cardiac diseases represent common highly morbid conditions for which underlying molecular mechanisms remain incompletely understood. Here, we leveraged 1,459 protein measurements in 44,313 UK Biobank participants to characterize the circulating proteome associated with incident coronary artery disease, heart failure, atrial fibrillation, and aortic stenosis. Multivariable-adjusted Cox regression identified 820 protein-disease associations—including 441 proteins—at Bonferroni-adjusted $P < 8.6 \times 10^{-6}$. *Cis*-Mendelian randomization suggested causal roles that aligned with epidemiological findings for 6% of proteins identified in primary analyses, prioritizing novel therapeutic targets for different cardiac diseases (e.g., interleukin-4 receptor for heart failure and spondin-1 for atrial fibrillation). Interaction analyses identified seven protein-disease associations that differed Bonferroni-significantly by sex. Models incorporating proteomic data (vs. clinical risk factors alone) improved prediction for coronary artery disease, heart failure, and atrial fibrillation. These results lay a foundation for future investigations to uncover novel disease mechanisms and assess the clinical utility of protein-based prevention strategies for cardiac diseases.

Abstract word count: 150 / 150

1 **Main**

2 Cardiac diseases represent the leading global cause of morbidity and mortality¹, with
3 coronary artery disease, heart failure, atrial fibrillation, and aortic stenosis collectively accounting
4 for more than 90% of cardiac deaths^{1,2}. The prevention of these diseases often relies on risk
5 assessment and early pharmacotherapy, yet most high-risk individuals remain undetected until
6 they experience their first clinical event^{3,4}. In addition, even under optimal treatment conditions,
7 there remains significant residual risk that is incompletely captured by traditional risk factors^{5,6}.
8 The development of effective screening tools and discovery of novel therapeutic targets could
9 significantly enhance early detection and improve treatment outcomes for different cardiac
10 diseases.

11 Large-scale genetic analyses have identified thousands of sequence variants associated
12 with cardiac diseases, enabling insights into disease etiology and facilitating the development of
13 risk prediction tools⁷⁻¹¹. However, for many genetic associations, molecular effectors remain
14 incompletely understood. For instance, recent genome-wide association studies of coronary
15 artery disease suggest that only ~50% of identified genetic loci confer risk through known risk
16 factors such as lipids, obesity, and blood pressure^{7,8}. Molecular pathways through which the
17 remaining loci (with signals typically residing in non-coding DNA sequence) exert their effects
18 remain less clear. However, they ultimately may represent ideal potential targets to reduce
19 residual cardiovascular disease risk.

20 The circulating proteome is a dynamic network that encompasses all proteins in the
21 bloodstream, reflecting both genetic background and external factors such as environmental
22 exposures and lifestyle alterations. Smaller studies have demonstrated that sparse protein-
23 based risk scores can improve the prediction of recurrent events in individuals with established
24 atherosclerotic cardiovascular disease^{12,13}. Furthermore, targeted analyses of specific
25 biomarkers suggest that integrating proteomic and genetic data can nominate causal protein-
26 disease associations and reveal actionable drug targets in the bloodstream^{14,15}. Whether

27 population-scale, agnostic analyses of the circulating proteome can provide novel insights into
28 disease biology and improve the prediction of first clinical events across a range of cardiac
29 disease subtypes remains unclear.

30 In the present study, we perform a proteomic analysis of cardiac diseases in the UK
31 Biobank Pharma Proteomics Project (UKB-PPP)¹⁶, a population-based cohort including 54,306
32 individuals with longitudinal follow-up (**Fig. 1**). Using multivariable-adjusted time-to-event
33 models, we test the associations of 1,459 circulating proteins with incident coronary artery
34 disease, heart failure, atrial fibrillation, and aortic stenosis. We leverage Mendelian
35 randomization (MR) to infer causal roles among the identified proteins to prioritize novel
36 therapeutic targets. Finally, we present sex-specific protein-disease associations and evaluate
37 whether protein-based scores can improve disease prediction beyond the use of traditional risk
38 factors.

39

40 **Results**

41 *Study population*

42 The present study included 44,313 unrelated UKB-PPP participants without a history of
43 coronary artery disease, heart failure, atrial fibrillation, or aortic stenosis at enrolment. Participant
44 characteristics are summarized in **Table 1**. Baseline blood samples underwent proteomic
45 profiling using the *Olink Explore 1536* platform, providing data on 1,459 unique proteins
46 (**Supplementary Table 1**) available for >90% of UKB-PPP participants (**Supplementary Table**
47 **2**).

48 The majority of participants were female ($n=24,701$ [55.7%]) and self-reported as white
49 ($n=41,481$ [93.6%]), with a mean (standard deviation [SD]) age of 56.4 (8.2) years. A total of
50 4,610 participants (10.4%) experienced at least one cardiac event over a median (interquartile
51 range, IQR) follow-up of 11.1 (10.4-11.8) years. Coronary artery disease had the highest
52 cumulative incidence (6.2% [$n=2,729/44,313$]), followed by atrial fibrillation (4.8%

53 [$n=2,107/44,313$]), heart failure (2.3% [$n=1,014/44,313$]), and aortic stenosis (0.7%
54 [$n=326/44,313$]) (**Supplementary Fig. 1**).

55

56 *Associations of circulating proteins with incident heart diseases*

57 Primary analyses tested the associations of 1,459 circulating proteins with each of the
58 incident heart diseases (coronary artery disease, heart failure, atrial fibrillation, and aortic
59 stenosis) using multivariable-adjusted Cox regression models. These analyses identified 820
60 protein-disease associations—reflecting 441 unique proteins—at Bonferroni-corrected
61 $P<0.05/5,836$ (**Fig. 2, Supplementary Table 3**). Heart failure had the highest number of
62 proteomic associations ($n=384$), followed by coronary artery disease ($n=259$), atrial fibrillation
63 ($n=156$), and aortic stenosis ($n=21$). Among proteins with one or more significant associations,
64 261 (59.2%) were shared across multiple outcomes and 15 (3.4%) were shared across all four
65 outcomes (**Supplementary Fig. 2**).

66 The strongest protein-disease associations (by P -value) were observed for atrial
67 fibrillation, with N-terminal pro-B type natriuretic peptide (NT-proBNP) and B-type natriuretic
68 peptide (NPPB, also known as BNP) yielding hazard ratios (HRs) of 1.74 (95% confidence
69 interval, CI: 1.67-1.81; $P=8.7\times 10^{-173}$) and 1.62 (95%CI: 1.54-1.69; $P=5.6\times 10^{-95}$), respectively, for
70 each SD increase in circulating protein levels. NT-proBNP was also the second strongest for
71 association with heart failure, with an HR of 1.57 (95%CI: 1.48-1.66; $P=5.4\times 10^{-56}$) per SD. The
72 biomarker most strongly associated with heart failure was WAP four-disulfide core domain
73 protein 2 (WFDC2), a fibroblast-derived mediator of fibrosis also known as human epididymis
74 protein 4 (HE4)¹⁷, with an HR of 1.62 (95%CI: 1.54-1.72; $P=4.1\times 10^{-65}$) per SD. The proteins
75 most strongly associated with incident coronary artery disease were growth differentiation factor
76 15 (GDF15; HR, 1.31 [95%CI: 1.26-1.36] per SD; $P=2.0\times 10^{-45}$) and matrix metalloproteinase-12
77 (MMP12; HR, 1.29 [95%CI: 1.24-1.34] per SD; $P=1.1\times 10^{-39}$); those most strongly associated

78 with aortic stenosis were GDF15 (HR, 1.44 [95%CI: 1.29-1.60] per SD; $P=2.7\times 10^{-11}$) and
79 WFDC2 (HR, 1.40 [95%CI: 1.26-1.55] per SD; $P=7.3\times 10^{-10}$).

80 To gain insights into biological pathways associated with the identified proteins, we
81 carried the 820 observed protein-disease associations forward for pathway enrichment analysis
82 using the *Gene Ontology* resource¹⁸ via *Enrichr*¹⁹. The highest-scoring pathways for coronary
83 artery disease and heart failure included inflammatory and immune-related processes involving
84 leukocyte/lymphocyte chemotaxis and cellular response to cytokines (**Supplementary Fig. 3,**
85 **Supplementary Table 4**). Participants with coronary artery disease or heart failure during
86 follow-up were also enriched for apoptosis-related proteins such as those from the tumor
87 necrosis factor (TNF) receptor family. Proteins associated with aortic stenosis demonstrated
88 enrichment for peptidase inhibitor activity, consistent with recent work suggesting an important
89 role for certain peptidases in the progression of calcific aortic stenosis²⁰. Significantly enriched
90 cellular components included the endoplasmatic reticulum lumen (for all cardiac outcomes),
91 collagen-containing extracellular matrix (for coronary artery disease, heart failure, and atrial
92 fibrillation), and intracellular organelle lumen (for atrial fibrillation and aortic stenosis).

93

94 *Cis-Mendelian randomization analyses*

95 Next, we performed *cis*-MR analyses to infer causal effects of the identified proteins on
96 coronary artery disease, heart failure, atrial fibrillation, and aortic stenosis. Of 441 unique
97 Bonferroni-significant proteins in primary analyses, 430 (97.5%; corresponding to 802 protein-
98 disease associations) had at least one valid *cis*-pQTL (± 200 kilobases, $P < 5 \times 10^{-6}$, $R^2 < 0.1$)
99 (**Supplementary Table 5**). *F*-statistics were > 10 for all proteins other than myoglobin, which was
100 excluded from *cis*-MR analyses to minimize the risk of weak instrument bias. Median (IQR) *F*-
101 statistics and R^2 estimates (representing phenotypic variance explained by genetic instruments)
102 were 1515 (454-4050) and 3.2% (1.0-7.9%), respectively (**Supplementary Table 6**). Consistent

103 with the use of *cis*-pQTLs²¹, Steiger filtering did not identify any variants explaining more
104 variance in the outcome than the exposure (**Supplementary Table 7**).

105 Of 801 protein-disease associations examined in *cis*-MR analyses, 225 (28.1%;
106 representing 172/429 [40.1%] proteins) showed suggestive evidence of causality with $P < 0.05$
107 (**Fig. 3, Supplementary Table 8**). Because it is routinely recommended to evaluate *cis*-MR
108 findings across P -value and R^2 thresholds¹⁵, we performed multiple sensitivity analyses (using
109 genetic variants at $P < 5 \times 10^{-4} / < 5 \times 10^{-6} / < 5 \times 10^{-8}$ and $R^2 < 0.001 / < 0.01 / < 0.1 / < 0.2$) to evaluate the
110 robustness of the observed genetic associations (**Supplementary Table 9**). We further
111 performed MR-Egger (**Supplementary Table 9**), one-sample MR (**Supplementary Table 10**),
112 and multivariable-adjusted MR adjusting for proteins with shared pQTLs (**Supplementary**
113 **Tables 11-12**). A total of 97/225 (43.1%) genetic protein-disease associations were robust
114 across all sensitivity analyses. Genetic and observational analyses showed directional
115 consistency for 48/97 (49.5%) robust genetic associations, corresponding to 6.0% of all protein-
116 disease pairs tested in *cis*-MR analyses. These protein-disease associations all had positive
117 effect estimates, implying that increased protein concentrations may promote cardiac disease
118 risk and lowering would reduce risk.

119 Because proprotein convertase subtilisin/kexin type 9 (PCSK9) is an established causal
120 biomarker and therapeutic target for coronary artery disease²², we used this protein as a positive
121 control for our *cis*-MR analyses (despite PCSK9 not reaching Bonferroni significance for
122 coronary artery disease in epidemiological analyses). Each SD increase in genetically predicted
123 PCSK9 was associated with 1.28-fold odds of coronary artery disease (95%CI, 1.23-1.34;
124 $P = 8.6 \times 10^{-30}$), supporting a *cis*-MR strategy for the identification of potential causal protein-
125 disease associations. The proteins with the strongest epidemiological associations did not
126 generally show strong genetic associations with cardiac diseases. For instance, neither genetic
127 associations for WFDC2 nor GDF15 reached nominal significance (**Supplementary Table 8**).
128 Genetically predicted MMP12, which was among the most strongly associated proteins for

129 incident coronary artery disease in epidemiological analyses but was only modestly associated
130 with a protective effect on coronary artery disease risk in primary *cis*-MR analyses (OR, 0.98
131 [95%CI, 0.96-0.99] per SD; $P=0.002$).

132 Among all robust *cis*-MR analyses, the strongest genetic association was observed for
133 the interleukin-4 receptor (IL4R), which was associated with 1.07-fold odds of heart failure
134 (95%CI, 1.05-1.08; $P=3.9\times 10^{-16}$) per SD of genetically predicted levels, consistent with previous
135 work implicating heightened interleukin-4 signaling in profibrotic myocardial remodeling.²³ The
136 Kunitz-type protease inhibitor 1 (SPINT1; also known as hepatocyte growth factor activator
137 inhibitor type 1) had the strongest directionally concordant and robust genetic association for
138 coronary artery disease (OR, 1.13 [95%CI, 1.07-1.18] per SD; $P=1.3\times 10^{-6}$). The senescence-
139 and fibrosis-related insulin-like growth factor binding protein 7 (IGFBP7) was the most significant
140 protein for aortic stenosis (OR, 1.10 [95%CI, 1.05-1.14] per SD; $P=8.6\times 10^{-6}$). Spondin-1
141 (SPON1; OR, 1.10 [95%CI, 1.06-1.15]; $P=3.0\times 10^{-6}$) and adrenomedullin (ADM; OR, 1.23
142 [95%CI, 1.11-1.35]; $P=5.4\times 10^{-5}$) were among the strongest directionally concordant genetic
143 associations for atrial fibrillation; notably, colocalization analyses suggested shared causal
144 genetic variants between these two proteins and atrial fibrillation (posterior probability for shared
145 causal variants [H_4] >0.80). Colocalization analyses yielded inconclusive results for most other
146 protein-disease associations (**Supplementary Table 13**).

147

148 *Sex-specific protein-disease associations*

149 Although previous work suggested sex differences in the concentrations of
150 cardiovascular biomarkers²⁴, it remains unclear whether proteins assessed proteome-wide affect
151 cardiovascular disease risk differently in men vs. women. Therefore, we tested the multivariable-
152 adjusted associations of all 1,459 proteins with cardiac diseases in men ($n=19,612$) vs. women
153 ($n=24,701$). A total of 467 protein-disease associations met the primary significance threshold
154 ($P<0.05/5,836$) for men vs. 314 for women (**Supplementary Table 14**). Protein-disease

155 associations (for all 1,459 tested biomarkers) showed strong correlation between sexes,
156 indicated by a Pearson correlation coefficient (r) of 0.71 (**Supplementary Fig. 4**). The
157 correlation between sexes was strongest for heart failure ($r=0.79$) whereas it was comparatively
158 weaker for aortic stenosis ($r=0.47$).

159 We formally tested for sex interactions across all protein-disease associations reaching
160 significance ($P<0.05/5,836$) in at least one sex ($n=566$) (**Fig. 4, Supplementary Table 14**). Six
161 protein-disease associations had a Bonferroni-significant ($P<0.05/566$) sex-differential effect for
162 atrial fibrillation, including T-cell surface glycoprotein CD1c (CD1C; HR_{male}, 1.05 [95%CI, 0.99-
163 1.11] vs. HR_{female}, 0.90 [95%CI, 0.84-0.96] per SD; $P_{\text{interaction}}=6.9\times 10^{-5}$), cyclic ADP-ribose
164 hydrolase (CD38; HR_{male}, 1.01 [95%CI, 0.94-1.09] vs. HR_{female}, 1.22 [95%CI, 1.12-1.34];
165 $P_{\text{interaction}}=3.3\times 10^{-6}$), cathepsin L2 (CTSV; HR_{male}, 0.98 [95%CI, 0.22-1.05] vs. HR_{female}, 0.87
166 [95%CI, 0.80-0.94]; $P_{\text{interaction}}=7.3\times 10^{-5}$), NT-proBNP (HR_{male}, 1.69 [95%CI, 1.61-1.77] vs.
167 HR_{female}, 1.86 [95%CI, 1.74-1.99]; $P_{\text{interaction}}=3.7\times 10^{-5}$), paired immunoglobulin-like type 2 receptor
168 β (PILRB; HR_{male}, 1.05 [95%CI, 0.99-1.11] vs. HR_{female}, 1.20 [95%CI, 1.12-1.30]; $P_{\text{interaction}}=4.1\times 10^{-5}$),
169 and WFDC2 (HR_{male}, 1.20 [95%CI, 1.14-1.28] vs. HR_{female}, 1.32 [95%CI, 1.23-1.42];
170 $P_{\text{interaction}}=7.7\times 10^{-5}$). We also observed a sex differential effect for chymotrypsin C (CTRC; HR_{male},
171 1.04 [95%CI, 0.99-1.09] vs. HR_{female}, 0.88 [95%CI, 0.82-0.93]; $P_{\text{interaction}}=1.9\times 10^{-5}$) on coronary
172 artery disease. For all of these associations, P -values were lower and effect sizes greater in
173 women vs. men.

174

175 *Protein-based prediction of cardiac diseases*

176 We next derived and tested the predictive accuracy of protein-based risk scores in
177 addition to clinical risk factors in the UKB-PPP. We constructed protein-based, clinical, and
178 combined (using proteomic and clinical variables) risk scores in the training set (80%; $n=35,450$)
179 using least absolute shrinkage and selection operator (LASSO) regression with 10-fold cross-
180 validation. Protein-based risk scores included 64 proteins for coronary artery disease, 38 for

181 heart failure, 92 for atrial fibrillation, and 21 for aortic stenosis (**Supplementary Table 15,**
182 **Supplementary Fig. 5**). The prediction models' highest-weighted biomarkers were largely
183 overlapping with those showing the strongest associations in primary analyses.

184 Analyses in the testing cohort (20%; $n=8,863$) revealed that protein-based risk scores
185 effectively stratified the risk of incident events across outcomes (**Fig. 5A-B**). The protein-based
186 risk scores were strong independent predictors of incident events in multivariable-adjusted Cox
187 regression models, with HRs of 1.83 (95%CI, 1.62-2.06; $P=3.1\times 10^{-23}$) per SD increase for
188 coronary artery disease, 2.33 (95%CI, 1.99-2.73; $P=1.4\times 10^{-25}$) for heart failure, 2.22 (95%CI,
189 2.00-2.48; $P=7.5\times 10^{-48}$) for atrial fibrillation, and 1.91 (95%CI, 1.39-2.64; $P=7.5\times 10^{-5}$) for aortic
190 stenosis. The top 10% vs. bottom 90% of protein-based risk scores was associated with HRs of
191 1.72 (95%CI, 1.38-2.15; $P=1.8\times 10^{-6}$) for coronary artery disease, 3.62 (95%CI, 2.56-5.11;
192 $P=3.9\times 10^{-13}$) for heart failure, 4.02 (95%CI, 3.23-5.02; $P=3.9\times 10^{-35}$) for atrial fibrillation, and 2.50
193 (95%CI, 1.31-4.77; $P=5.5\times 10^{-3}$) for aortic stenosis.

194 ROC curve analyses revealed that adding proteomic data improved the prediction of
195 incident coronary artery disease, heart failure, and atrial fibrillation (**Fig. 5C**). The increment in
196 predictive accuracy compared with the clinical model—quantified using the area under the ROC
197 curve (AUC)—was most pronounced for atrial fibrillation (AUC, 0.801 [95%CI, 0.779-0.822] vs.
198 0.749 [95%CI, 0.727-0.772]; DeLong test: $P=2.0\times 10^{-10}$) and heart failure (AUC, 0.799 [95%CI,
199 0.769-0.830] vs. 0.732 [95%CI, 0.698-0.766]; $P=1.7\times 10^{-6}$), followed by coronary artery disease
200 (AUC, 0.757 [95%CI, 0.738-0.776] vs. 0.734 [95%CI, 0.714-0.754]; $P=1.4\times 10^{-4}$). There was no
201 statistically significant difference for aortic stenosis (AUC, 0.803 [95%CI, 0.754-0.852] vs. 0.789
202 [95%CI, 0.738-0.840]; $P=0.35$).

203 Given the disproportionately high weights for NT-proBNP in the protein-based risk scores
204 for atrial fibrillation and heart failure (**Supplementary Table 15**), we further evaluated the
205 performance of models including NT-proBNP alone vs. those incorporating all other biomarkers
206 in predicting these outcomes. We also excluded NPPB from the latter set of protein-based risk

207 scores since NPPB and NT-proBNP are encoded by the same gene and released in the
208 circulation in equimolar quantities²⁵. Compared with the score based on clinical factors alone
209 (0.749 [95%CI, 0.727-0.772]), inclusion of NT-proBNP significantly improved the prediction of
210 atrial fibrillation (AUC, 0.788 [95%CI, 0.766-0.811]; $P=1.1\times 10^{-6}$), resulting in a greater increment
211 in predictive accuracy than the score incorporating all proteins other than NT-proBNP and NPPB
212 (AUC, 0.777 [95%CI, 0.756-0.799]; $P=9.2\times 10^{-7}$) (**Supplementary Fig. 6**). In contrast, for heart
213 failure, the score incorporating all proteins except NT-proBNP and NPPB was associated with a
214 greater improvement in predictive accuracy vs. the clinical score (AUC, 0.786 [95%CI, 0.754-
215 0.818] vs. 0.732 [95%CI, 0.698-0.766]; $P=1.4\times 10^{-5}$) than the score incorporating NT-proBNP
216 alone (AUC, 0.756 [95%CI, 0.722-0.790]; $P=0.07$) (**Supplementary Fig. 7**).

217

218 Discussion

219 In a population-based cohort of ~45,000 middle-aged adults with circulating protein
220 measurements and longitudinal follow-up, we characterized the proteomic architecture of
221 incident coronary artery disease, heart failure, atrial fibrillation, and aortic stenosis. We identified
222 820 significant protein-disease associations with important roles (potentially mediating or
223 marking disease presence) for natriuretic peptides (e.g., NT-proBNP), inflammatory mediators
224 (e.g., MMP12), and apoptosis-related factors (e.g., GDF15) as predictors of cardiac diseases.
225 Genetic analyses suggested causal or mediating roles—either protective or deleterious—for a
226 substantial proportion of biomarkers identified in observational analyses. Sex-based analyses
227 suggested generally preserved associations between men and women, albeit with varying
228 weights of prediction including several biomarkers with strong sex interactions. Finally, we
229 constructed sparse protein-based risk scores that improved the prediction of cardiac disease
230 development in the general population. Our findings provide novel insights into the biology of
231 cardiac diseases, with implications for the prediction of incident cardiovascular diagnoses, and
232 with the potential for targeted prevention and treatment of these conditions.

233 The findings from this study offer insights into potential causal roles of proteins
234 associated with incident cardiac diseases. We found that ~6% of proteins identified in primary
235 analyses (and tested in *cis*-MR analyses) had putative causal associations that were
236 directionally discordant with those derived from epidemiological models (i.e., primary analyses).
237 However, we also identified many proteins (another ~6%) with genetic associations that were
238 robust across sensitivity analyses yet discordant with epidemiological estimates. By
239 systematically integrating observational and genetic data, our study corroborates and extends
240 previous studies reporting similar discrepancies between genetic and epidemiological
241 associations for selected proteins^{15,26}. For instance, consistent with prior research^{15,26}, primary
242 *cis*-MR analyses suggested a protective effect of genetically predicted MMP12 on coronary
243 artery disease risk, despite observational analyses indicating strong associations of higher
244 MMP12 levels with the same outcome. Similarly, higher measured but lower genetically
245 predicted levels of layilin (LAYN)—a relatively poorly understood protein that binds hyaluronic
246 acid and has roles in cellular adhesion and immunity²⁷—were strongly and robustly associated
247 with increased coronary artery disease risk. Whether these seemingly discordant observations
248 reflect inherent differences between disease onset versus progression or compensatory
249 response to subclinical disease requires further investigation. Nevertheless, several proteins had
250 consistent genetic and observational effects. One such example was SPON1, for which higher
251 levels (both measured and genetically predicted) were associated with increased atrial fibrillation
252 risk. SPON1 is an extracellular protein expressed in tissues such as the heart and brain that has
253 been implicated in Alzheimer’s dementia through its role in amyloid- β precursor protein
254 processing²⁸. Previous protein-focused analyses in patients with heart failure showed that the
255 presence of atrial fibrillation was associated with activation of amyloid- β -related pathways, with
256 SPON1 as one of the most strongly upregulated proteins in those with atrial fibrillation²⁹. These
257 data, together with colocalization findings indicating shared causal variants for SPON1 and atrial
258 fibrillation, collectively suggest that SPON1 not only marks presence of a pro-arrhythmic

259 substrate, but could also represent an upstream therapeutic target for preventing and/or treating
260 atrial fibrillation. Given the paucity of identified robust biomarkers mediating atrial fibrillation risk,
261 more data are needed regarding SPON1 and its role in arrhythmogenesis.

262 Biomarkers identified in proteomic analyses are often markers of already established
263 disease, rather than mediators of disease biology. In this regard, our analyses identified several
264 inflammation- and apoptosis-related proteins as strong predictors marking risk for cardiac
265 disease but were unlikely causal biomarkers. WFDC2 (also known as HE4)—a profibrotic
266 protease inhibitor with a potential role in natural immunity¹⁷—emerged as the strongest
267 proteomic predictor of heart failure. Previous research in hospitalized heart failure patients
268 demonstrated associations of circulating WFDC2 with disease severity as well as kidney
269 function³⁰. As WFDC2 is expressed exclusively in non-cardiovascular tissues such as the
270 respiratory tract, male and female genitourinary system, and kidneys³¹, it is likely that the strong
271 associations of WFDC2 with cardiac outcomes stem from peripheral organ responses rather
272 than indicating direct cardiac dysfunction or vascular damage. Similarly, GDF15—another
273 pleiotropic protein expressed across multiple organ systems³²—was the strongest biomarker for
274 coronary artery disease. As a member of the transforming growth factor- β (TGF- β) superfamily,
275 GDF15 is upregulated in response to external stressors (e.g., inflammation, hypoxia, and
276 oxidative stress) and is believed to reflect the cumulative impact of both acute and chronic
277 exposure to cellular stressors³². Recent data suggest GDF15 as an independent prognostic
278 biomarker for individuals with established atherosclerotic cardiovascular disease³³. The present
279 study builds upon this work, extending the applicability of WFDC2 and GDF15 as cardiac
280 biomarkers to the general population. Nevertheless, *cis*-MR analyses found no evidence of
281 causality in the associations of these proteins with cardiac diseases. These findings collectively
282 suggest that inflammation- and apoptosis-related biomarkers such as WFDC2 and GDF15
283 represent early disease markers without causal involvement in the pathogenesis of cardiac

284 diseases, consistent with their pleiotropic and non-specific effects in response to tissue damage
285 across organs. Much the same may be said about the natriuretic peptides.

286 In this regard, findings of this study corroborate the role of natriuretic peptides as
287 important cardiac biomarkers, grossly up-regulated in those with already prevalent but sub-
288 clinical cardiovascular disease states. In addition to the established association of natriuretic
289 peptide elevation with so-called “pre-heart failure”³⁴, the result of this study support the use of
290 NT-proBNP to guide atrial fibrillation screening. NT-proBNP is the N-terminal fragment of
291 proBNP (i.e., the precursor of biologically active NPPB, also known as BNP) and is primarily
292 released by ventricular and/or atrial cardiomyocytes in response to elevated myocardial wall
293 tension²⁵. Proteomic association analyses identified NT-proBNP as one of the strongest
294 predictors of heart failure, consistent with its routine clinical use in diagnosing and risk-stratifying
295 this condition²⁵. Despite its widespread use in the setting of heart failure (where it was a strong
296 marker), the results of this study suggest that NT-proBNP’s strongest disease associations were
297 observed for atrial fibrillation; ROC curve analyses also revealed that NT-proBNP contributed
298 more to the prediction of atrial fibrillation than to that of heart failure. These results extend prior
299 work demonstrating strong associations of circulating NT-proBNP with incident atrial fibrillation³⁵
300 and align with recent data from the LOOP trial suggesting that individuals with elevated NT-
301 proBNP levels may derive more clinical benefit from atrial fibrillation screening than those with
302 lower levels³⁶. Collectively, these findings provide support for the use of NT-proBNP to guide
303 atrial fibrillation screening in the general population.

304 Another important finding from this study was evidence for biological sex differences
305 underlying cardiac disease risk in men and women. The strongest sex interaction across all
306 tested proteins was observed for CD38, which was significantly more strongly associated with
307 incident atrial fibrillation in female than in male participants. CD38 is a glycoprotein expressed
308 across various immune cells including lymphocytes and plasma cells³⁷. Previous research
309 suggests that CD38 is causally implicated in autoimmune diseases such as rheumatoid arthritis

310 and systemic lupus erythematosus^{37,38}. As a history of autoimmune diseases represents a risk
311 factor for atrial fibrillation that affects women more strongly than men³⁹, it could be possible that
312 the observed sex differences for CD38 reflect a more important role for immunity-related
313 pathways in women vs. men. Furthermore, some of the largest differences in protein-disease
314 associations between sexes were observed for aortic stenosis. Prior histological work in aortic
315 stenosis patients suggests distinct tissue composition differences between men and women,
316 with women showing less valvular calcification but more fibrosis than men⁴⁰. Building upon these
317 findings, we identified several sex-specific senescence-associated biomarkers (e.g., IGFBP7
318 and TNF) associated with incident aortic stenosis in female, but not male, participants. Notably,
319 *cis*-MR analyses indicated a potential causal effect of IGFBP7 on aortic stenosis risk. IGFBP7
320 plays a role in the senescence-associated secretory phenotype, promoting cell cycle arrest upon
321 tissue damage and leading to apoptosis and tissue fibrosis⁴¹. These findings support an
322 important role for fibrotic valve remodeling in aortic stenosis, particularly among women,
323 identifying IGFBP7 as a mediator and potential therapeutic target in this population.

324 While this study benefits from a large sample size and the use of state-of-the-art
325 proteomic profiling methods, findings must be interpreted in the context of limitations. First, the
326 strength and quantity of protein-disease associations for each outcome were influenced by
327 statistical power and, consequently, the number of cases per outcome. Conditions with lower
328 incidence rates during follow-up (such as aortic stenosis) had fewer proteomic associations.
329 Second, the study population was predominantly white, precluding generalization to other
330 races/ethnicities. Third, any causal inference using MR relies on the validity of the underlying
331 instrumental variable assumptions. This study utilized a robust *cis*-MR framework (facilitating the
332 adherence to these assumptions^{21,42}) and tested the robustness of the genetic associations
333 through many sensitivity analyses, suggesting potential causal roles for the identified proteins.
334 Nevertheless, prioritized therapeutic targets remain to be evaluated in animal experiments and
335 eventually human trials. Fourth, not all proteins identified in primary analyses had strong *cis*-

336 pQTLs, precluding adequate *cis*-MR analyses for these proteins. Additionally, genetic instrument
337 strength varied across proteins. Instruments with more variants have greater power to detect
338 statistically significant genetic protein-disease associations, potentially leading to an
339 underestimation of associations for instruments with fewer variants. Nevertheless, we minimized
340 type II error by adopting a lenient *P*-value threshold ($P < 0.05$) to indicate statistical significance
341 for primary *cis*-MR analyses and prioritizing genetic protein-disease associations that were
342 robust to many sensitivity analyses.

343 Leveraging a population-based cohort of ~45,000 participants with comprehensive
344 genotypic, proteomic, and longitudinal clinical profiling, this study characterized the circulating
345 proteome associated with incident coronary artery disease, heart failure, atrial fibrillation, and
346 aortic stenosis. The study findings support new applications for established biomarkers (e.g.,
347 atrial fibrillation surveillance using NT-proBNP) and identify strong and potentially useful
348 predictors of cardiac diseases (e.g., WFDC2 for heart failure). Additionally, this study prioritized
349 potential therapeutic targets within the circulating proteome and elucidated biological sex
350 differences. These results demonstrate the potential of proteomics in refining cardiovascular risk
351 assessment and lay a foundation for future investigations to uncover novel disease mechanisms
352 and assess the clinical utility of protein-based prevention strategies for cardiac diseases.

353

354 **Methods**

355 *Study design and participants*

356 The study design is illustrated in **Fig. 1**. The UK Biobank is a population-based cohort of
357 ~500,000 volunteers aged 40-69 years at the time of study enrolment, recruited from 22
358 assessment centers across the United Kingdom during 2006-2010.⁴³ At enrolment, participants
359 provided informed consent; underwent physical examination; provided details on
360 sociodemographic characteristics, lifestyle factors, medical history, and medication use; and

361 donated blood samples. Follow-up for incident events occurred via linkage to electronic health
362 records through March 2020.

363 The UK Biobank Pharma Proteomics Project (UKB-PPP) is a precompetitive consortium
364 of 13 biopharmaceutical companies funding the generation of blood-based proteomic data in a
365 subset of UK Biobank participants.^{16,43} Upon release, the sponsors have no direct role in
366 research activities of these features as is the case for the present work. The UKB-PPP includes
367 54,306 participants, of whom 46,673 (85.9%) were randomly selected from baseline, 6,385
368 (11.8%) were preselected by UKB-PPP consortium members based on certain characteristics of
369 interest (e.g., disease status or genetic ancestry), and 1,268 (2.3%) were selected because they
370 attended multiple visits of the COVID-19 case-control imaging study.¹⁶ We considered 52,705
371 participants with baseline proteomic data passing quality control for inclusion in the present
372 study (**Fig. 1**). Participants were excluded if they had missing data for >10% of assay
373 measurements or if they had missing data on self-reported race/ethnicity or genetic ancestry.
374 We also excluded individuals inferred to be related (closer than third degree; kinship coefficient
375 >0.0884) and those with self-reported or physician-ascertained coronary artery disease, heart
376 failure, atrial fibrillation, or aortic stenosis at baseline (see below for disease definitions).

377 The UK Biobank was approved by the North West Multi-center Research Ethics
378 Committee. All analyses were conducted under UK Biobank application number 7089. The Mass
379 General Brigham Institutional Review Board approved the secondary use of these data.

380

381 *Protein measurements and proteomic data processing*

382 A detailed overview of the protein measurement strategy and proteomic data processing
383 steps in the UKB-PPP can be found in the **Supplementary Note** and has been described
384 previously.⁴⁵ In brief, blood samples donated by UKB-PPP study participants at study baseline
385 underwent proteomic profiling using the *Olink Explore 1536* platform (Olink Proteomics, Inc;
386 Waltham, MA), which uses proximity extension assay technology to measure 1,472 protein

387 analytes across four different panels (the *Cardiometabolic, Inflammation, Neurology, and*
388 *Oncology* panels) representing 1,463 unique proteins (**Supplementary Table 1**). We excluded
389 proteins with >10% missingness in the final study cohort (cathepsin S, nucleophosmin, tumor-
390 associated calcium signal transducer 2, and procollagen C-endopeptidase enhancer 1;
391 **Supplementary Table 2**) and imputed the remaining 1.1% of missing protein values using *K*-
392 nearest neighbors ($K=10$) as done previously.⁴⁴ All 1,459 protein markers underwent *Z*-score
393 transformation prior to analysis.

394

395 *Outcome ascertainment*

396 Follow-up for incident outcomes occurred through linkage to national health records until
397 March 2020. Incident events were defined by the occurrence of (1) at least one qualifying ICD-9
398 or ICD-10 code for a corresponding in- or outpatient diagnosis (as either a primary or secondary
399 disease diagnosis); or (2) at least one *Office of Population Censuses and Surveys Classification*
400 *of Surgical Operations and Procedures* (OPCS) code for a qualifying procedure (e.g., coronary
401 artery revascularization for coronary artery disease). The specific codes used to define each
402 outcome are listed in **Supplementary Table 16** and were used as done previously.^{46,47}

403

404 *Proteomic association analyses*

405 Primary analyses tested the associations of circulating protein levels with incident cardiac
406 events using Cox proportional hazards models adjusted for age, age², sex, self-reported
407 race/ethnicity, the first ten principal components of genetic ancestry, smoking, normalized
408 Townsend deprivation index, BMI, systolic blood pressure, antihypertensive medication use,
409 total cholesterol, HDL cholesterol, cholesterol-lowering medication use, serum creatinine (as a
410 measure of kidney function), and prevalent type 2 diabetes. In addition, to increase the
411 specificity of the detected protein associations for a given disease (e.g., coronary artery
412 disease), we included the other cardiac outcomes (e.g., heart failure, atrial fibrillation, and aortic

413 stenosis) as time-varying covariates using the *tmerge()* function in R (*survival* package).⁴⁸ A
414 detailed description of the study covariates can be found in the **Supplementary Note**.
415 Bonferroni-corrected $P < 0.05/5,836$ ($0.05/[1,459 \times 4]$ or $\sim 8.6 \times 10^{-6}$) indicated statistical significance
416 for the primary analyses.

417

418 *Pathway enrichment analyses*

419 Pathway enrichment analyses evaluated whether certain protein groups representing
420 biologically distinct pathways were disproportionately up- or downregulated in individuals with
421 incident cardiac events. Top biological functions, molecular pathways, and cellular components
422 were queried for each outcome using the *Gene Ontology* resource¹⁸ via *Enrichr*.¹⁹ Enrichment
423 tests were performed against a background gene set including the genes corresponding to all
424 1,459 proteins tested in primary analyses. Gene sets with a false discovery rate-adjusted $P < 0.05$
425 were considered statistically significant.

426

427 *Cis-Mendelian randomization analyses*

428 We performed two-sample MR analyses to explore the causal roles of proteins that were
429 statistically significantly associated with one or more cardiac outcomes in epidemiological
430 models (i.e., primary analyses). These analyses tested the associations of protein quantitative
431 trait loci (pQTLs; genetic variants associated with circulating protein levels) with coronary artery
432 disease, heart failure, atrial fibrillation, and aortic stenosis. We obtained pQTL data from 35,571
433 UKB-PPP participants who had their circulating proteomes profiled using the *Olink Explore 1536*
434 platform.¹⁶ FinnGen (freeze 9; <https://r9.finngen.fi/>) was used for genetic association data for
435 coronary artery disease (cases/total participants: $n=43,518/377,277$), heart failure
436 ($n=27,304/377,277$), atrial fibrillation/flutter ($n=45,766/237,690$), and operated calcific aortic
437 stenosis ($n=9,153/377,277$). All genetic data were derived from individuals of European
438 ancestry, and there was no overlap between the exposure and outcome study cohorts.

439 Because the use of *cis*-pQTLs (pQTLs that map near the protein-encoding gene)
440 facilitates adherence to MR assumptions,^{21,26} we only used region-wide significant, largely
441 independent *cis*-variants to construct genetic instruments ($P < 5 \times 10^{-6}$, $R^2 < 0.1$, ± 200 kilobases). All
442 proteins with at least one valid *cis*-pQTL were tested. We used the inverse-variance-weighted
443 (IVW) method for genetic instruments with more than one *cis*-pQTL and the Wald ratio estimator
444 for instruments with only one *cis*-pQTL. IVW estimates were adjusted for residual correlation
445 between genetic variants.^{49,50} Sensitivity analyses were performed for all genetic protein-disease
446 associations with at least nominal significance (unadjusted $P < 0.05$) in primary *cis*-MR analyses.
447 These sensitivity analyses tested whether the observed genetic associations were robust to (1)
448 different instrument construction parameters ($R^2 < 0.001 / < 0.01 / < 0.1 / < 0.2$ and $P < 5 \times 10^{-4} / < 5 \times 10^{-6} / < 5 \times 10^{-8}$); (2) MR-Egger (which tests for horizontal pleiotropy); (3) one-sample MR (in an
449 independent UKB sample); and (4) multivariable-adjusted MR adjusting for genetic instruments
450 of proteins which may also exert *trans*-pQTL effects.
451

452 Genetic associations were considered robust if the effect estimates were directionally
453 consistent across all primary and sensitivity analyses and MR-Egger suggested no horizontal
454 pleiotropy. Additional details of the MR methodology can be found in the **Supplementary Note**.

455

456 *Protein-based prediction models*

457 We constructed protein-based risk scores to predict incident cardiac events in the UKB-
458 PPP. We created three risk scores for each cardiac outcome using logistic least absolute
459 shrinkage and selection operator (LASSO) regression, including (1) a score based on clinical
460 risk factors, (2) a score based on circulating proteins, and (3) a combined score (i.e., using
461 clinical risk factors and circulating proteins). The clinically evaluable variables used as covariates
462 in primary analyses (age, sex, self-reported race/ethnicity, smoking, BMI, systolic blood
463 pressure, antihypertensive medication use, total cholesterol, HDL cholesterol, cholesterol-
464 lowering medication use, serum creatinine, and type 2 diabetes) were fed into LASSO models

465 for the clinical risk scores. Circulating levels of the 1,459 proteins tested in primary analyses
466 were fed into LASSO models for the protein-based risk scores.

467 The study cohort was randomly divided into a training (80%; $n=35,450$) and testing (20%;
468 $n=8,863$) set. All risk scores were constructed in the training set using LASSO regression with
469 10-fold cross-validation. The performance of each prediction model was evaluated by receiver-
470 operating characteristic (ROC) curve analysis. Additional information on the construction and
471 evaluation of all prediction models can be found in the **Supplementary Note**.

472 References

- 473 1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and
474 injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global
475 Burden of Disease Study 2019. *The Lancet* **396**, 1204–1222 (2020).
- 476 2. Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V. & Roth, G. A. The Global
477 Burden of Cardiovascular Diseases and Risk: A Compass for Future Health. *J Am Coll*
478 *Cardiol* **80**, 2361–2371 (2022).
- 479 3. Mortensen, M. B., Nordestgaard, B. G., Afzal, S. & Falk, E. ACC/AHA guidelines superior
480 to ESC/EAS guidelines for primary prevention with statins in non-diabetic Europeans: the
481 Copenhagen General Population Study. *Eur Heart J* **38**, 586–594 (2017).
- 482 4. Nurmohamed, N. S. *et al.* Proteomics and lipidomics in atherosclerotic cardiovascular
483 disease risk prediction. *Eur Heart J* **44**, 1594–1607 (2023).
- 484 5. Fruchart, J. C. *et al.* The Residual Risk Reduction Initiative: A Call to Action to Reduce
485 Residual Vascular Risk in Patients with Dyslipidemia. *Am J Cardiol* **102**, 1K-34K (2008).
- 486 6. Figtree, G. A. *et al.* Mortality in STEMI patients without standard modifiable risk factors: a
487 sex-disaggregated analysis of SWEDEHEART registry data. *The Lancet* **397**, 1085–1094
488 (2021).
- 489 7. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes
490 for coronary artery disease in over a million participants. *Nat Genet* **54**, 1803–1815
491 (2022).
- 492 8. Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery
493 disease in genetically diverse populations. *Nat Med* **28**, 1679–1692 (2022).
- 494 9. Levin, M. G. *et al.* Genome-wide association and multi-trait analyses characterize the
495 common genetic architecture of heart failure. *Nat Commun* **13**, 1–15 (2022).
- 496 10. Miyazawa, K. *et al.* Cross-ancestry genome-wide analysis of atrial fibrillation unveils
497 disease biology and enables cardioembolic risk prediction. *Nat Genet* **55**, 187–197
498 (2023).
- 499 11. Small, A. M. *et al.* Multiancestry Genome-Wide Association Study of Aortic Stenosis
500 Identifies Multiple Novel Loci in the Million Veteran Program. *Circulation* **147**, 942–955
501 (2023).
- 502 12. Nurmohamed, N. S. *et al.* Targeted proteomics improves cardiovascular risk prediction in
503 secondary prevention. *Eur Heart J* **43**, 1569–1577 (2022).
- 504 13. Ganz, P. *et al.* Development and Validation of a Protein-Based Risk Score for
505 Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA*
506 **315**, 2532–2541 (2016).
- 507 14. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in
508 30,931 individuals. *Nat Metab* **2**, 1135–1148 (2020).
- 509 15. Henry, A. *et al.* Therapeutic Targets for Heart Failure Identified Using Proteomics and
510 Mendelian Randomization. *Circulation* **145**, 1205–1217 (2022).
- 511 16. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank
512 participants. *bioRxiv* <https://www.biorxiv.org/content/10.1101/2022.06.17.496443v1>
513 (2022).
- 514 17. LeBleu, V. S. *et al.* Identification of human epididymis protein-4 as a fibroblast-derived
515 mediator of fibrosis. *Nat Med* **19**, 227–231 (2013).

- 516 18. Carbon, S. *et al.* The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*
517 **49**, D325–D334 (2021).
- 518 19. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server
519 2016 update. *Nucleic Acids Res* **44**, W90–W97 (2016).
- 520 20. Choi, B. *et al.* Dipeptidyl peptidase-4 induces aortic valve calcification by inhibiting insulin-
521 like growth factor-1 signaling in valvular interstitial cells. *Circulation* **135**, 1935–1950
522 (2017).
- 523 21. Swerdlow, D. I. *et al.* Selecting instruments for Mendelian randomization in the wake of
524 genome-wide association studies. *Int J Epidemiol* **45**, 1600–1616 (2016).
- 525 22. Sabatine, M. S. PCSK9 inhibitors: clinical evidence and implementation. *Nat Rev Cardiol*
526 **16**, 155–165 (2018).
- 527 23. Peng, H. *et al.* Profibrotic Role for Interleukin-4 in Cardiac Remodeling and Dysfunction.
528 *Hypertension* **66**, 582–589 (2015).
- 529 24. Lau, E. S. *et al.* Sex Differences in Circulating Biomarkers of Cardiovascular Disease. *J*
530 *Am Coll Cardiol* **74**, 1543–1553 (2019).
- 531 25. Kim, H. N. & Januzzi, J. L. Natriuretic Peptide Testing in Heart Failure. *Circulation* **123**,
532 2015–2019 (2011).
- 533 26. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nat*
534 *Commun* **11**, 1–12 (2020).
- 535 27. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79
536 (2018).
- 537 28. Mahuron, K. M. *et al.* Layilin augments integrin activation to promote antitumor immunity.
538 *Journal of Experimental Medicine* **217**, (2020).
- 539 29. Ho, A. & Südhof, T. C. Binding of F-spondin to amyloid- β precursor protein: A candidate
540 amyloid- β precursor protein ligand that modulates amyloid- β precursor protein cleavage.
541 *Proc Natl Acad Sci U S A* **101**, 2548–2553 (2004).
- 542 30. Santema, B. T. *et al.* Pathophysiological pathways in patients with heart failure and atrial
543 fibrillation. *Cardiovasc Res* **118**, 2478–2487 (2022).
- 544 31. De Boer, R. A. *et al.* The WAP Four-Disulfide Core Domain Protein HE4: A Novel
545 Biomarker for Heart Failure. *JACC Heart Fail* **1**, 164–169 (2013).
- 546 32. Galgano, M. T., Hampton, G. M. & Frierson, H. F. Comprehensive analysis of HE4
547 expression in normal and malignant human tissues. *Modern Pathology* **19**, 847–853
548 (2006).
- 549 33. Wollert, K. C., Kempf, T. & Wallentin, L. Growth Differentiation Factor 15 as a Biomarker
550 in Cardiovascular Disease. *Clin Chem* **63**, 140–151 (2017).
- 551 34. Kato, E. T. *et al.* Growth differentiation factor 15 and cardiovascular risk: individual patient
552 meta-analysis. *Eur Heart J* **44**, 293–300 (2023).
- 553 35. Bozkurt, B. *et al.* Universal Definition and Classification of Heart Failure: A Report of the
554 Heart Failure Society of America, Heart Failure Association of the European Society of
555 Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal
556 Definition of Heart Failure. *J Card Fail* **27**, 387–413 (2021).
- 557 36. Patton, K. K. *et al.* N-Terminal Pro-B-Type Natriuretic Peptide Is a Major Predictor of the
558 Development of Atrial Fibrillation. *Circulation* **120**, 1768–1774 (2009).

- 559 37. Xing, L. Y. *et al.* Effects of Atrial Fibrillation Screening According to N-Terminal Pro-B-
560 Type Natriuretic Peptide: A Secondary Analysis of the Randomized LOOP Study.
561 *Circulation* **147**, 1788–1797 (2023).
- 562 38. Cole, S. *et al.* Integrative analysis reveals CD38 as a therapeutic target for plasma cell-
563 rich pre-disease and established rheumatoid arthritis and systemic lupus erythematosus.
564 *Arthritis Res Ther* **20**, 1–14 (2018).
- 565 39. Ostendorf, L. *et al.* Targeting CD38 with Daratumumab in Refractory Systemic Lupus
566 Erythematosus. *N Engl J Med* **383**, 1149–1155 (2020).
- 567 40. Tilly, M. J. *et al.* Autoimmune diseases and new-onset atrial fibrillation: a UK Biobank
568 study. *EP Europace* **25**, 804–811 (2023).
- 569 41. Simard, L. *et al.* Sex-related discordance between aortic valve calcification and
570 hemodynamic severity of aortic stenosis. *Circ Res* **120**, 681–691 (2017).
- 571 42. Severino, V. *et al.* Insulin-like growth factor binding proteins 4 and 7 released by
572 senescent cells promote premature senescence in mesenchymal stem cells. *Cell Death*
573 *Dis* **4**, e911–e911 (2013).
- 574 43. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
575 *Nature* **562**, 203–209 (2018).
- 576 44. Gadd, D. A. *et al.* Blood protein levels predict leading incident diseases and mortality in
577 UK Biobank. *medRxiv* <https://www.medrxiv.org/content/10.1101/2023.05.01.23288879v2>
578 (2023).
- 579 45. Zhong, W. *et al.* Next generation plasma proteome profiling to monitor health and
580 disease. *Nat Commun* **12**, 1–12 (2021).
- 581 46. Pirruccello, J. P. *et al.* Genetic analysis of right heart structure and function in 40,000
582 people. *Nat Genet* **54**, 792–803 (2022).
- 583 47. Honigberg, M. C. *et al.* Association of Premature Natural and Surgical Menopause With
584 Incident Cardiovascular Disease. *JAMA* **322**, 2411–2421 (2019).
- 585 48. Therneau, T. M., Lumley, T., Atkinson, E. & Crowson, C. Package 'survival'. [https://cran.r-](https://cran.r-project.org/web/packages/survival/survival.pdf)
586 [project.org/web/packages/survival/survival.pdf](https://cran.r-project.org/web/packages/survival/survival.pdf) (2023).
- 587 49. Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple
588 instrumental variables in Mendelian randomization: comparison of allele score and
589 summarized data methods. *Stat Med* **35**, 1880–1906 (2016).
- 590 50. Burgess, S., Zuber, V., Valdes-Marquez, E., Sun, B. B. & Hopewell, J. C. Mendelian
591 randomization with fine-mapped genetic data: Choosing from large numbers of correlated
592 instrumental variables. *Genet Epidemiol* **41**, 714 (2017).

595 **Acknowledgements**

596 All UK Biobank analyses were performed under application N° 7089.

597

598 **Acknowledgements**

599 A.S. is supported by the Belgian American Educational Foundation. J.L.J. is supported in
600 part by the Adolph M. Hutter MD Professorship. M.C.H. is supported by the U.S. NHLBI
601 (K08HL166687) and AHA (940166, 979465). P.N. is supported by the Paul & Phyllis Fireman
602 Endowed Chair in Vascular Medicine from the Massachusetts General Hospital and the U.S.
603 NHLBI (R01HL127564).

604

605 **Ethics declarations**

606 J.L.J. is a Trustee of the American College of Cardiology and reports board membership
607 of Imbria Pharmaceuticals; grant support from Abbott Diagnostics, Applied Therapeutics,
608 AstraZeneca, Bayer, HeartFlow, Innolife and Roche; previous consulting income from Abbott
609 Diagnostics, AstraZeneca, Bayer, Beckman-Coulter, Jana Care, Janssen, Novartis, Quidel,
610 Roche Diagnostics and Siemens; clinical endpoint committee/data safety monitoring board
611 membership for Abbott, AbbVie, CVRx, Pfizer and Takeda. M.C.H. reports consulting fees from
612 CRISPR Therapeutics and Comanche Biopharma; advisory board service for Miga Health; and
613 grant support from Genentech. P.N. reports research grants from Allelica, Amgen, Apple, Boston
614 Scientific, Genentech / Roche, and Novartis; personal fees from Allelica, Apple, AstraZeneca,
615 Blackstone Life Sciences, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly & Co,
616 Foresite Labs, Genentech / Roche, GV, HeartFlow, Magnet Biomedicine, Merck, and Novartis;
617 scientific advisory board membership of Esperion Therapeutics, Preciseli, and TenSixteen Bio;
618 scientific co-foundership of TenSixteen Bio, equity in MyOme, Preciseli, and TenSixteen Bio;
619 and spousal employment at Vertex Pharmaceuticals (all unrelated to the present work).

620

621 **Data availability**

622 Data supporting the results of the present study are available from the UK Biobank
623 (<https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>) to bona fide researchers
624 with institutional review board and UK Biobank approval. These analyses were performed using
625 the UK Biobank resource under application number 7089. The secondary use of these data was
626 approved by the Mass General Brigham institutional review board. Pathway enrichment
627 analyses were performed using the Gene Ontology resource via Enrichr
628 (<https://maayanlab.cloud/Enrichr/>). The UK Biobank Pharma Proteomics Project was used for
629 genetic association data for circulating proteins (i.e., protein quantitative trait locus data) through
630 Synapse (<https://doi.org/10.7303/syn51364943>). FinnGen was used for genetic association data
631 for coronary artery disease, heart failure, atrial fibrillation/flutter, and operated calcific aortic
632 stenosis (freeze 9; <https://r9.finnngen.fi/>).

633

634 **Code availability**

635 Code used for the main analyses of this study can be accessed at
636 https://github.com/aschuerm/ukbPPP_cardiac_diseases.

637

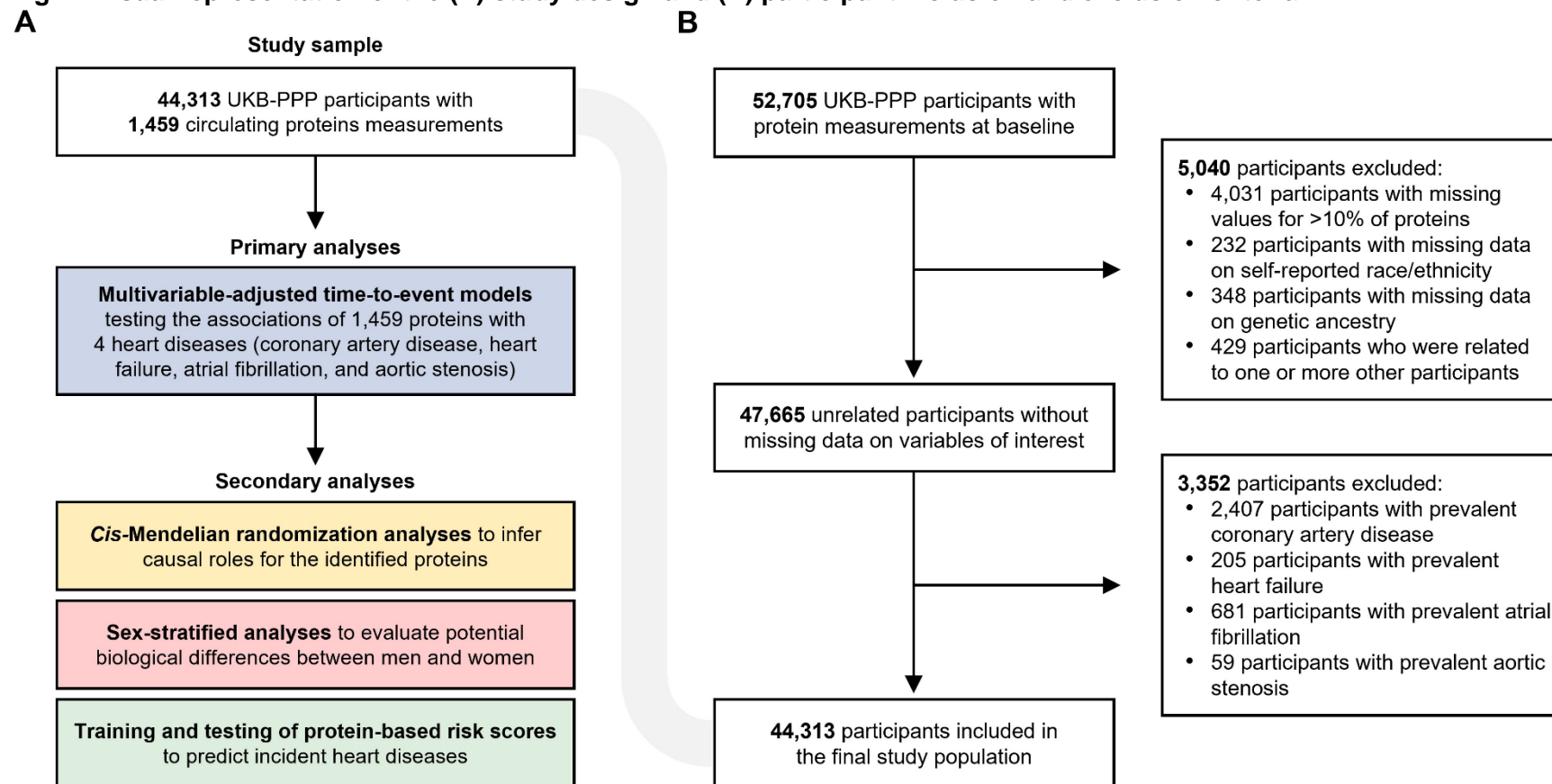
638 **Inclusion and ethics**

639 Inclusion and ethics standards have been reviewed where applicable.

640
641
642

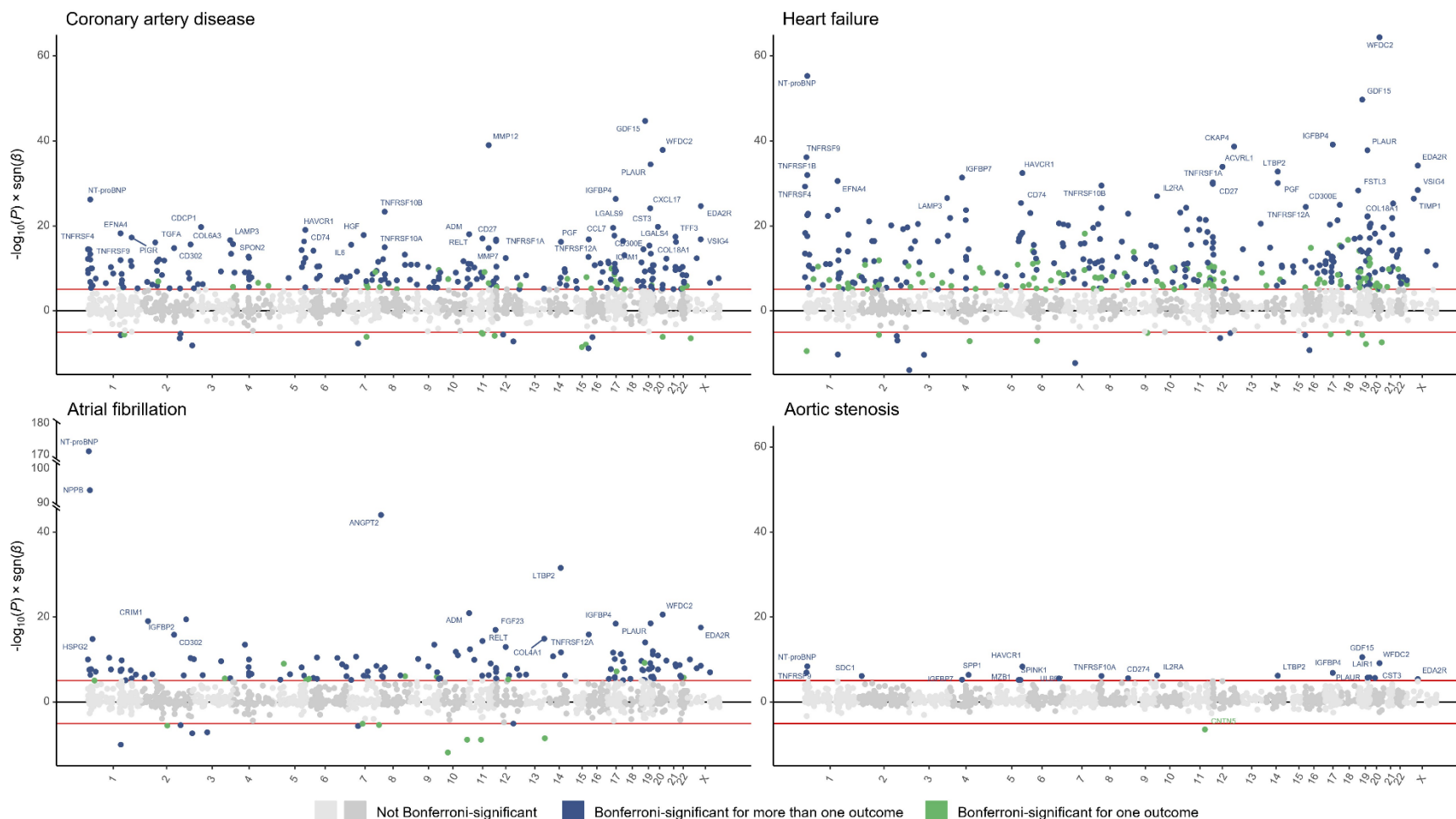
Figures

Fig. 1. Visual representation of the (A) study design and (B) participant inclusion and exclusion criteria.



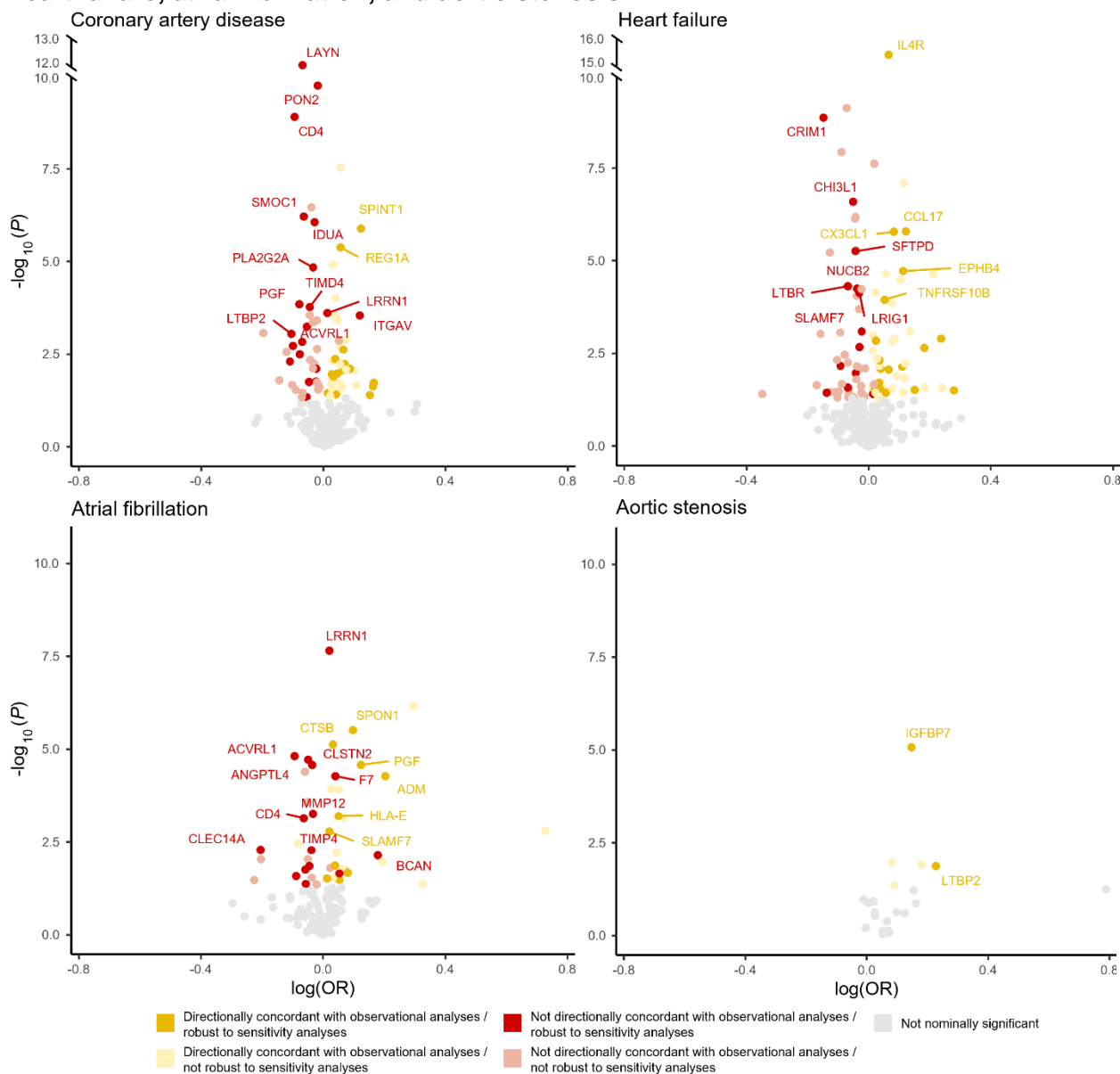
643 The present study tested the associations of circulating proteins with common cardiac diseases (coronary artery disease, heart
644 failure, atrial fibrillation, aortic stenosis) in the UK Biobank Pharma Proteomics Project (UKB-PPP). Primary analyses tested the
645 epidemiological associations of 1,459 circulating proteins with cardiac diseases in 44,313 UKB-PPP participants without these
646 diseases at baseline. Secondary analyses performed *cis*-Mendelian randomization analyses, tested for sex-specific effects, and
647 trained and tested protein-based risk scores.

648 **Fig. 2. Associations of circulating protein levels with incident coronary artery disease, heart failure, atrial fibrillation, and**
 649 **aortic stenosis.**



650 The Miami plots visualize the associations of all 1,459 Olink proteins with coronary artery disease, heart failure, atrial fibrillation, and
 651 aortic stenosis, tested using multivariable-adjusted Cox proportional hazards models (see *Methods*). The Y-axis indicates the $-\log_{10}(P)$
 652 value for each association, multiplied by 1 if the association was positive (i.e., $\beta > 0$) or -1 if the association was negative (i.e.,
 653 $\beta < 0$). The X-axis indicates the genetic position of each protein's encoding gene. Protein-disease associations with Bonferroni-
 654 corrected $P < 0.05$ (i.e., $P < 0.05/5,836$ or $\sim 8.6 \times 10^{-6}$) are shown in blue (if the protein was associated with more than one outcome) or
 655 green (if the protein was not associated with more than one outcome).

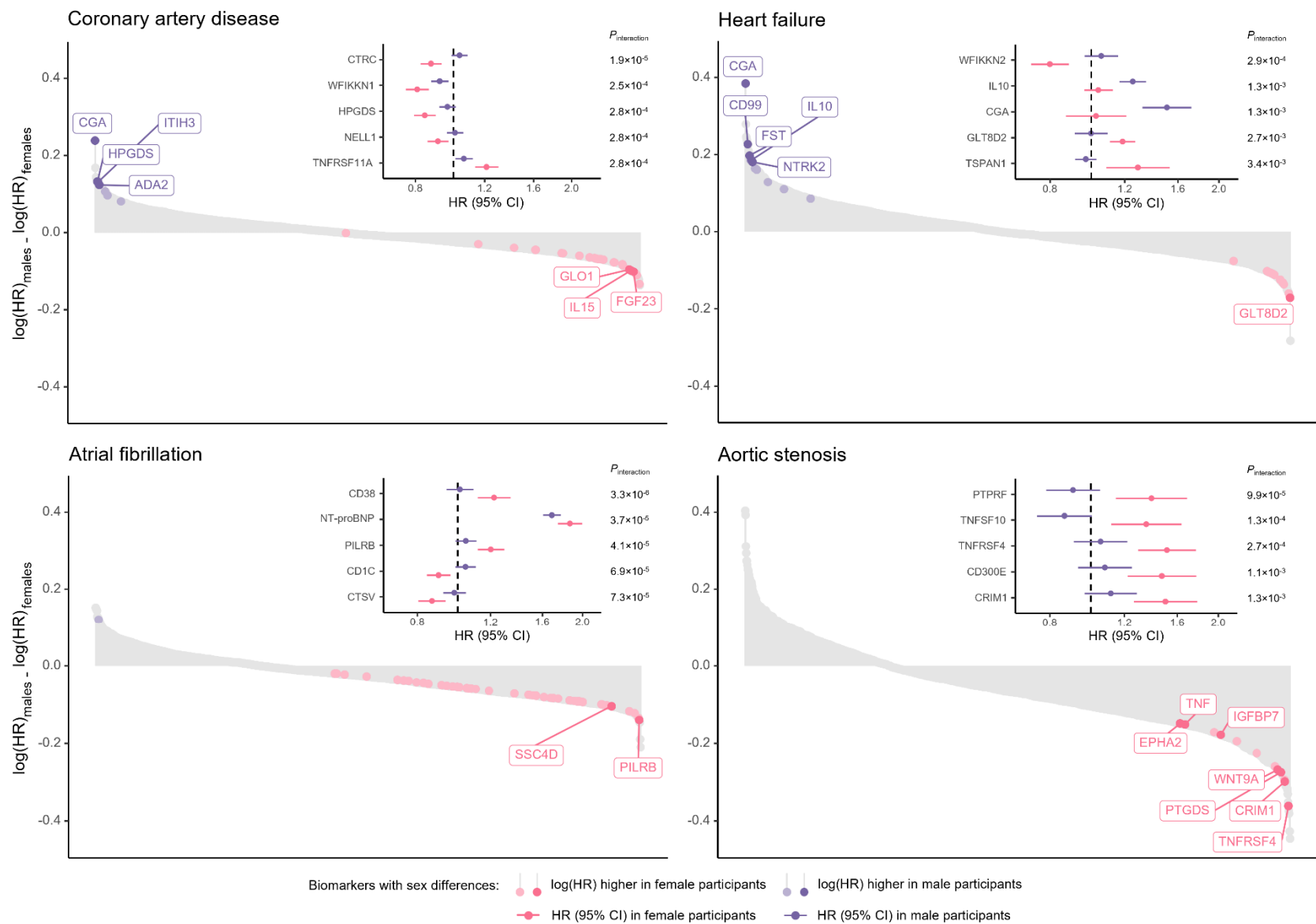
656 **Fig. 3. Associations of genetically predicted protein levels with coronary artery disease,**
 657 **heart failure, atrial fibrillation, and aortic stenosis.**



658 The volcano plots visualize the genetic associations of all proteins identified in primary analyses
 659 with their corresponding outcomes, by plotting each association's $-\log_{10}(P)$ against the
 660 corresponding $\log(\text{OR})$ per standard deviation increase in genetically predicted protein levels. All
 661 analyses represent *cis*-Mendelian randomization (MR) analyses using the inverse-variance-
 662 weighted (IVW; for instruments with two or more variants) or Wald ratio method (for instruments
 663 with one variant). Genetic instruments were constructed using *cis*-variants associated with
 664 circulating protein levels at $P < 1 \times 10^{-4}$ clumped at $R^2 < 0.1$. Associations with $P < 0.05$ are shown in
 665 yellow (if the primary *cis*-MR analysis was directionally consistent with the observational
 666 analysis) or red (if the primary *cis*-MR analysis was not directionally consistent with the
 667 observational analysis). Bright colors and protein labels indicate robustness against sensitivity
 668 analyses (see *Methods*), whereas dull colors indicate no robustness against sensitivity
 669 analyses. CI indicates confidence interval; OR, odds ratio.

670
671

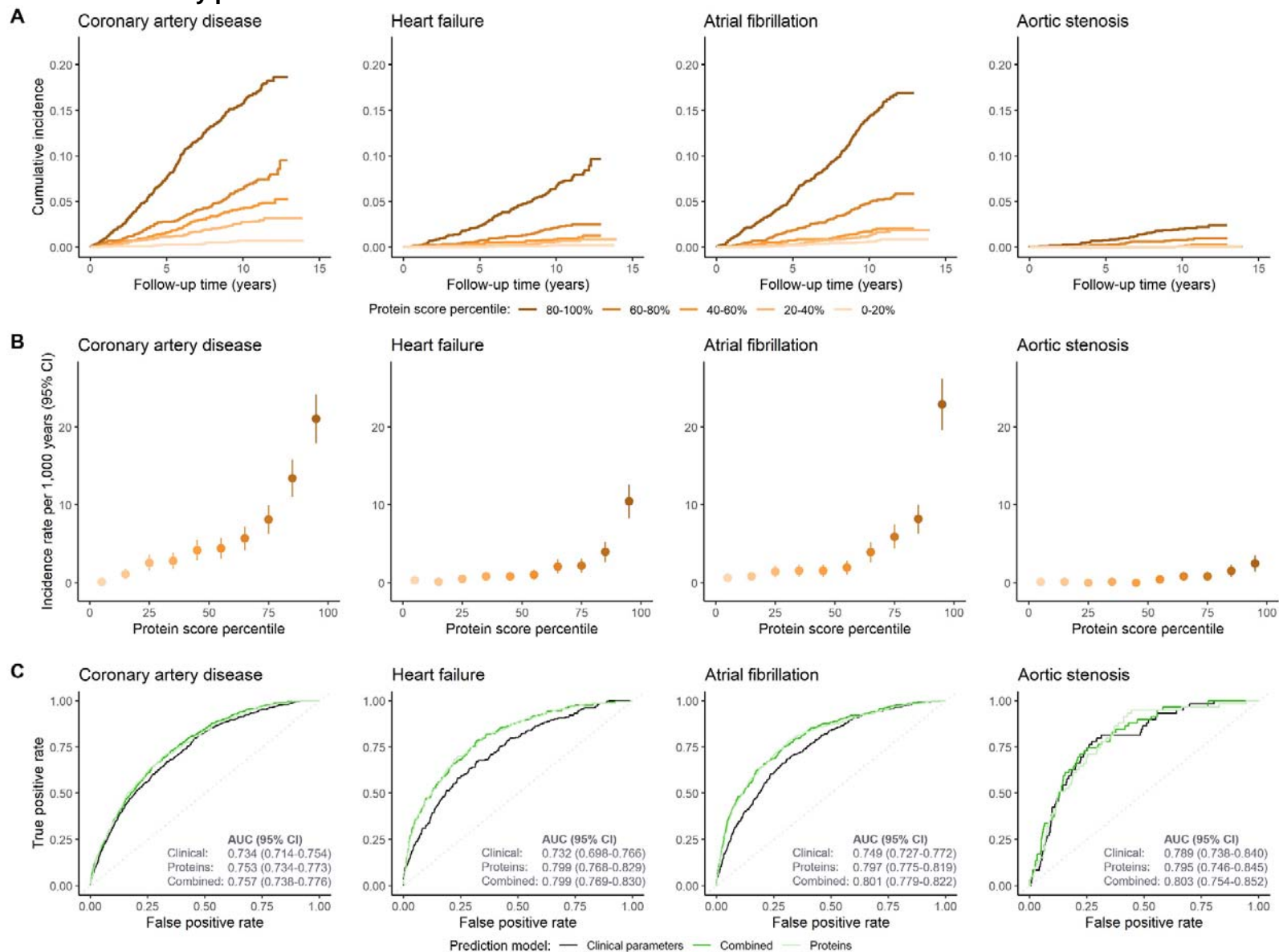
Fig. 4. Sex-specific protein-disease associations and protein-by-sex interactions for coronary artery disease, heart failure, atrial fibrillation, and aortic stenosis.



672 The lollipop plots depict the differences in effect sizes between male and female participants ($\log[\text{HR}]_{\text{males}} - \log[\text{HR}]_{\text{females}}$) for all tested
673 protein-disease associations. Bright colors with labels represent proteins with $P < 0.05/5,836$ in one sex without nominal significance
674 ($P > 0.05$) in the other sex; dull colors represent proteins with $P < 0.05/5,836$ in one sex and at least nominal significance ($P < 0.05$) in
675 the other sex. In addition, all proteins indicated in color had suggestive evidence for interaction by sex ($P_{\text{interaction}} < 0.05$). Forest plots
676 depict the sex-stratified protein-disease associations (purple for men, pink for women) for the five proteins with the strongest sex-
677 protein interactions. $P_{\text{interaction}}$ indicates the P -value for the interaction term between “sex” and the indicated protein on the
678 corresponding outcome. All associations were tested using multivariable-adjusted Cox proportional hazards models (see Methods).

679
680

Fig. 5. Risk (A-B) stratification and (C) prediction of incident coronary artery disease, heart failure, atrial fibrillation, and aortic stenosis by protein-based risk scores.



681 The indicated plots depict (A) the cumulative incidence of each outcome (calculated using the Kaplan-Meier method) by protein-
682 based score quintiles; (B) incidence rate estimates with 95% confidence intervals (CIs) according to protein-based score deciles; and
683 (C) the accuracies of the clinical, proteomic, and combined risk scores in predicting the indicated outcomes (quantified using the area
684 under the receiver-operating characteristic curve [AUC] with corresponding 95%CI). All analyses were performed in the UK Biobank
685 Pharma Proteomics Project (UKB-PPP) testing set ($n=8,863$). Participants were followed for a median (interquartile range) follow-up
686 of 11.1 (10.4-11.8) years.

687 **Tables**

688

689 **Table 1. Baseline characteristics of UK Biobank Pharma Proteomics Project (UKB-PPP)**
 690 **participants included in the present study (N=44,313).**

	UKB-PPP participants (N=44,313)
Age at blood draw, y	56.4 ± 8.2
Female, <i>n</i>	24,701 (55.7%)
Race/ethnicity, <i>n</i>	–
Asian	942 (2.1%)
Black	1,051 (2.4%)
White	41,481 (93.6%)
Mixed	300 (0.7%)
Other	539 (1.2%)
Smoking status, <i>n</i>	–
Never	24,709 (55.8%)
Previous	14,932 (33.7%)
Current	4,672 (10.5%)
Body mass index, kg/m ²	27.3 ± 4.7
Blood pressure, mmHg	–
Systolic blood pressure	139.5 ± 19.6
Diastolic blood pressure	82.3 ± 10.6
Blood biochemistry, mg/dL	–
Total cholesterol	221.4 ± 43.8
LDL cholesterol	138.3 ± 33.4
HDL cholesterol	56.4 ± 14.7
Triglycerides	129.8 (92.0 to 188.1)
Creatinine	0.82 ± 0.19
Type 2 diabetes mellitus, <i>n</i>	1,218 (2.7%)
Medication use, <i>n</i>	–
Cholesterol-lowering medication use	6,544 (14.8%)
Antihypertensive medication use	6,214 (14.0%)
Townsend deprivation index	-2.08 (-3.63 to 0.70)

691 Continuous variables are summarized as mean ± standard deviation or median (interquartile
 692 range), as appropriate. Categorical variables are summarized as *n* (%).

A**Study sample**

44,313 UKB-PPP participants with
1,459 circulating proteins measurements

**Primary analyses**

Multivariable-adjusted time-to-event models
testing the associations of 1,459 proteins with
4 heart diseases (coronary artery disease, heart
failure, atrial fibrillation, and aortic stenosis)

**Secondary analyses**

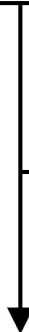
Cis-Mendelian randomization analyses to infer
causal roles for the identified proteins

Sex-stratified analyses to evaluate potential
biological differences between men and women

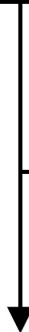
Training and testing of protein-based risk scores
to predict incident heart diseases

B

52,705 UKB-PPP participants with
protein measurements at baseline



47,665 unrelated participants without
missing data on variables of interest



44,313 participants included in
the final study population

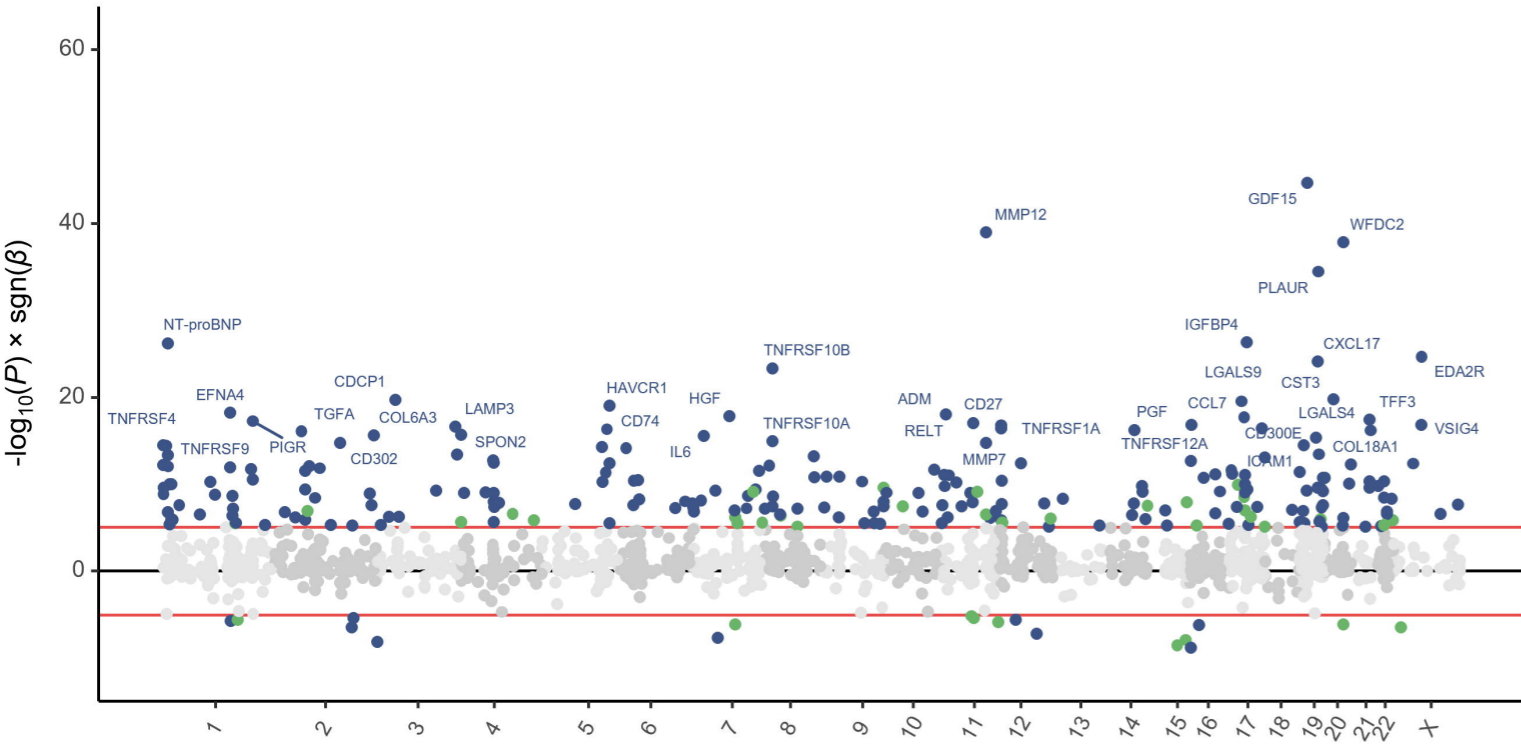
5,040 participants excluded:

- 4,031 participants with missing values for >10% of proteins
- 232 participants with missing data on self-reported race/ethnicity
- 348 participants with missing data on genetic ancestry
- 429 participants who were related to one or more other participants

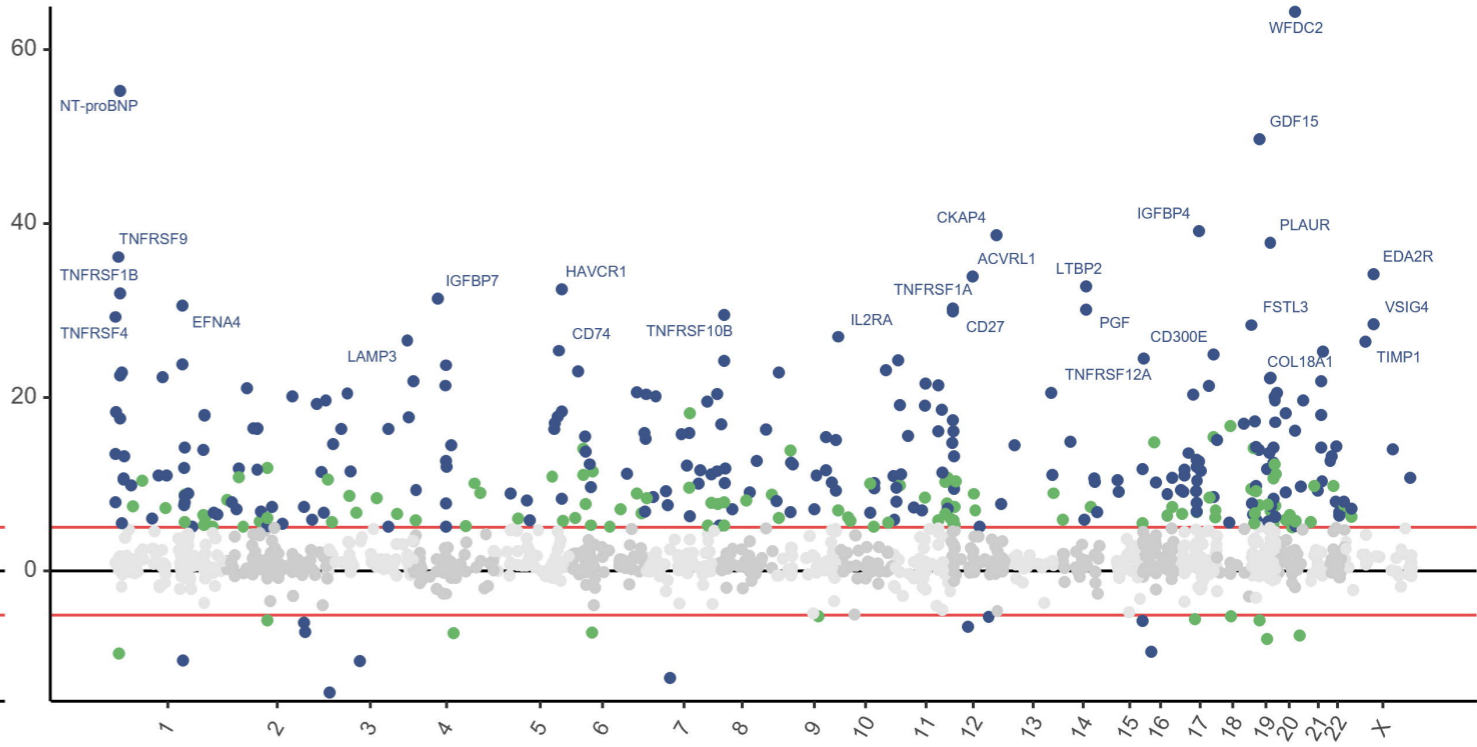
3,352 participants excluded:

- 2,407 participants with prevalent coronary artery disease
- 205 participants with prevalent heart failure
- 681 participants with prevalent atrial fibrillation
- 59 participants with prevalent aortic stenosis

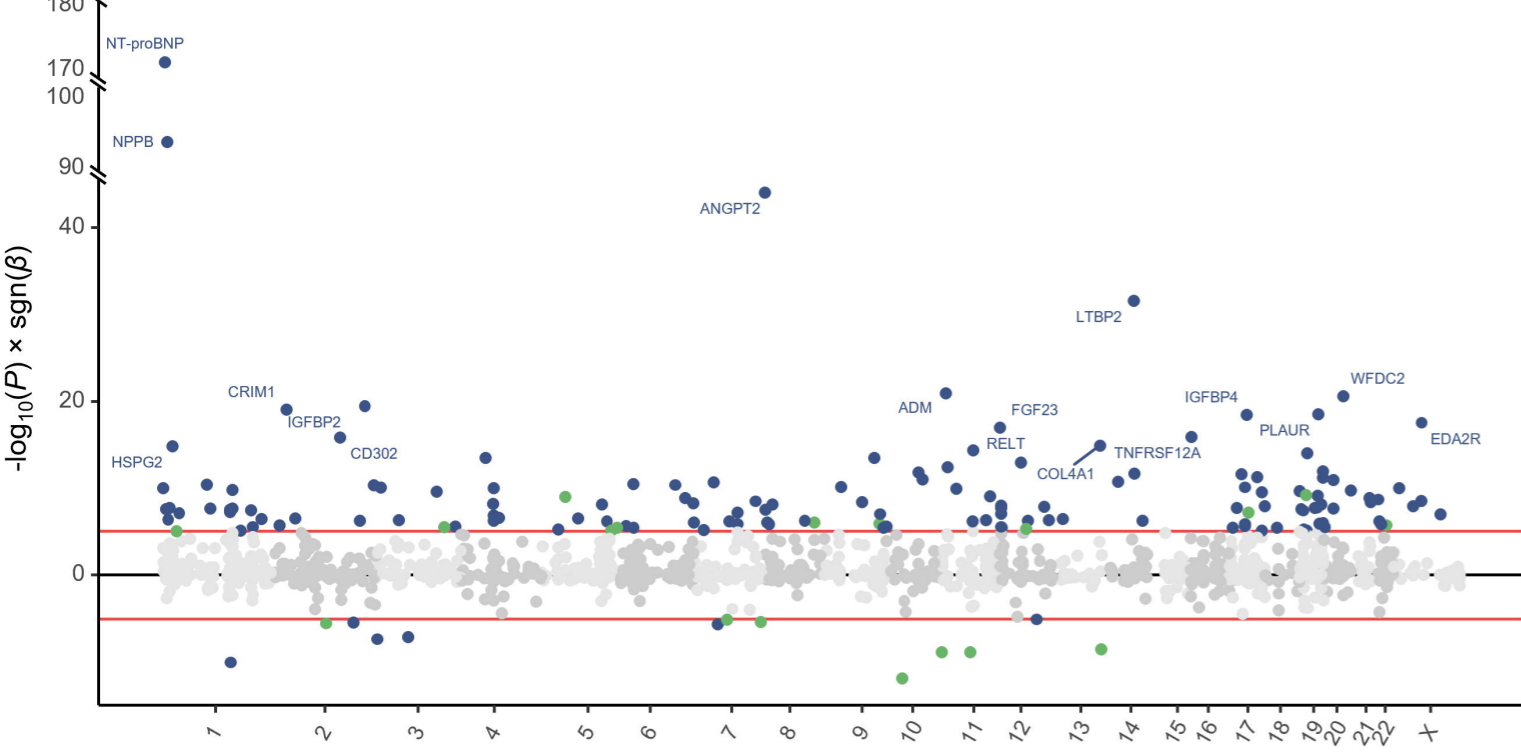
Coronary artery disease



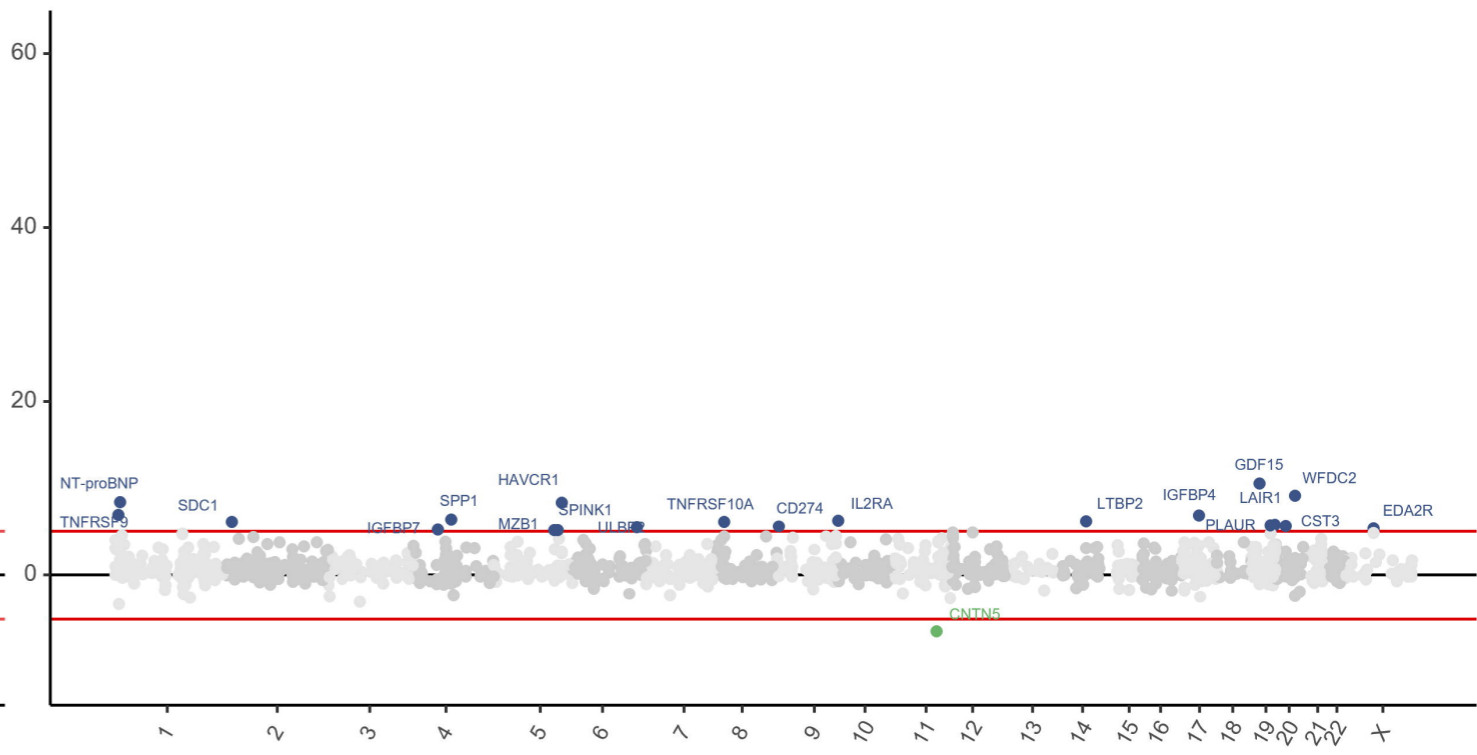
Heart failure



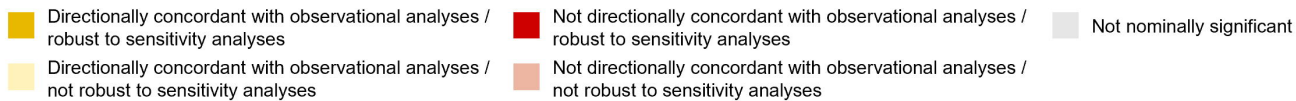
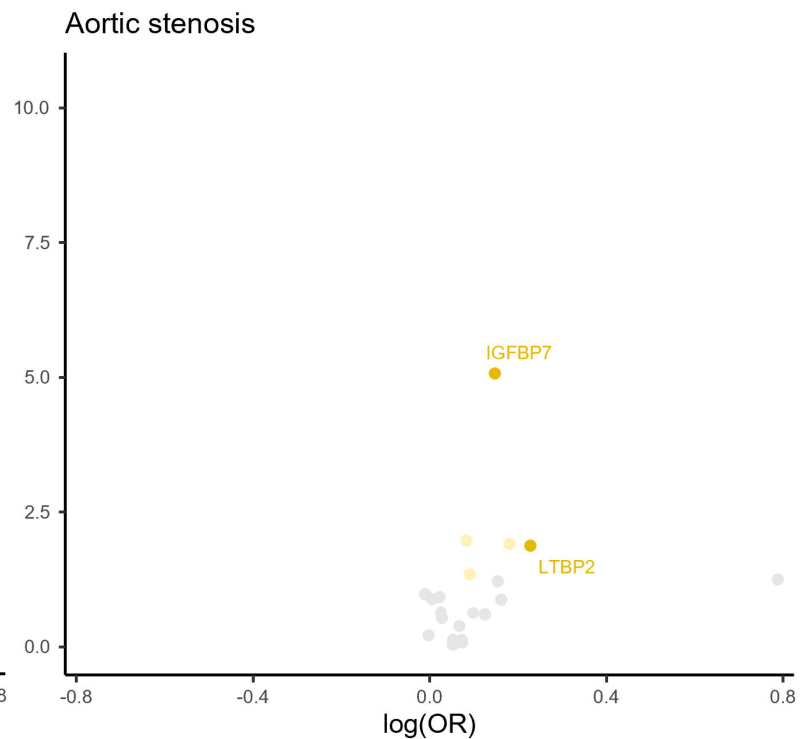
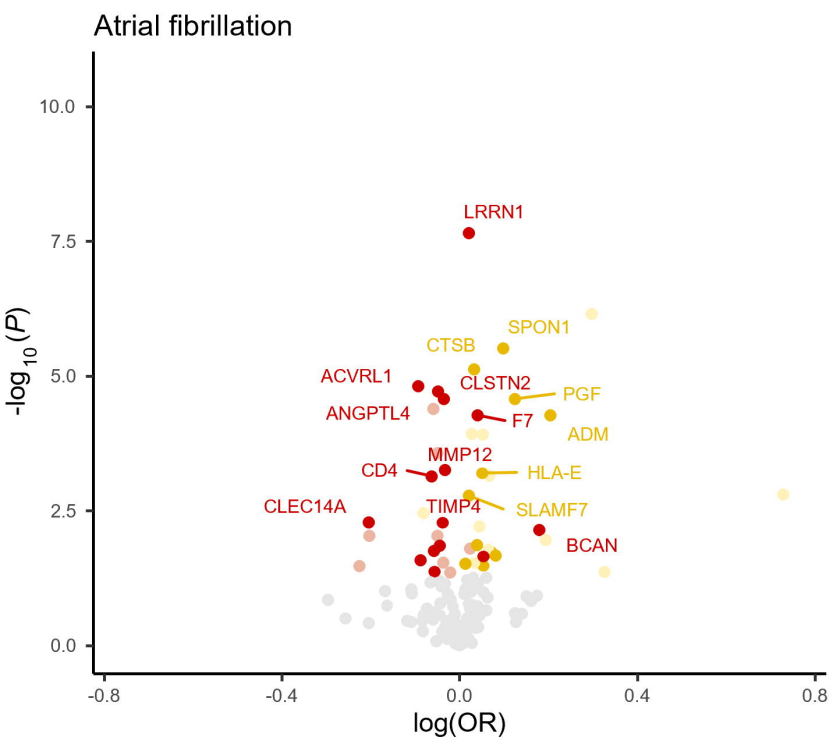
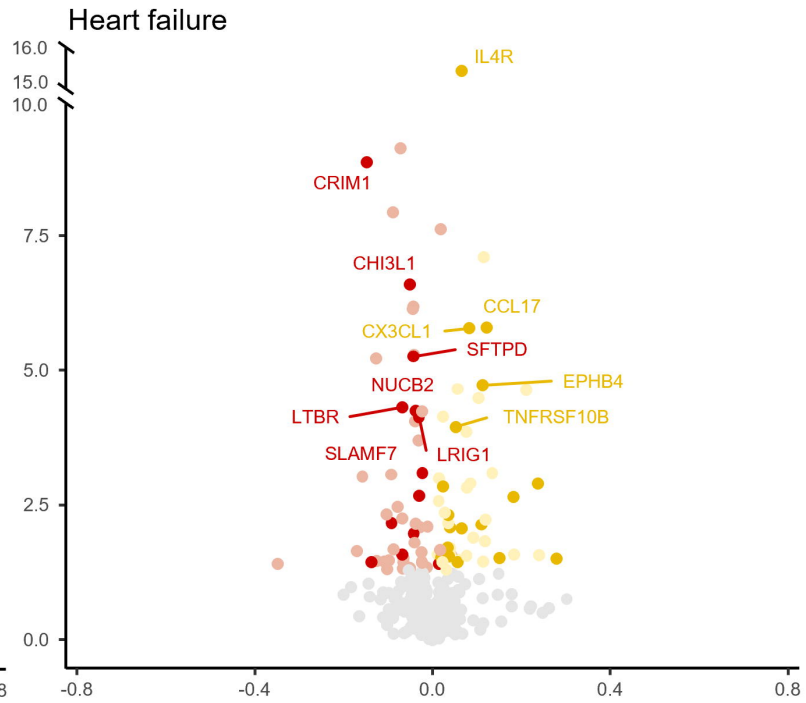
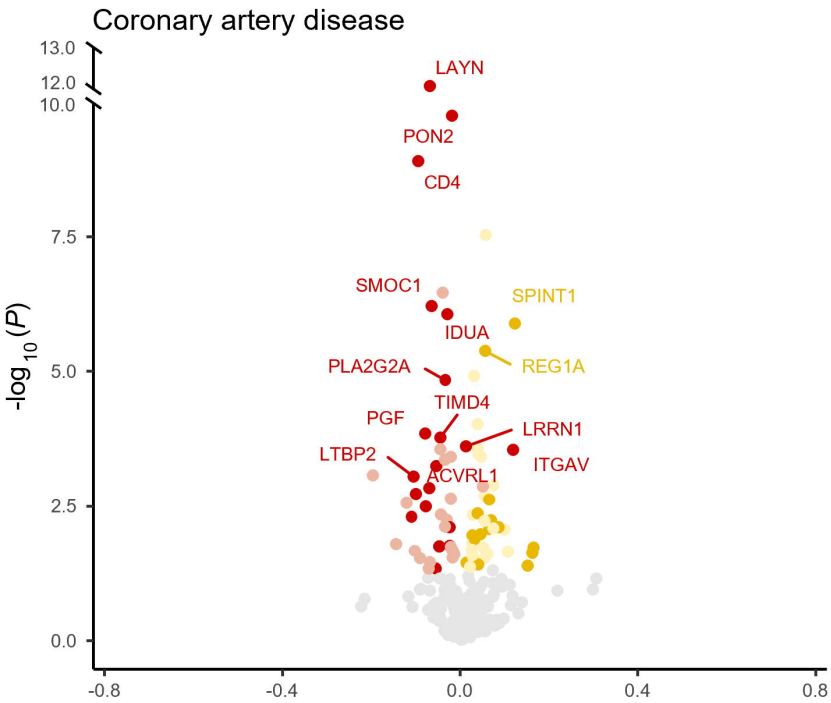
Atrial fibrillation



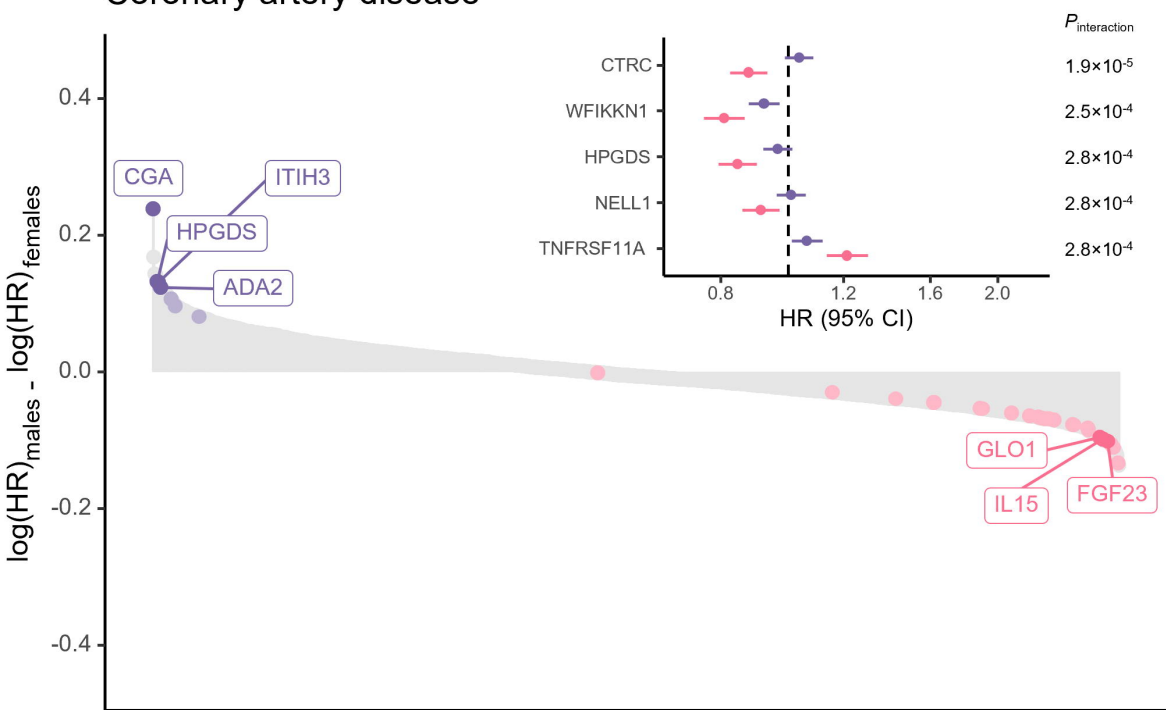
Aortic stenosis



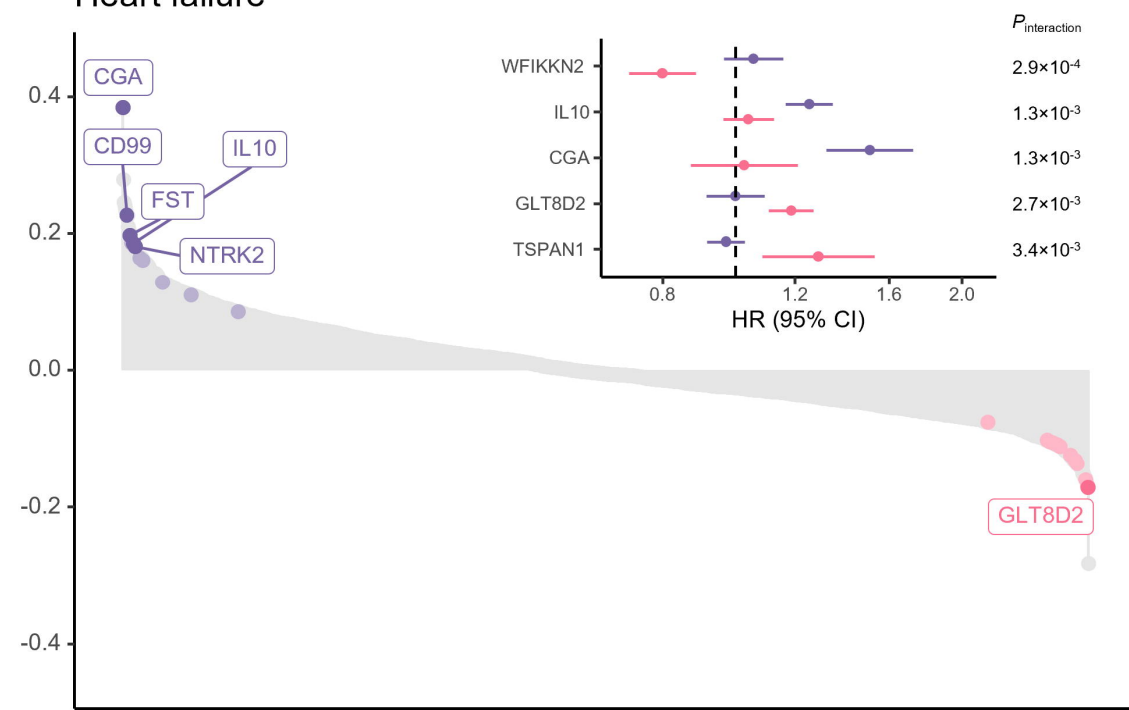
Not Bonferroni-significant Bonferroni-significant for more than one outcome Bonferroni-significant for one outcome



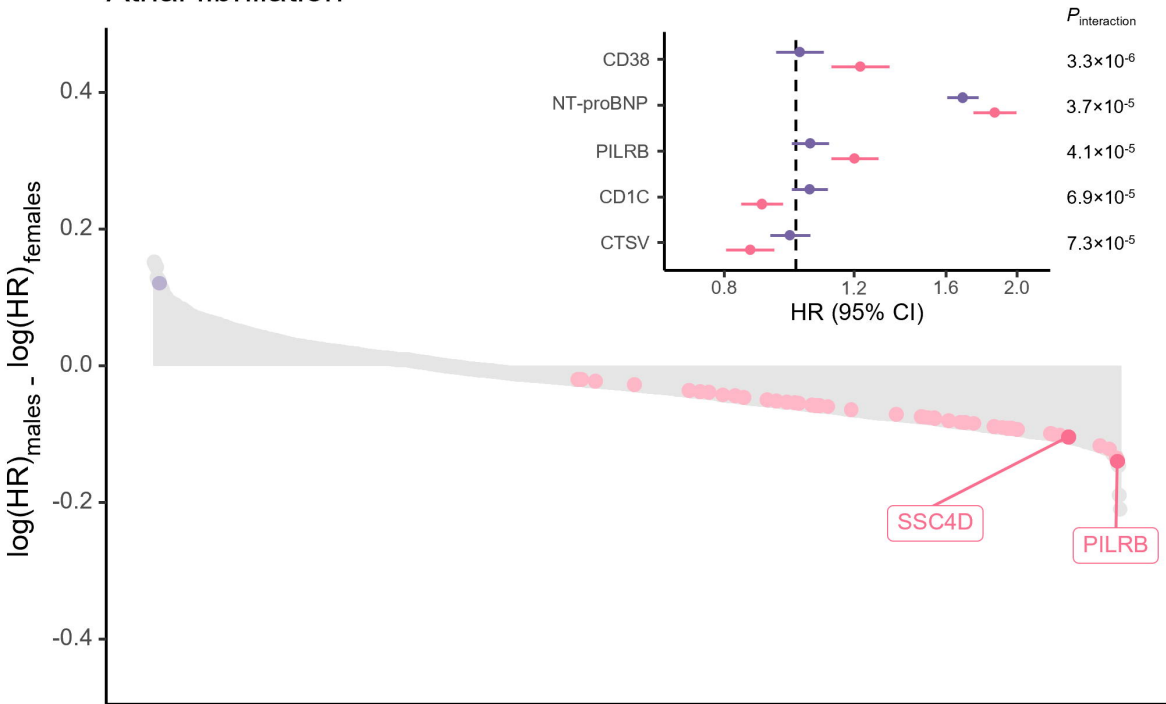
Coronary artery disease



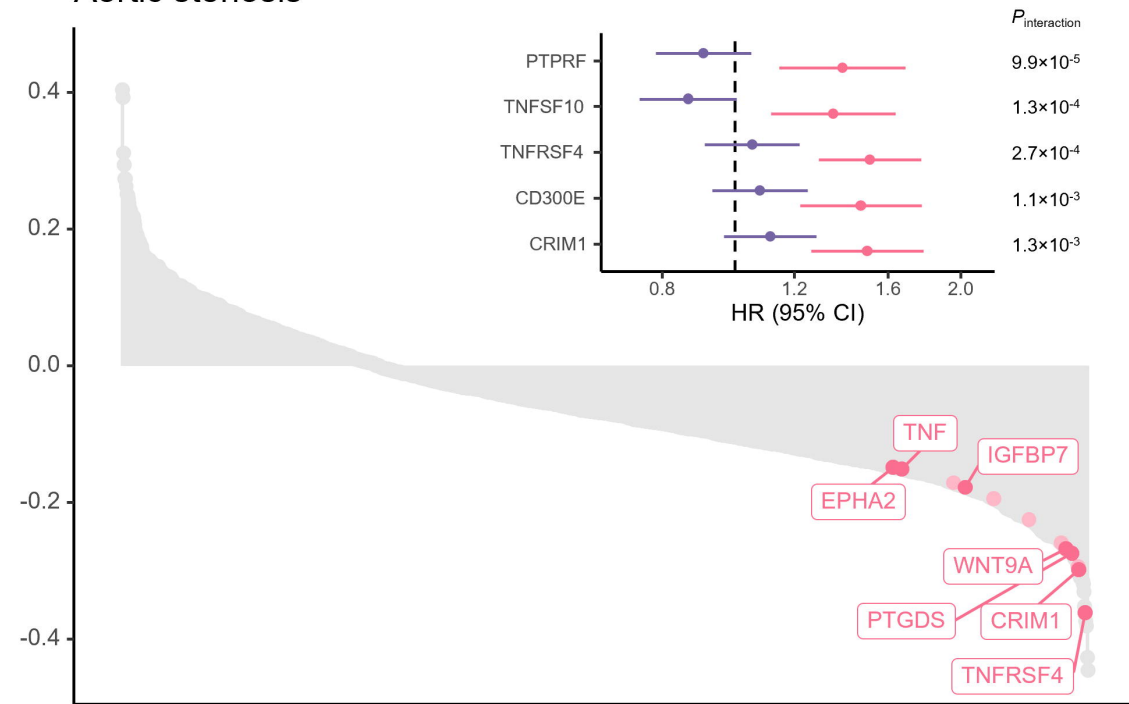
Heart failure



Atrial fibrillation



Aortic stenosis



Biomarkers with sex differences:
● ● $\log(HR)$ higher in female participants ● ● $\log(HR)$ higher in male participants
— — HR (95% CI) in female participants — — HR (95% CI) in male participants

