

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

Received: 13 December 2024

Accepted: 10 June 2026

Published online: 29 June 2026

Cite this article as: Ometto, S., Chatterjee, S., Vergani, A.M. *et al.* Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture. *Nat Commun* (2026). <https://doi.org/10.1038/s41467-026-74575-y>

Sara Ometto, Soumick Chatterjee, Andrea Mario Vergani, Arianna Landini, Sodbo Sharapov, Edoardo Giacomuzzi, Alessia Visconti, Emanuele Bianchi, Federica Santonastaso, Emanuel M. Soda, Francesco Cisternino, Carlo Andrea Pivato, Francesca Ieva, Emanuele Di Angelantonio, Nicola Pirastu & Craig A. Glastonbury

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

Sara Ometto^{*1}, Soumick Chatterjee^{*1,2}, Andrea Mario Vergani^{1,3}, Arianna Landini¹, Sodbo Sharapov¹, Edoardo Giacomuzzi¹, Alessia Visconti^{1†}, Emanuele Bianchi¹, Federica Santonastaso¹, Emanuel M. Soda^{1,3}, Francesco Cisternino^{1,4}, Carlo Andrea Pivato^{5,6}, Francesca Ieva^{1,3}, Emanuele Di Angelantonio¹, Nicola Pirastu¹, and Craig A. Glastonbury^{†1}

¹ Human Technopole, Viale Rita Levi-Montalcini 1, 20157, Milan, Italy

² Faculty of Computer Science, Otto von Guericke University Magdeburg, Universitätsplatz 2, 39106, Magdeburg, Germany

³ Department of Mathematics, Politecnico di Milano, Via Bonardi 9, 20133, Milan, Italy

⁴ European School of Molecular Medicine (SEMM), Milan, Italy

⁵ Department of Biomedical Sciences, Humanitas University, Via Rita Levi Montalcini 4, 20072, Pieve Emanuele, Milan, Italy

⁶ IRCCS Humanitas Research Hospital, Via Manzoni 56, 20089, Rozzano, Milan, Italy

[†] Current address: Centre for Biostatistics, Epidemiology, and Public Health, University of Turin, Regione Gonzole 10, 10043, Orbassano, Turin, Italy

{sara.ometto, soumick.chatterjee, craig.glastonbury}@fht.org

Abstract. Biobank-scale imaging provides an unprecedented opportunity to characterise how thousands of organ phenotypes vary in populations. However, deriving specific phenotypes from imaging data requires time-consuming expert annotation, limiting scalability. In this study, we develop a 3D diffusion autoencoder to derive latent phenotypes from temporally resolved cardiac MRI data of 71,017 UK Biobank participants. These phenotypes are reproducible, heritable ($h^2 = [4 - 18\%]$), and significantly associated with cardiometabolic traits. To establish the genetic basis of such traits, we perform a genome-wide association study, identifying 89 significant common variants ($P < 2.3 \times 10^{-9}$) across 42 loci, including seven novel loci. Extensive multi-trait colocalisation analyses (PP.H₄ > 0.8) link variants across phenotypic scales, from intermediate cardiac traits to cardiac disease endpoints. In conclusion, this study showcases the use of diffusion autoencoding methods as powerful tools for unsupervised phenotyping, genetic discovery and disease risk prediction using cardiac MRI data.

Introduction

Biobank-scale non-invasive medical imaging coupled with advances in unsupervised machine learning enables unbiased phenotyping at population scale. For example, UK Biobank (UKB) cardiac Magnetic Resonance Imaging (CMRI) acquisitions contain rich structural and functional information that remains unexplored. Previous research efforts focused on extracting single quantitative "*image-derived phenotypes*" (IDPs), such as Left Ventricle Ejection Fraction (LVEF) [1, 2]. These traits can be linked with electronic health records and common and rare genetic variation through genome-wide association studies (GWAS) to identify disease associations and genetic determinants. However, each trait requires time-consuming expert annotation and specific model development, limiting scalability and downstream genetic characterisation. Moreover, 3D or temporal MRI acquisitions are extremely information rich and therefore deriving single phenotypes of organ function is reductive for such complex systems.

Recent breakthroughs in machine learning (e.g. diffusion) allow interpretable image phenotyping without any specific annotations. These models learn to compress high-dimensional imaging data into a low-dimensional representation (z) [3], where each dimension of z can be thought of as capturing a specific imaging feature or "*latent phenotype*". We define a "*latent phenotype*" as a data-driven, unobserved feature represented by one dimension of a learnt latent space, capturing complex image patterns without explicit human definition. In contrast, IDPs are predefined interpretable traits extracted from images using supervised or classical methods, often through segmentation, representing a specific anatomical or functional measurement. Whilst such approaches have been applied to 2D natural images [4], static 3D medical images [5] and other high-dimensional data modalities [6, 7], they have not been extended to temporally resolved imaging for phenotype discovery.

Variational AutoEncoders (VAE) have demonstrated that latent representations can capture high-level semantic concepts and have been applied to imaging-based phenotyping for phenotypic discovery and genetic analyses [8, 9, 10, 11].

* These authors contributed equally to this work.

† Lead contact.

However, VAEs are limited by constrained latent capacity and suboptimal reconstruction fidelity, curtailing interpretability. Diffusion autoencoders (DiffAE) [4] address these limitations, combining the flexibility of autoencoders with high fidelity reconstructions. Here, we extend diffusion autoencoders to work with 3D, time-resolved cardiac magnetic resonance imaging, enabling high throughput and interpretable phenotyping in 71,017 UKB participants (**Figure 1A**).

We demonstrate that learnt latent phenotypes recapitulate both known and novel cardiac traits and disease associations, both epidemiologically and genetically (**Figure 1B-C**). Using multi-trait colocalisation, we provide a high resolution map of shared common variants that influence traits across phenotypic scales, from low-level cardiac endophenotypes (e.g. heart rate), risk factors (LDL cholesterol levels), and cardiac disease outcomes (myocardial infarction) (**Figure 1D**) and how such variants are enriched in cardiomyocyte specific cell populations (**Figure 1E**). Finally, as imaging provides a means of unbiased prognostic modelling, we show that polygenic risk scores built from latent factors can pinpoint individuals at substantially elevated risk of cardiac-relevant diseases (up to 26-fold increased cumulative hazard, **Figure 1F**). In summary, we have conducted the largest systematic characterisation of cardiac imaging, intermediate cardiac traits and cardiac diseases to date and demonstrate that unsupervised phenotyping of cardiac MRIs using diffusion processes provides a tractable means for novel genetic discovery, risk prediction and interpretation.

Results

Phenotyping by 3D diffusion autoencoders of cardiac MRIs

To derive unbiased phenotypes in a high-throughput manner, we exploited 4-chamber view Electrocardiogram (ECG)-synchronised temporal cardiac MRI acquisitions in UK Biobank, which provide detailed characterisation of cardiac function. Whilst multiple studies have focused on deriving single phenotypes from such images [12, 13, 14], unsupervised and self-supervised methods can represent complex images as a compressed latent vector, where each dimension corresponds to a specific image feature (here referred to as *latent phenotype*) [15, 9, 10, 11]. We therefore extracted latent phenotypes from all 71,017 UK Biobank individuals with available cardiac MRIs for diagnostic and prognostic risk modelling and downstream genetic characterisation.

After quality control, images were cropped to centre the heart (see Methods). As the data are dynamic 2D MRIs over 50 time points, we extended diffusion autoencoders to 3D (3DDiffAE) by stacking them across time points as a single 3D volume, capturing time as the third dimension and therefore modelling the heart through one cardiac cycle. Compared to Variational Autoencoders (VAEs), this approach decouples high-level semantic latents from low-level stochastic details, thereby overcoming inherent trade-offs between regularisation and expressiveness. This architecture yielded near-perfect reconstruction fidelity (**Figure 2A, Supplementary Figure 1**), achieving a median Structural Similarity Index Measure (SSIM) of 0.96 ± 0.02 , a metric that quantifies how similar the reconstructed image is to that of the original input image, significantly surpassing the VAE benchmark (**Figure 2B**; see **Supplementary Figure 2A-B** for 3DDiffAE scores across multiple model runs). This is a prerequisite for reliable latent-based interpretability and richer phenotypic space generation.

We selected a latent dimension of $d = 128$ based on benchmarking ($d \in \{64, 128, 256\}$) which optimised compression (mean SSIM (0.962), mean PSNR (30.03)), **Supplementary Figure 2C-D**). We further ensured replicability by retaining only those phenotypes consistently present across multiple independent model runs ($R \geq 0.8$), resulting in a final set of 182 robust latent phenotypes. Furthermore, no orthogonality constraint was enforced, and correlation between phenotypes reflects both shared biological signal and methodological consistency, with some latent phenotypes showing moderate to high correlation (**Supplementary Figure 3**). This is expected, as cardiac traits are inherently highly correlated (**Supplementary Figure 4**). Latent phenotypes are indexed by dimension and model run (e.g. Z10_S1 is the 10th latent phenotype, discovered in the first run). Together, our ability to extract unbiased phenotypes at scale and reconstruct cardiac MRIs with high fidelity enables us to characterise cardiac structure and function in whole populations without the need for time-consuming annotations.

Latent phenotypes associate with cardiometabolic traits

To explore the diagnostic and prognostic potential of our cardiac MRI latent phenotypes, we tested their association with diseases recorded in at least 200 participants, using GP clinical event records, hospital inpatient (HI) admissions and self-reported (SR) medical conditions. DiffAE-derived latent phenotypes outperform VAE-derived ones in capturing clinically relevant information (**Figure 2C-D, Supplementary Figure 5A-B, Supplementary**

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

Data 1). Taking forward the 182 DiffAE latent phenotypes replicated across 5 model runs, we found 334 significant associations across diseases ($P\text{-value} \leq 1.96 \times 10^{-5}$), with 77% of all latent phenotypes associated with at least one condition (**Figure 3A**). The most significant association was between Z106_S4 and atrial fibrillation ($\beta = 0.29$, $P\text{-value} = 2 \times 10^{-51}$), a widespread cardiac arrhythmia which was also the most frequently associated disease overall. Type 2 diabetes, a common cardiometabolic condition known to increase the risk of cardiovascular disease (CVD), showed 55 significant latent phenotype associations, followed by heart failure (27) and myocardial infarction (22). Importantly, several latent phenotypes exhibited cardiac disease specificity; for example, Z49_S1 was exclusively associated with atherosclerotic conditions: coronary heart disease ($\beta = -0.21$, $P\text{-value} = 4.6 \times 10^{-15}$), myocardial infarction ($\beta = -0.27$, $P\text{-value} = 6.21 \times 10^{-15}$) and angina pectoris ($\beta = -0.20$, $P\text{-value} = 7.7 \times 10^{-10}$), the latter two representing the strongest associations across all latent phenotypes (**Supplementary Data 2** contains the full list of associations).

We next assessed how well our latent phenotypes simultaneously captured specific heart and aortic endophenotypes [16]. Remarkably, despite the absence of supervision, all 182 latent phenotypes demonstrated associations with at least one trait (**Figure 3B, Supplementary Data 3**), with Z107_S1 associated with all the 28 cardiac and aortic measures. Some latent phenotypes captured both imaging traits and cardiac diseases. For example, Z82_S1 was associated with 25 MRI-derived measurements, including Left Atrium (LA) maximum volume ($\beta = -0.37$, $P\text{-value} = 1.86 \times 10^{-354}$) and Left Ventricle (LV) stroke volume ($\beta = -0.43$, $P\text{-value} = 1.47 \times 10^{-339}$) alongside 16 additional traits ($P\text{-values} \leq \times 10^{-100}$). It also showed the highest number of disease associations (8, together with Z93_S1), namely type 2 diabetes ($\beta = 0.31$, $P\text{-value} = 1.93 \times 10^{-26}$), myocardial infarction ($\beta = -0.21$, $P\text{-value} = 6.97 \times 10^{-10}$), coronary heart disease ($\beta = -0.16$, $P\text{-value} = 2.45 \times 10^{-9}$), atrial fibrillation ($\beta = -0.16$, $P\text{-value} = 5.65 \times 10^{-9}$), hypertension ($\beta = -0.07$, $P\text{-value} = 6.11 \times 10^{-7}$), mitral problems ($\beta = -0.25$, $P\text{-value} = 4.09 \times 10^{-6}$), high cholesterol ($\beta = 0.07$, $P\text{-value} = 1.63 \times 10^{-5}$) and angina pectoris ($\beta = -0.14$, $P\text{-value} = 1.65 \times 10^{-5}$).

Sensitivity analyses excluding waist circumference and standing height from the covariates preserved the association patterns, with expected increases in significance for adiposity-related traits (**Supplementary Figure 5C-D**). Comparison with existing cardiac MRI-derived endophenotypes in terms of links to cardiometabolic diseases showed broadly comparable associations (**Supplementary Data 4**). Atrial fibrillation remained the strongest association, with some latent phenotypes (Z106_S4, Z32_S1, Z65_S1) having larger effect sizes than right atrium endophenotypes (min/max volume, ejection fraction) derived from the same long-axis 4 chamber view MRI dataset (**Supplementary Figure 5E**). As expected, left atrium traits (derived from combined long-axis 2ch and 4ch views) showed the strongest association with atrial fibrillation, while short-axis left ventricular myocardial measures were most associated with hypertension (**Supplementary Figure 5F-G**). In summary, these latent phenotypes capture a continuum of cardiovascular biology, linking low-level heart endophenotypes and disease outcomes.

Latent phenotype interpretation using latent manipulation

As 3DDiffAE is generative, we can modify the magnitude of each latent phenotype to understand what it specifically captures in a cardiac MRI. To do so, we trained linear and logistic Lasso regression models to predict cardiac traits and diseases from learnt latent phenotypes. We first confirmed that $d = 128$ maximised predictive performance relative to alternative dimensions (**Supplementary Figure 6**). Despite pruning the latent space by 3.5-fold (182/640), we verified that the selected latent phenotypes retained most of the total variance explained ($R^2 = 95\%$, **Supplementary Figure 7**). These selected latent phenotypes were predictive of several time-dynamic traits, including left ventricle end diastolic volume (mean cross-validation $R^2 = 0.58$), demonstrating the model captures temporal aspects of cardiac function (**Supplementary Figure 8, Supplementary Data 5**). For disease prediction, latent phenotypes achieved a performance comparable to traditional derived cardiac measures (mean cross-validation $AUC = [0.63 - 0.77]$ latent phenotypes vs $[0.61 - 0.74]$ cardiac measures; **Supplementary Datas 6, 7**). Furthermore, adding latent phenotypes to the predictive model based on cardiac measures consistently improved prediction across all diseases (max improvement: 6.3% AUC, mean improvement: $2.5 \pm 1.8\%$, $P\text{-value} < 0.05$). When restricting to the 11 cardiac IDPs derived from 4-chamber view, for a fairer comparison, improvements were larger (max: 15.5% AUC, mean: $9.6 \pm 2.7\%$ AUC, $P\text{-value} < 0.05$; **Supplementary Figure 9, Supplementary Data 8**). These results confirm that latent phenotypes capture complementary information about cardiac structure and function beyond previously derived IDPs.

To investigate biological relevance further, we identified latent phenotypes predictive of a given trait or disease, and manipulated the corresponding latent dimensions to visualise what they capture in the MRI (see Methods). **Figure 3C** shows manipulation of latent phenotypes selected by the Lasso model to predict LV end diastolic volume in two subjects. Strikingly, increasing the manipulation magnitude m $[-1.5, 1.5]$ produced a progressive increase in LV end diastolic volume, with a visibly larger heart in the reconstructed MRIs (additional examples: **Supplementary**

Figure 10, Supplementary Videos 1-5). Across five subjects (**Supplementary Figure 10**), mean LV end diastolic volume decreased from 138.95 ± 29.65 to 127.82 ± 23.87 with $m = -0.75$, reaching 118.00 ± 23.65 at $m = -1.5$, while increasing to 145.53 ± 31.96 when $m = 0.75$, and to 167.13 ± 23.96 at $m = 1.5$. Similarly, **Figure 3D** illustrates manipulation of latent phenotypes predictive of LV mean myocardial wall thickness, where increasing m resulted in visibly thicker myocardium (additional examples: **Supplementary Figure 10, Supplementary Videos 6-10**). Again, quantifying the trait changes showed myocardial wall thickness decreasing by $-4.09\% \pm 7.84\%$ with $m = -1.5$, and increasing by $15.59\% \pm 9.16\%$ when $m = 1.5$, with smaller but directionally consistent effects at $m = \pm 0.75$ ($-2.30\% \pm 6.19\%$ and $0.30\% \pm 10.96\%$ respectively). Together, these results demonstrate that latent phenotypes can be tied back to visually interpretable features in the MRI.

Latent phenotypes are associated with novel genetic variants

Having established that latent phenotypes capture multivariate aspects of the heart's structure and function, we next mapped their genetic architecture leveraging UKB genotyping data, assessing shared common variants across biological scales, from expression quantitative trait loci (eQTLs) to disease outcomes. Latent phenotype heritability ranged from 4% to 18% (mean h^2 of 9.6%, SD = 0.9). The 182 prioritised latent phenotypes (Pearson's $R \geq 0.8$) selected for robustness across model runs exhibited significantly higher heritability than the remaining 458 (one-sided Mann-Whitney U test P-value = 6.5×10^{-3} , **Supplementary Figure 11A**). Compared with MRI-derived cardiac endophenotypes, latent phenotypes showed comparable heritability ranges, considering right atrium measures and LV longitudinal strain that are also derived from the long-axis 4-chamber view (mean $h^2 = 12.5\%$, SD = 0.2). This persisted in the extended comparison to IDPs from both two- and four-chamber long-axis views (mean $h^2 = 13\%$, SD = 0.3, **Supplementary Figure 11B**).

To identify associated common genetic variants, we performed genome-wide association studies (GWAS) on 182 latent phenotypes in 47,740 Caucasian individuals following quality control (**Supplementary Figure 12** and Methods). **Supplementary Data 9** contains baseline characteristics and covariates of the discovery cohort. At a study-wide significance threshold of 2.27×10^{-9} [17], we detected 95 significant associations. Additionally, 366 single nucleotide polymorphisms (SNPs) reached genome-wide significance (5×10^{-8}). Conditional analysis refined these to 89 conditionally independent trait-wise associations (**Supplementary Data 10**) mapping to 44 unique lead SNPs within 42 genomic loci (**Supplementary Data 11**, see Methods). The Manhattan plot in **Figure 4A** summarises these 182 GWAS results, highlighting significant loci annotated with their nearest or overlapping genes. Of the 44 lead SNPs, 7 were previously unreported in the GWAS Catalog [18]. To categorise significant loci, we assigned a label based on the most prevalent trait association category reported in the GWAS Catalog. The dotplot in **Figure 4B** reveals that most variants have been previously associated with cardiac traits and diseases, followed by anthropometric measures, biomarkers and other cardiometabolic risk factors. Associations with lung and bone phenotypes were also observed, consistent with the inclusion of surrounding anatomical structures in the cropped cardiac MRI. Four lead variants overlapped with findings from a smaller GWAS on latent embeddings learnt from a cross-modal ECG and MRI autoencoder [19] (labelled "MRI latents" in **Figure 4B**). Of the 33 reported variants in their prior GWAS, 26 common SNPs were testable in our data (MAF ≥ 0.01), of which 12 reached suggestive significance ($P < 1 \times 10^{-4}$) in our analysis. Despite methodological differences, this overlap supports the robustness of unsupervised latent representations, while the additional variants identified here highlight the increased power of our larger dataset and analytical framework.

Examining conditionally independent associations, Z82_S1 (together with Z86_S3) showed the strongest association signal, being associated with 7 of the 89 SNPs across 7 loci, including the novel *CRIL* locus, as well as *CCDC141*, *SLC35F1*, *HSF2*, *MYH6* and *KIAA1755*. Lead SNPs within these loci have previously been linked to cardiac phenotypes. Interestingly, the *SOX5* locus was shared between Z49_S1 and Z82_S1, both of which were also associated with multiple cardiac diseases and MRI-derived traits. The lead variant rs4963772 (minimum P-value = 3.55×10^{-14}) has previously been associated with blood pressure, ECG traits and atrial fibrillation.

Annotation of the nearest genes highlighted several well-characterised cardiac genes. For example, Myosin heavy chain 6 (*MYH6*), encoding a key sarcomeric protein involved in cardiac contraction, has been associated with atrial fibrillation (rs422068, P-value = 3.5×10^{-12} [20]). *GJA1* (Gap Junction Alpha 1 or Connexin 43), mediating electrical coupling between cardiomyocytes, has been previously linked to pulse rate and resting heart rate (rs9388001, P-value = 4.4×10^{-110} [20]). We also identified a novel association near *LCORL* (ligand-dependent nuclear receptor corepressor-like): whilst the intronic variant (4:17956213 TG>T) has not been linked to cardiac traits in humans, *LCORL*-knockout mice show decreased heart rate and prolonged RR interval [21], warranting further investigation into *LCORL*'s precise role.

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

To test for sex specific effects, we performed SNP-by-sex interaction GWAS across the 182 selected latent phenotypes. At the study-wide significance threshold (2.27×10^{-9}), one locus on chromosome 11 reached significance, with the lead variant, rs72925197 (P-value = 1.34×10^{-11}), being an intronic SNP in *NRXN2* (**Supplementary Figure 13**). Although not previously linked to cardiac or sex-specific traits, this SNP is an eQTL for *VEGFB* in heart and adipose tissues [22] and has been implicated in cardiovascular development [23]. Nine additional loci demonstrated suggestive associations at 5×10^{-8} (**Supplementary Data 12**).

Replication in a held out UKB cohort ($N = 12,396$, with approximately two-thirds non-Caucasian; **Supplementary Data 13**) confirmed that associations with MRI-derived traits were preserved (**Supplementary Figure 14, Supplementary Data 14**). Replication GWAS in this ancestrally diverse cohort showed that of the 89 conditionally independent significant trait-SNP associations, 84 had concordant effect directions (**Supplementary Figure 15, Supplementary Data 15**), and 51 replicated at nominal significance (P-value < 0.05).

Finally, we assessed shared genetic architecture using multivariable LD score regression [24] (**Supplementary Figure 16**). Latent phenotype Z82_S1 exhibited the strongest positive genetic correlation with resting heart rate ($r = 0.78 \pm 0.08, p = 8.7 \times 10^{-25}$) and pulse rate ($r = 0.83 \pm 0.09, p = 8 \times 10^{-21}$), alongside negative correlations with several MRI-derived measurements. Consistent with this, Z82_S1 was also strongly associated with multiple cardiac diseases, including myocardial infarction (P-value = 7×10^{-10}) and atrial fibrillation (P-value = 5.6×10^{-9}). As expected, no significant genetic correlation was observed for the negative control (hair colour, mean absolute correlation = 0.054, minimum P-value = 0.135; **Supplementary Figure 17**). Overall, we identify biologically meaningful and replicable genetic associations with derived latent phenotypes.

Latent phenotype genes are expressed in cardiac cells

To prioritise the most likely target genes at associated loci, we combined positional and eQTL mapping in cardiac tissues using FUMA [25], linking 82 protein-coding genes to the 44 lead SNPs. *Enrichr* over-representation analysis [26] showed significant enrichment for cardiac-related traits including heart rate and cardiomyopathies, as well as cardiomyocytes (**Figure 5A, Supplementary Data 16**). Gene-set enrichment analysis using MAGMA [27] further highlighted enrichment in processes related to heart morphogenesis and atrial septation defects (**Supplementary Figure 18, Supplementary Data 17**). To assess tissue specificity, MAGMA gene-property analysis using GTEx v8 expression data revealed significant enrichment in left ventricle and right atrial appendage, as well as colon sigmoid, consistent with its smooth muscle content (**Figure 5B, Supplementary Data 18**). Chromatin enrichment analysis (CHEERS) [28] demonstrated that latent phenotype-associated variants are enriched in regulatory regions active in fibroblasts and atrial cardiomyocytes (**Supplementary Figure 19**), supporting cardiac cell-type specific regulatory activity.

To link genetic associations to gene expression in the heart, we performed a transcriptome-wide association study (TWAS) using FUSION. We identified significant associations for 82 genes in right atrial appendage (AA) and 81 in left ventricle (LV), with 14 genes showing concordant effects in both tissues (**Supplementary Figure 20A**). Among the strongest signals was *BCKDHA* (AA: $Z = 7.28, P = 1.86 \times 10^{-9}$), encoding a key enzyme in branched-chain amino acid (BCAA) degradation, previously linked to heart failure [29, 30]. Additional prioritised genes included *WWP2* (AA: $Z = -7.16, P = 4.79 \times 10^{-9}$), an E3 ubiquitin protein ligase known to regulate cardiac fibrosis [31, 32], and *SCN5A* (AA: $Z = -6.12, P = 5.39 \times 10^{-6}$), a cardiac sodium channel in which mutations are extensively linked to cardiomyopathies [33]. Overall, 27 TWAS significant genes overlapped with FUMA-prioritised genes (**Figure 5C**), with substantial cross tissue sharing (π_1 , **Supplementary Figure 20B, Supplementary Data 19**).

Finally, using single-nucleus RNA sequencing data (snRNA-seq) from human heart tissue [34], we found that the 82 MAGMA-prioritised genes showed significantly higher expression in atrial and ventricular cardiomyocytes (**Figure 6A-B**). Canonical markers were identified: *MYH6* for atrial (ACMs) and *MYH7* for ventricular (VCMs) cardiomyocytes, respectively; whilst *TTN* was strongly expressed in both (**Supplementary Data 20**). Several genes mapped to specific cardiomyocyte populations (**Figure 6C**), including *MLF1* (Log fold change (LFC) = 2.68 in ACMs; 3.0 in VCMs) and *CCDC141* (LFC = 4.08 in VCMs vs LFC = 1.38 in ACMs). In summary, we demonstrate that genes linked to latent phenotype-associated variants are enriched in and functionally relevant to specific cardiac cell types, particularly cardiomyocytes.

Latent phenotypes demonstrate extensive colocalisation

Given that many lead variants have been previously linked to cardiometabolic traits (e.g. heart rate, ECG measures, atrial fibrillation), we assessed whether shared genetic variants underlie both latent phenotypes and these traits. For

instance, rs142556838 (MAF = 0.09, Z82_S1 association $P = 3.26 \times 10^{-12}$), an intronic variant in *CCDC141* (nearest gene *TTN*), has been associated with diastolic blood pressure [35], pulse rate, heart rate and ECG measures. We establish here that the intron with rs142556838 overlaps a DNase hypersensitivity peak in cardiac right atrium tissue and H3K4me1 and H3K27ac CHIP-seq peaks in heart left ventricle, ascending aorta and right atrium auricular tissue [36], suggesting it is part of an active enhancer in right atrium auricular tissue and implicating it in the regulatory activity of the neighbouring gene *TTN* [36]. Similarly, rs72801474 (MAF = 0.09, Z1_S5 association $P = 1.03 \times 10^{-9}$) near *HSPA4* is a pleiotropic variant associated with triglyceride levels and hypertension, and lies within a candidate *cis*-regulatory element (cCRE) [36]. To systematically characterise shared genetic architecture, we performed multi-trait colocalisation across 54 existing GWAS and eQTL datasets spanning 12 tissues (**Supplementary Data 21**). In total, 34 loci showed significant colocalisation ($PP.H_4 > 0.8$; **Figure 7A-B**, **Supplementary Data 22**). Of these, 19 loci colocalised with at least one MRI-derived cardiac phenotype, ECG trait or blood pressure measure, 10 with cardiac eQTLs (heart and/or artery), and 7 with cardiac diseases. Notably, only 6 loci colocalised with MRI-derived traits despite testing 27 cardiac and aortic IDPs, of which just 9 exhibited any significant colocalisation, suggesting that latent phenotypes capture complementary biological signals not detected by standard MRI-derived measures.

Nine loci were cardiac-specific, colocalising exclusively with cardiac eQTLs or cardiac GWAS traits. For example, at the *KIAA1755* locus, Z82_S1 colocalised with ventricular stroke volume, pulse pressure, PR interval and resting heart rate (**Figure 7B**). At the *MYH6* locus, the variant is shared between three latent phenotypes, multiple MRI-derived traits (including descending aorta maximum and minimum area, ventricular stroke volume, right ventricle ejection fraction), pulse pressure, and atrial fibrillation (**Supplementary Figure 21**).

We also identified three novel loci from our latent phenotype GWAS that colocalised with known cardiac traits (**Supplementary Figure 22**). For example, rs11118288 (chromosome 1) colocalised with a pulse rate locus, while the intronic deletion in *LCORL* (4:17956213 TG>T) colocalised with hypertension, pulse pressure and bone mineral density, and acts as an eQTL for *DCAF16* in coronary artery, blood and skeletal muscle. The chromosome 6 locus colocalised with eQTLs for *CENPW* in heart atrial appendage and subcutaneous adipose tissue.

Approximately one-third of lead variants (14/44) colocalised across multiple trait categories, including cardiac diseases, cardiometabolic risk factors and cardiovascular traits (**Supplementary Figure 23**). For instance, at the *HSPA4* locus, Z1_S5 shares rs72801474 with cardiovascular disease (CVD), hypertension, pulse pressure, triglycerides and HDL cholesterol (**Figure 7C**).

These results demonstrate the power of MRI-derived latent phenotypes to capture multivariate aspects of cardiac function and to uncover relevant genetic signals that bridge different biological scales.

Burden tests of latent phenotypes implicate cardiac genes

To assess the contribution of large-effect rare variation, we conducted a single-variant exome-wide association study (EWAS) in 46,034 individuals from the discovery cohort, conditioning on significant common variants from GWAS. Five associations reached exome-wide significance, four with MAF ≤ 0.01 (**Supplementary Data 23**). However, due to the low minor allele count, further studies are warranted to support these findings.

To increase power, we conducted gene-based burden tests in 46,134 individuals using the SKATO-ACAT as implemented in REGENIE, applying three variant masks including Loss of Function (LoF) and missense variants. This identified 17 significant genes at the exome-wide threshold (4.7×10^{-07} ; **Figure 8**; **Supplementary Data 24**), including four not previously reported in Genebase [37] (*SMCP*, *PRUNE1*, *FAM218A* & *LINC7*). Several genes have prior cardiometabolic relevance: pLoF mutations in *JAGN1* have been associated with stroke and vascular or heart problems (Genebase). Mouse knockouts of *ATP5L*, *PHC3* and *BTBD8* show an enlarged heart and/or abnormal heart morphology [21], and *ATP5L* knockout mice exhibit increased cardiomyocyte necrosis and infarct size [38]. *ATP5L*, encoding subunit G of mitochondrial ATP synthase, is also specifically expressed in cardiac tissues (GTEx [22]). Of the four novel genes, *PRUNE1*, a member of the DHH phosphoesterase superfamily, has been linked to hypertrophic cardiomyopathy by the Human Phenotype Ontology (HPO) project [39]. Classical burden tests with the same masks replicated all significant genes (**Supplementary Figure 24**, **Supplementary Data 25**). Because several of the reported variants have low minor allele counts, further investigation of their role in cardiac structure and function requires replication in external imaging and genetics cohorts of comparable scale.

PRS of latent phenotypes predict cardiometabolic diseases

We assessed whether latent phenotype Polygenic Risk Scores (PRS) could stratify participants into distinct risk groups for cardiac diseases (**Supplementary Data 26**). After constructing a PRS for each latent phenotype, we

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

assessed whether individuals in the top percentiles of the most predictive PRS for each disease had higher disease incidence. Across all diseases considered, we observed that the mean cohort disease prevalence was 14.7%, with a higher prevalence in males (18.6%) compared to females (11.5%). Strikingly, the average increase in prevalence when comparing the *bottom* 5% to the *top* 5% PRS percentiles for atherosclerotic conditions was substantial, at 26.5% in the combined population, 25.1% in males, and 26.6% in females (**Figure 9A**). Similarly, **Figure 9D** shows the same for metabolic syndrome, whilst all other diseases are presented in **Supplementary Figures 25, 26**.

Given there are significant differences in disease prevalence across the PRS distribution, we investigated whether there was a difference in both the cumulative diagnoses and hazard rates up to ten years after the collection of baseline characteristics. To do so, we fit disease-specific Generalised Linear Models (GLMs) using each latent PRS as a predictor (one model for each latent PRS). **Figure 9B** and **C** respectively present the cumulative number of new diagnoses and the cumulative hazard rate for atherosclerotic diseases, and **Figure 9E** and **F** respectively for metabolic syndrome, whilst the results for all other diseases are presented in **Supplementary Figures 25, 26**. Across all diseases, myocardial infarction showed the largest increase (20.6%) in cumulative diagnoses when comparing the high risk group to the low risk group.

Focusing on atherosclerotic diseases, the average increase across time points was 16.1%, with a maximum increase of 26%. These results indicate that individuals in the *Top risk* group experienced a significantly higher cumulative incidence of new diagnoses compared to those in the *Low risk* group. For atherosclerotic diseases, the cumulative hazard rate in the top-risk group was 6.3-fold higher at two years and 4.8-fold at ten years relative to the low-risk group.

To assess combined predictive power, we aggregated individual latent PRS using a LASSO model to create a multi-latent PRS, and compared its performance against each single PRS and a baseline of age, sex, BMI and smoking status. We observed modest but statistically significant improvements ($P < 0.05$) in AUC for both diagnosis and prognosis over the baseline model (box plots in **Supplementary Figure 27A–B**; full results in **Supplementary Datas 27–29**). Hazard ratios per standard deviation are reported in **Supplementary Data 30**.

We next contextualised performance against established disease-specific PRS [40], comparing our multi-PRS with standard UKB PRS for two cardiac conditions (atrial fibrillation and cardiovascular disease) and two cardiometabolic conditions (hypertension and type 2 diabetes). Unlike standard approaches requiring disease-specific polygenic models, our framework uses a single set of latent PRS across multiple diseases. Whilst the method did not outperform these specialised benchmarks, it achieved comparable performance across all four conditions (**Supplementary Figure 28, Supplementary Data 31**). We further compared our latent multi-PRS against a multi-PRS derived from cardiac IDPs; predictive performance was virtually indistinguishable across all four conditions (**Supplementary Figure 29, Supplementary Data 32**).

Discussion

In this study, we applied a 3D diffusion autoencoder (DiffAE) to cardiac MRI data from 71,017 UK Biobank participants to derive latent phenotypes representing cardiac structure and function. Our approach enabled the extraction of latent cardiac phenotypes at the population level without manual annotation or segmentation, capturing a broad representation of cardiac morphology and dynamics. By integrating these latent phenotypes with clinical and genetic data, we identified significant associations with cardiometabolic traits and diseases, including atrial fibrillation and myocardial infarction, and demonstrated that latent phenotypes are predictive of cardiac traits such as left ventricular volume and myocardial mass. Through GWAS, we uncovered 89 significant common variant associations across 42 loci, including seven novel loci not previously associated with cardiac phenotypes. We thoroughly characterised these variants and demonstrated their biological relevance through enrichment analyses and TWAS. Whilst recent work has utilised multivariate approaches (e.g. MANOVA) to capture aggregate genetic signal, such methods often require the construction of linear probes to decipher phenotypic associations and typically lack simple effect directionality [41, 42]. We therefore prioritised a univariate framework to provide granular interpretability of risk versus protective alleles, and to generate standard summary statistics essential for integration into downstream analyses and direct comparison against published studies (i.e. colocalisation).

One of the key strengths of our approach is the ability to interpret latent phenotypes through latent space manipulation and visualisation of image reconstructions. This enables direct interpretation of how specific latent phenotypes influence cardiac structure and function. By manipulating latent factors associated with left ventricular end-diastolic volume, we demonstrated visible increases and decreases in left ventricular size in reconstructed images. Similarly, manipulation of factors related to myocardial wall thickness produced visible changes in wall thickness, relevant to

conditions such as hypertrophy. However, we emphasise that these manipulations are intended as qualitative explorations; no quantitative analysis was performed to assess whether such latent traversals remain within high-density regions of the learnt data manifold (i.e. avoiding potential "latent holes").

The latent polygenic risk scores demonstrated predictive power for a range of cardiometabolic diseases. Individuals in the top risk quantiles had a significantly higher cumulative hazard for diseases such as metabolic syndrome and atherosclerotic conditions, with up to a 26-fold increase compared to those in the low-risk quantiles at the 10-year mark. We interpret these large effect sizes with caution: rather than implying immediate clinical applicability, these results primarily demonstrate the genetic relevance and biological validity of our imaging-derived latent phenotypes.

We acknowledge that our analyses were restricted to the Caucasian population; we adopted this design to maximise statistical power and minimise confounding from population stratification. Although our replication cohort included individuals of diverse ancestries, the sample sizes for non-European subsets were insufficient to support robust independent analysis or cross-ancestry PRS validation. Given the known challenges regarding the transferability of polygenic scores across populations, this remains a limitation of the current work.

Whilst the framework presented here was applied to cardiac MRI, it is generalisable to other imaging modalities. Applying this approach to brain or abdominal MRI acquisitions would enable characterisation of multi-organ phenotypes and their shared genetic underpinnings, a line of work we are actively pursuing. External validation in other large-scale imaging cohorts would further assess generalisability across populations and imaging protocols. Additionally, integrating longitudinal imaging data could enable the study of temporal changes in cardiac structure and function, potentially identifying early imaging biomarkers of disease progression before clinical symptoms arise.

In conclusion, our study demonstrates the utility of diffusion-based autoencoding models for phenotyping and genetic discovery. By capturing complex imaging features without manual annotation, our approach provides a scalable means to characterise the genetic architecture and pervasive pleiotropy of cardiac-specific genes, traits and diseases. The identification of novel genetic loci, coupled with the ability to visualise and interpret latent phenotypes, underscores the potential of this methodology for advancing understanding of cardiac structure and function in relation to disease.

Methods

Study participants

The UK Biobank is a prospective cohort study including over 500,000 individuals aged between 40 and 70, living across the United Kingdom at the time of recruitment (2006-2010) [43, 44]. The UK Biobank received ethical approval from the UK Biobank Research Ethics Committee (reference number: 11/NW/0382).

Cardiac MRI

Cardiac magnetic resonance imaging (CMR) was performed at 1.5 Tesla using a MAGNETOM Aera scanner by Siemens Healthcare, Erlangen, Germany. Long axis CMRs (UK Biobank data-field ID: 20208) were available for 71,017 subjects with ongoing consent (release: October 2023). These images were acquired in sagittal views using the TRUFI (True Fast Imaging with Steady-State Free Precession) sequence. Full details regarding the CMR protocol and acquisition can be found elsewhere [45, 46]. The dataset includes 2D dynamic images capturing 50 distinct cardiac phases throughout one cardiac cycle per subject, obtained via retrospective electrocardiogram (ECG) gating. The Long-axis CMRs were measured in three directions: the two-chamber (2Ch), three-chamber (3Ch), and four-chamber (4Ch) views, further categorised into three planes (transverse, coronal, and sagittal). This study focuses specifically on the four-chamber view, as its angle provides visibility of both the atria and ventricles. For consistency across subjects, only the transverse plane, representing the most common orientation, was used in the analysis. The images were cropped at the heart using a segmentation mask as guide, with a final dimension of 128 x 128 x 50 time steps. MRIs were normalised by dividing the intensity by the maximum value for that subject. 8,455 subjects were discarded prior to DiffAE training owing to shape mismatch during cropping, absence of the 4Ch transversal plane, or fewer than 50 time steps. The majority of participants were imaged only once, either at instance 2 (UKB encoding of the first imaging visit) or at instance 3 (repeated imaging visit). For the subjects having two repeated acquisitions, only the MRIs from the first visit were kept. Images from 62,562 participants were used for the model training.

To detect and remove low-quality images (in terms of image acquisition) post DiffAE, we used a pre-trained ResNet-18 model from Torchvision (trained on ImageNet) to extract features from each slice. We then applied principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) to generate a final two-dimensional embedding for each image, with outliers identified by applying k-nearest neighbours (KNN) to remove

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

the top 5% with the largest distances to their fifth nearest neighbour, resulting in the removal of 71 images. The remaining 62,491 subjects were then divided into 48,427 and 14,064 discovery and replication latent-phenotyped cohorts, respectively.

Regarding UK Biobank’s phenotypic data, subjects with missing values were removed. Then, those exhibiting outlier values for a certain phenotype were excluded utilising the Elliptic Envelope algorithm [47], which operates under the assumption that the outlier-free data follows a Gaussian distribution and ascertains the deviation of a data point from the central tendency using the Mahalanobis distance. The algorithm was applied with a contamination factor of 0.01 — i.e., it assumes that 1% of the data points are anomalies and adjusts the decision function accordingly to accurately identify outlier boundaries, and with a support fraction of 0.7, indicating that a randomly selected 70% of the data points would be used to estimate the covariance matrix.

Considering only the subjects from the latent-phenotyped cohort that also had phenotypic data available in the UKB, the final discovery and replication cohorts for genetic analyses resulted in $N_{disc} = 47854$ (**Supplementary Data 9**) and $N_{replic} = 12396$ (**Supplementary Data 13**) subjects, respectively. The sample filtering process is illustrated in **Supplementary Figure 7**. The discovery cohort includes only Caucasian subjects determined by genetic ancestry based on a PCA of the genotypes (UK Biobank data-field ID: 22006, as described in [48]), which were released before June 2023. The replication cohort includes subjects of all ancestries, regardless of their release date ($N=8319$) and Caucasian subjects released between June 2023 and October 2023 ($N=4077$). The complete ethnic breakdown (UK Biobank data-field ID: 21000) is presented in **Supplementary Data 33**.

Inferring latent phenotypes

Given the CMR data acquired at different time points, let $\mathbf{X}_t \in \mathbb{R}^{m \times n}$ represent the MRI slice at time point t , where $m \times n$ is the dimensionality of the image, and a set of such MRI slices over time with T time points can be represented as $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T\}$. The objective is to learn a latent representation $\mathbf{Z} \in \mathbb{R}^d$ (with d being the dimensionality of the latent space) such that $\mathbf{Z}$ captures the underlying structural information and temporal dynamics across all time points. This can be formulated as $\mathbf{Z} = f(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T; \theta_{2D})$, where the function $f(\cdot)$, with parameters θ_{2D} , maps the set of MRI slices over time to the latent space $\mathbf{Z}$. We could represent the MRI slices over time as a 3D volume by stacking the slices along with temporal axis as $\mathbf{V} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T] \in \mathbb{R}^{m \times n \times T}$, and re-formulate the task as $\mathbf{Z} = f(\mathbf{V}; \theta_{3D})$. The latent representation can be expressed as $\mathbf{Z} = [z_1, z_2, \dots, z_d]^\top$, where each $z_i \in \mathbb{R}$ is a latent factor.

$\mathbf{Z}$, or any of its subsets, may be used for downstream tasks such as disease diagnosis prediction. More importantly, each z_i can be used to perform GWAS by considering it as a latent phenotype, without the need for any hand-crafted features or annotation. Multiple methods can be employed to learn this function $f(\cdot)$, such as Autoencoders (AEs) and Variational Autoencoders (VAEs). Diffusion Autoencoders (DiffAE) have demonstrated superior performance in both latent representation and image generation tasks [4], utilising a diffusion-based approach to generate high-quality images while learning a meaningful latent representation, outperforming traditional methods like AE and VAE in reconstructing fine details and preserving global image structures. We have implemented a 3D diffusion autoencoder to learn a 3D feature set θ_{3D} for $f(\cdot)$, in order to obtain the latent representation $\mathbf{Z}$ from a given $\mathbf{V}$. Formally, the 3D diffusion autoencoder is defined as a conditional generative model comprising a semantic encoder E_ϕ and a time-dependent denoising network ϵ_ψ . The latent representation $\mathbf{Z} \in \mathbb{R}^d$ is obtained via the deterministic mapping:

$$\mathbf{Z} = E_\phi(\mathbf{V}) \quad (1)$$

Unlike standard autoencoding, $\mathbf{Z}$ is optimised to minimise the reweighted variational lower bound of the diffusion process. Specifically, $\mathbf{Z}$ serves as the conditioner for the score estimation function ϵ_ψ , such that the training objective is:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t \sim [1, T], \mathbf{V}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon - \epsilon_\psi(\sqrt{\bar{\alpha}_t} \mathbf{V} + \sqrt{1 - \bar{\alpha}_t} \epsilon, t, \mathbf{Z})\|^2] \quad (2)$$

where $\mathbf{V}$ represents the input volume drawn from the data distribution, t is the diffusion timestep, $\bar{\alpha}_t$ follows a fixed variance schedule, and $\epsilon_\psi(\cdot, t, \mathbf{Z})$ estimates the noise component conditioned on the latent topology $\mathbf{Z}$. Thus, $\mathbf{Z}$ captures the high-level semantic information required to guide the reconstruction of $\mathbf{V}$ from isotropic Gaussian noise. Further details can be found in the Supplementary Methods section.

Implementation training and evaluation of 3D DiffAE

We used 62,562 subjects with 4-chamber view CMRs along the long axis of the heart, having 50 time points, to train the models, and then infer the embeddings using the trained models. The dataset was divided into training ($N_{train} = 23,776$), validation ($N_{validation} = 3,840$), and held-out test sets ($N_{test} = 35,714$). Before providing the images as input to the model, the data was pre-processed using a two-step normalisation process: first, min-max normalisation is applied based on the input volume's own range, followed by z-score normalisation with a mean of 0.5 and a standard deviation of 0.5.

In our implementation of the 3D DiffAE, we set the network's base channel size to 32 and the dimension of the latent space $d = 128$. We empirically validated this choice through the benchmarking experiments detailed in the Results section (**Supplementary Figures 2C-D** and **6**), which demonstrated that $d = 128$ yields superior reconstruction quality and phenotypic association compared to $d = 64$ or $d = 256$.

The implementation was carried out using PyTorch, and the complete code is publicly available (see Code Availability) [49, 50]. We trained the 3D DiffAE model, which comprises approximately 20.8 million trainable parameters, by optimising the objective function (see **Supplementary Eq. 7**) using the Adam optimiser, with a batch size of two, for 200 epochs on an Nvidia Tesla V100-SXM2-32GB GPU, utilising 16-bit automatic mixed precision (AMP). We performed gradient clipping during training. Rather than using the entire dataset for each epoch, we randomly sampled 25% of the training and 2.5% validation samples during each epoch, thereby introducing additional randomness into the gradient estimation, which acts as a form of implicit regularisation. Furthermore, we applied minimalist data augmentation by randomly flipping images horizontally with a probability of 0.5, in order to enhance the model's generalisation capability.

Each epoch of training and validation took an average of 42 minutes, resulting in a total training time of approximately 140 hours (around six days). Upon completion of training, we selected the model state that achieved the best validation loss as the final model to avoid any possible overfitting, which was subsequently used for inference on the entire dataset. The full inference process, involving the acquisition of both embeddings and reconstructions using $t_step = 20$, took approximately 12.5 hours for 10,000 subjects.

It is well-established in the literature that factors models, including autoencoders are non-identifiable, thereby causing the latent factors to capture different features from the image during a given model run [51, 52]. This variability highlights the importance of conducting multiple runs, as reliance on a single run may introduce bias and lead to suboptimal representations that do not generalise effectively. By performing several runs, we can capture a broader range of latent features, enabling us to determine which aspects of the representations are robust across runs and which are more sensitive to initial conditions. This approach facilitates a more comprehensive exploration of the latent representation of the given input. Consequently, we initialised the model's weights using five different seeds: (1701, 1993, 2023, 42, and 1994), thereby obtaining five sets of latent factors.

We evaluated the perceptual quality of the reconstructions from the trained models across five seeds using the Structural Similarity Index (SSIM) [53], which accounts for changes in structural information, luminance, and contrast as well as Peak Signal-to-Noise Ratio (PSNR) by measuring noise introduced in the reconstruction.

Selection of Latent phenotypes across model runs

We obtained five sets of 128-dimensional latent factors from the entire dataset by training the models with five different seeds, resulting in a total of 640 latent factors. Across these sets of representations, while some features may be captured more frequently than others, certain factors might be identified uniquely in a single model run. To maximise reproducibility and select the most robust factors, we focused on latent factors consistently identified across multiple model runs for downstream analyses. The correlation coefficients between pairs of latent factors across the model runs were computed, and we subsequently selected those latent factors that exhibited a paired correlation of 0.8 or higher (see Supplementary Algorithm 2). This process resulted in 182 latent factors, which we considered the final set of latent phenotypes used for downstream analyses. We refer to them using the following nomenclature: Zn_Si , where n is the latent factor number from the original training (1 to 128), and i is a number from 1 to 5 to distinguish the model run.

Association of latent phenotypes with traits

Curated diseases were defined using three different data sources available in UK Biobank: GP clinical event records (UKB data-field ID: 42040), hospital inpatient (HI) admissions (UKB data-field ID: 41270) and self-reported (SR)

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

medical conditions (UKB data-field ID: 20002), converting ICD-9 codes to ICD-10 and discarding the mismatches. Tables were manually filtered, selecting records related to diseases of the cardiovascular system (including hypertension), Type 2 Diabetes and hypercholesterolaemia. Records of heart failure, left ventricular failure and cardiomegaly were grouped into the "Heart failure" category, while "Conduction block" included Tawara branches, atrioventricular block and fascicular block. Individuals without any record of these diagnoses were considered healthy. For the continuous traits, 28 MRI-derived structural and functional cardiac and aortic endophenotypes were downloaded from UK Biobank (category ID: 157) [16]. Linear regression with ordinary least squares (OLS) was employed to test for associations between each latent phenotype and incident diseases or continuous traits in the discovery cohort ($N = 48482$), including genetic sex (UKB data-field ID: 22006), age at acquisition (computed from the MRI scan date and the birth year/month, UKB data-field IDs: 34 and 52 respectively), standing height, waist circumference (data-field ID: 48), body surface area (computed from weight and standing height, UKB data-field IDs: 21002, 50 respectively), MRI acquisition year and MRI centre (taken from the DICOM header) as covariates. An additional analysis was performed removing waist circumference and standing height from the covariates, results are shown in **Supplementary Figure 5A-B**. Only diseases with a minimum of 200 diagnosed subjects within 5 years before and one year after the MRI acquisition were considered. Associations with very small magnitude effect sizes ($< |1 \times 10^{-6}|$) were discarded. The P-value threshold for significance was $\frac{0.05}{n}$, with n equal to the number of latent-trait pairs.

Linear and logistic regression models with L1 penalty were trained to predict the traits and diseases from our embeddings, cardiac IDPs, or both, depending on the specific analysis. We trained the models on the discovery cohort with 5-fold cross-validation, and tested on the replication cohort. To include only non-collinear covariates, we iteratively removed all features with a variance inflation factor > 10 . Moreover, features and continuous outcomes were standard scaled. Additional models were trained conditioning on LV end diastolic volume.

We performed latent manipulation for a certain target phenotype and latent factor $\mathbf{z} = (\mathbf{z}_{\text{sem}}, x_T)$ from a single-seeded DiffAE model following the formula:

$$\tilde{\mathbf{z}}_{\text{sem}}^{\text{norm}} = \mathbf{z}_{\text{sem}}^{\text{norm}} + m\sqrt{D}\beta \quad (3)$$

where m denotes the manipulation magnitude, D the dimension of $\mathbf{z}_{\text{sem}}$ (128 in our case), β the normalised regression coefficients for the considered target (possibly conditioned); the superscript *norm* is used to indicate standard scaled vectors [4]. The manipulated latent $\tilde{\mathbf{z}} = (\tilde{\mathbf{z}}_{\text{sem}}, x_T)$ was decoded, and the reconstruction was visually inspected. We repeated the latent manipulation process for different m , $\mathbf{z}$, and targets.

For cardiac trait quantification, an experienced cardiologist (C.A.P.) manually traced LV endocardial and epicardial borders at end-diastole on the 4Ch view for the original and all manipulated reconstructions used for latent space validation. We acknowledge that these quantifications were performed by a single expert observer. Whilst this approach ensures internal consistency, the absence of multiple raters precludes the assessment of inter-observer variability and may inadvertently introduce an element of subjective bias. LVEDV (millilitre) was computed using the single-plane area-length (prolate ellipsoid) method from the 4Ch view, combining the planimetered LV endocardial cavity area at end-diastole with the LV long-axis length from the midpoint of the mitral annular plane to the apex. Myocardial wall thickness was assessed on the same apical 4Ch view at end-diastole as the linear thickness of the mid inter-ventricular septum. For each manipulation magnitude m , mid-septal wall thickness was recorded for the original and manipulated reconstructions and expressed as absolute values and as percent change relative to the original reconstruction.

UK Biobank genotyping and quality control

Full details of UK Biobank genotyping and quality control have been described elsewhere [48]. Briefly, the samples were genotyped using a combination of two arrays: the UK Biobank Axiom array from Affymetrix and the UK BiLEVE array, then imputed with the Haplotype Reference Consortium and the UK10K +1000 Genomes haplotype panels. Quality control involved discarding genotyped variants with call rate $< 95\%$, imputed variants with INFO score < 0.3 , and variants with minor allele count (MAC) < 100 , corresponding to a minor allele frequency (MAF) of about 0.001. Variants with severe Hardy-Weinberg equilibrium (HWE) violations were removed (test P-value $< 10^{-30}$). After variant-level QC, 42,795,932 autosomal SNPs remained for the analysis, with positions reported in GRCh37 coordinates. Subjects without imputed genetic data, high missing rate ($\geq 10\%$) or having discordant sex annotations (self-reported and inferred) were excluded from the analysis.

GWAS of inferred latent phenotypes

Genome-wide association study (GWAS) on the discovery cohort was performed using an automated Nextflow pipeline developed in-house based on REGENIE [54] (v3.3.0). Autosomes were tested assuming an additive genetic model and including as covariates: genetic sex (UKB data-field ID: 22006), age at acquisition (computed from the MRI scan date and the birth year/month, having UKB data-field IDs: 34 and 52, respectively), genotyping batch (UKB data-field ID: 22000), body surface area (BSA, computed from weight and standing height, with the formula: $BSA(m^2) = 0.007184 \times \text{Weight}(kg)^{0.425} \times \text{Height}(cm)^{0.725}$ [55]), year in which the imaging was performed (2014–2023), assessment centre of the imaging visit, and visit number. All the information regarding the MRI acquisitions was extracted from the DICOM images. REGENIE performs genome-wide association studies in two steps. In the first step, a whole-genome regression model that captures part of the phenotypic variance due to genetic effects is fit using a subset of variants. This subset, which included 417,400 SNPs, was generated as suggested in [56], *i.e.*, by using only genotyped data, with minor allele frequency (MAF) $\geq 1\%$, call rate $\geq 99\%$, HWE test P-value $> 10^{-15}$, followed by linkage-disequilibrium (LD) pruning (1000 variant windows, 100 sliding windows and $R^2 > 0.8$, performed with PLINK [57], version 1.9). In the second step, REGENIE tests genetic variants for association conditioning on the prediction generated by the regression model fit in the first step, using a leave-one-chromosome-out (LOCO) scheme, thus avoiding proximal contamination. The autosomal panel of 42,795,932 common SNPs that passed QC having minor allele frequency $\geq 1\%$ was used in our genome-wide association study. Variants in the major histocompatibility complex (MHC) region (GRCh37 chr6:28,477,797–33,448,354) were excluded due to its long-range LD structure.

The GWAS discovery cohort included 47,740 Caucasian individuals that passed both phenotypic and genotype QC (**Supplementary Figure 12**). 182 latent phenotypes were tested after quantile normalisation. The associations were considered significant if the discovery P-value in Step 2 was lower than $5 \times 10^{-8}/22 = 2.27 \times 10^{-9}$, where 22 is the effective number of independent tests computed using the approach proposed by Li & Ji [17], taking into account the strong correlation among latent phenotypes. Genome-wide Manhattan plots were generated with *topR* R package [58] (v2.0.1). We annotated the nearest gene or the overlapping gene using the *closest* function from *bedtools* (v2.31.0). Positions are reported in GRCh37 coordinates.

Heritability and genetic correlation

SNP-based heritability of the latent phenotypes and genetic correlations against cardiac traits and diseases were computed using the R package *GenomicSEM* [59] (v0.0.5c) on the latent phenotypes GWAS summary statistics. *GenomicSEM* relies on a two-stage structural equation modelling approach based on a multivariable extension of LD score regression (LDSC) [24] that accounts for varying and potentially unknown degrees of sample overlap. Genetic correlations (r_g) were obtained from the standardised genetic covariance matrix. Statistical significance was assessed using two-sided Wald tests (r_g/SE), with corresponding χ^2 statistics (1 degree of freedom). The set of relevant cardiometabolic traits and diseases is the same as the one used for colocalisation analysis, explained below. A Bonferroni-adjusted significance threshold of $P < 4.21 \times 10^{-05}$ was applied, calculated as $0.05/(22 \times 54)$, where 22 is the effective number of independent tests for the latent phenotypes and 54 is the number of traits tested.

Statistical fine mapping

The 95 genome-wide significant associations refer to the last steps of the GWAS pipeline, which performs clumping with PLINK [57] (version 1.9). The genome-wide significant P-value was used as an index variant threshold, including in the clump all the SNPs within 250 kb having r^2 larger than 0.5. Significant variants were annotated with the closest gene using *bedtools* [60] (version 2.30). To dissect all the independent signals associated with each latent phenotype, conditional analysis was performed on raw summary statistics with GCTA's Conditional & joint (COJO) analysis [61] (v1.94.0 beta), considering a window of 10 Mb, collinearity = 0.9 and variants with a MAF $> 1\%$. Genotypes from 30,000 randomly selected UKB unrelated white British participants were used as reference panel for LD computations. The 89 variants having an association $P_j \leq 2.27 \times 10^{-9}$ in the COJO joint model were considered significant and independent. To define a number of globally significant loci across traits, genome-wide significant genomic regions overlapping among different latent phenotypes were merged considering linkage disequilibrium. LD was computed with PLINK for each variant within a window of 500kb from an independent SNP. If two significant and independent SNPs inside the same genomic window were in LD ($r^2 > 0.1$), only one signal was considered and the variant with the strongest P-value was selected as lead. On the other hand, if inside the same region two SNPs were not in high LD, they were both kept as independent signals. This led to the definition of 44 lead independent SNPs in 42 genomic loci.

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

GWAS Replication

Association study on the replication cohort was performed using REGENIE with the same configuration and set of covariates used in the discovery GWAS. The cohort included a total of 12,396 subjects of all ancestries, specifically 4,077 Caucasian and 8,319 non-Caucasian (Baseline characteristics in **Supplementary Data 13**, full ethnicity breakdown in **Supplementary Data 33**, filtering process of individuals and SNPs illustrated in **Supplementary Figure 12**). The allele frequency comparison with the discovery cohort is shown in **Supplementary Figure 15A**. Plots were created with the Python package *GwasLab* (v3.4.24) [62]. The 11,907 individuals passing genotype QC were included in the genetic analysis. Given the limitations of our replication cohort, characterised by both underpowered sample size and increased ethnic heterogeneity compared to the discovery cohort, we adopted a comprehensive approach to assess variant replication: lenient – how many variants have a concordant effect direction in the replication cohort; nominal – how many variants replicate at $P\text{-value} \leq 0.05$ with consistent direction of effect; and conservative – how many variants replicate with replication $P\text{-value} \leq 0.05/44 = 1.1 \times 10^{-3}$, where 44 is the number of lead independent SNPs identified in the previous step. Not only was sentinel lead SNP replication considered, but also SNPs in high LD ($r^2 > 0.6$) with the significant variants.

Curation of variants

We interrogated the NHGRI-EBI GWAS Catalog database through the *LDtrait* tool (inside *LDlink* [63] R package, v1.1.2) to identify any overlap between significantly associated SNPs discovered in our study (or in high linkage disequilibrium with them; $r^2 \geq 0.6$) and previously reported associations in published summary statistics. Associations were considered novel if previously unreported in the GWAS Catalog database. Additionally, Open Targets Genetics [20, 64] database (<https://genetics.opentargets.org> release October 2022, v22.10) was used to manually explore the lead variants.

Gene-by-sex interaction GWAS

To examine sex differences in the genetic architecture of latent phenotypes, a gene-by-sex interaction analysis was performed separately with REGENIE, testing for the interaction effect between being male and SNPs in the discovery cohort. We corrected for the same covariates as the discovery GWAS, excluding genetic variants that had a minor allele frequency < 0.01 . Also, the same Bonferroni-corrected threshold as discovery GWAS was used to determine significance. Conditional analysis was performed with the same configuration of the discovery GWAS, considering the suggestive significance of 5×10^{-8} as threshold (**Supplementary Data 8**).

Published GWAS summary statistics for colocalisation

A total of 27 summary statistics from published studies were downloaded from GWAS Catalog [18]. The downloaded traits span multiple biological scales and phenotypic categories, including cardiac, ECG and blood pressure traits, cardiac diseases and non-cardiac traits related to body measures and other organs. The summary statistics of "Dark brown hair colour" from UKB were included as a negative control. In addition, 27 summary statistics were produced with REGENIE for the MRI-derived cardiac and aortic endophenotypes from UK Biobank (Field ID: 157) [16], using the same process as the one previously described for GWAS on latent phenotypes. **Supplementary Data 21** lists all the relevant phenotypes used in downstream analyses with the respective number of significant loci, cohort size, proportion of cases and controls, PMID and URL for the download. All the study cohorts include European UKB individuals; however, some of the summary statistics derive from meta-analysis including also other cohorts. For 24 traits more than 80% of the genetic variants are shared with the UKB genotype data, whilst 3 GWAS summary statistics have an overlap of approximately 70%. Summary statistics were formatted and lifted to the hg38 reference genome build.

Conditional analysis and colocalisation

In addition to the 54 GWAS summary statistics described above, we included GTEx eQTLs v8 from 12 cardiovascular-related tissues and other organs (Heart - atrial appendage, Heart - left ventricle, Artery - aorta, Artery - coronary, Artery - tibial, Adipose - subcutaneous, Adipose - visceral omentum, Muscle - skeletal, Whole blood, Lung, Liver, Pancreas) in the colocalisation analysis. GWAS and eQTL summary statistics were harmonised and variants were filtered by minor allele frequency ($MAF > 0.01$).

Multi-trait colocalisation test was based on the Bayesian framework of *coloc* described by Giambartolomei et al. (2014) [65]. We defined trait-specific genomic regions by selecting all SNPs with a p-value threshold of 1×10^{-5} and setting loci boundaries between SNPs more than 250kb apart. We enlarged each genomic region by 100kb on both sides. Among such identified loci, only those having at least one SNP with a p-value $< 2 \times 10^{-9}$ for latent phenotypes, $< 2.5 \times 10^{-7}$ for GTEx (as suggested in their original paper [22]) and $< 5 \times 10^{-8}$ for the downloaded disease/trait GWAS, were considered significant and carried out to the following analyses. The loci which encompassed the MHC II region (GRCh38 chr6:28,510,120-33,480,577) were excluded from the analysis due to the difficulties in properly fine-mapping them.

For each locus, we tested if secondary conditionally independent associations existed and which were shared with other traits. To this end, we developed a statistical approach that divides each GWAS-significant genomic region into its component signals and then uses colocalisation across different traits to group association signals likely caused by the same SNP into single modulatory modules. We performed conditional analysis of each significant locus using GCTA-COJO software [61] in two steps. First, a stepwise forward conditional regression procedure was applied to select the independently associated SNPs at each locus identified as described above. The stopping criterion employed was that all conditional and joint p-values (pJ) were larger than 1×10^{-4} . This resulted in the identification of independent SNPs within the genomic region boundary. Then, for each of the identified variants, association analysis was performed conditional on all the other independent signals except for the target one (leave one out approach), generating a conditional dataset having only one independent association signal. This procedure is similar to that used by Robinson et al., but instead of using the step-wise conditioned datasets, it uses an ‘all but one’ approach. True independent signals were defined as those being genome-wide significant (according to the thresholds listed above) in the original GWAS and with a joint p-value $< 1 \times 10^{-6}$ or having a genome-wide significant joint p-value. When possible, we used in-sample LD for the conditional analysis. For the downloaded GWAS tables, the UKB LD reference containing our latent phenotyped individuals was used and DENTIST [66] was applied prior to COJO to detect and remove variants showing heterogeneity between the GWAS and LD reference samples. We thus generated a set of conditional datasets for each trait and locus where the association signal is driven by a single shared variant. We then fine-mapped each conditional dataset obtained by computing the 99% credible set, using the ‘finemap.abf’ function from the R package *coloc* (v5.2.3) [65].

To understand which loci were pleiotropic between traits, we ran pairwise colocalisation using *coloc* [65] for each combination of conditional datasets having at least one SNP in their 99% credible set in common. The colocalisation tests resulting in $PP.H_4 > 0.8$ (posterior probability for the hypothesis of same underlying variant contributing to both traits variation) were considered significant. Finally, since the *coloc* package allows only for pairwise testing, we followed up colocalisation analysis with network analysis (using the R package *igraph* v2.0.2 [67]) to identify, among multiple pairs of colocalising conditional datasets ($PP.H_4 > 0.8$), larger regulatory modules of traits all sharing the same variant. Locuszoom plots were created adapting the functions from *LocusCompareR* package (v1.0.0) [68].

Functional enrichments

To facilitate coherent and computationally feasible post-GWAS functional analyses, we consolidated multiple sets of GWAS summary statistics into a single dataset. This was achieved by selecting the minimum P-value for each SNP across the 182 GWAS summary statistics, referred to here as “minP GWAS”, which represents the strongest association for each SNP across all phenotypes. This approach prioritises the most significant genetic signals, irrespective of the phenotype in which they appear, thereby maximising the likelihood of identifying SNPs with true associations. Gene, gene-set and tissue expression analysis were performed with the MAGMA tool (v1.08) [27] as implemented inside FUMA software (<https://fuma.ctglab.nl/> v1.5.2) [25], using the SNP2GENE process on the aggregated minP GWAS summary statistic. A total of 17,005 gene sets were tested, obtained from MSigDB (v2023.1Hs) for “Curated gene sets” and “GO terms”, with window size set to ± 1 kb and using the full distribution of SNP p-values. Gene sets with Bonferroni-adjusted P-value ≤ 0.05 were considered significant. To assess tissue specificity, we used MAGMA’s gene-property analysis to test the relationships between latent-gene associations and tissue-specific expression profiles (RPKM) from 53 tissue types in GTEx v8. This one-sided test evaluated average gene expression per tissue, conditioned on the average expression across all tissues. Tissue types with Bonferroni-adjusted P-value ≤ 0.05 were considered significant.

FUMA was also employed to map genes through position (maximum distance 10 kb) and eQTL (tissues: heart atrial appendage and left ventricle). The resulting list of 85 mapped protein coding genes was used as input for an over-representation analysis with Enrichr [26], performed through the Python package *GSEAPy* (v1.1.3) [69].

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

The Chromatin Element Enrichment Ranking by Specificity (CHEERS) method [28] was employed to assess whether latent-associated variants were enriched in chromatin accessible regions of nine cardiac cell types. We used the read counts per peak matrix from a snATAC-seq cardiac atlas comprising atrial and ventricular cardiomyocyte, smooth muscle, endothelial, adipocyte, macrophage, fibroblast, lymphocyte and nervous cells [70]; together with the 42 genomic risk loci ($LD\ r^2 > 0.1$) of 500 kb from the GWAS summary statistics of latent phenotypes. Before computing the enrichment, CHEERS removed the bottom decile of peaks with the lowest read counts and obtained a cell type specificity score through Euclidean normalisation. The resulting enrichment one-sided P-value was corrected for the number of cell types tested (0.05/9).

TWAS

Transcriptome-wide association study was performed to identify associations between our latent phenotypes and the imputed tissue-specific gene expression from GTEx v8 [22]. The FUSION tool [71] was used on the aggregated minP GWAS summary statistic with its default settings, downloading the precomputed transcript expression reference weights of 12 tissues (Heart - atrial appendage, Heart - left ventricle, Artery - aorta, Artery - coronary, Artery - tibial, Adipose - subcutaneous, Adipose - visceral omentum, Muscle - skeletal, Whole blood, Lung, Liver, Pancreas) for European samples from FUSION authors' website <http://gusevlab.org/projects/fusion/>. To account for multiple hypotheses, a Bonferroni-corrected P-value threshold ($0.05/n$ of genes tested) was used to define significant TWAS associations. The π_1 statistic was computed to assess the degree of overlap in significant genes between tissue types. For each tissue pair, we filtered the significant genes in the target tissue to those overlapping with the reference tissue. We then computed the q-value object for this subset of P-values and derived the π_1 as $1 - \pi_0$, where π_0 represents the proportion of null hypotheses that are true.

Heart single-nucleus RNA sequencing

Single-nucleus RNA sequencing (snRNA-seq) data of six anatomical adult heart regions (left and right atrium, left and right ventricles, apex and inter-ventricular septum) was obtained from a study conducted by Litviňuková et al. [34]. The clustering and annotation of the cell types were available within the log-normalised H5AD file provided by the authors. The primary analysis and visualisation were conducted using the Python package *Scanpy* (v1.10.1). We computed the gene set score of the 82 MAGMA-prioritised protein-coding genes using the `score_genes` function, and then tested whether these 82 genes were differentially expressed in the two cardiomyocytes populations compared to the other cell types with the `rank_genes_groups` function using the T-test with overestimated variance.

UK Biobank exome sequencing

Single-variant exome-wide association study (EWAS) and rare variant association tests, utilising the whole exome sequencing (WES) data (UKB data-field ID: 23159) on the discovery cohort, were performed on the UKB Research Analysis Platform (RAP, cf. <https://ukbiobank.dnanexus.com>) using REGENIE [54] (version 3.4.1). WES data available on the RAP are the raw calls after merging sample-level data with GLNexus [72], and QC [73, 74] with bcftools [75] was performed before using them. Low-quality genotypes were set to missing, specifically, SNVs with a depth (DP) less than 7 and indels with a DP less than 10. Selected variants met the following criteria: at least 75% of non-missing genotypes with quality ≥ 20 , ratio of balanced heterozygous calls to total heterozygous calls ≥ 0.5 , missing call rate $< 10\%$, and HWE test P-value $> 10^{-15}$. Subjects with sex discordance were also removed. A total of 26,388,600 variants aligned to GRCh38 passed QC and were utilised for EWAS and the rare variant tests.

Rare-variant association tests

Regarding the exome wide-association study, the first step of REGENIE was identical to the GWAS (as described in the previous sub-section). For the second step, the WES release tranche (UKB data-field ID: 32050) was used as an additional covariate to correct for batch effects, along with the covariates used for GWAS. EWAS was performed on variants with $MAC > 5$, integrating the conditionally independent genome-wide significant variants obtained after performing GWAS as covariates for each latent-chromosome pair (for the latent-chromosome pairs without any significant hit, no conditioning was performed). A total of 46,034 individuals from the discovery cohort passing genotype QC were included in the EWAS analysis, while for the burden tests 46,134 subjects were included (see **Supplementary Figure 12**). For the burden tests, helper files (available on the UKB RAP) containing variant annotations and gene sets collapsing rare variants ($MAC > 1$) were utilised. Specifically, variants were annotated using SnpEff [76] based on the Ensembl 85 gene model, selecting the most severe consequence for each variant across

all protein-coding transcripts [73, 74]. For each gene, three categories of rare variant masks were considered: M1 with Loss-of-Function (LoF) variants only [LoF], M2 with LoF variants and likely deleterious missense variants - predicted to be deleterious by all five scoring algorithms - [LoF, missense (5/5)], and finally, M3 with LoF variants and possibly deleterious missense variants (predicted to be deleterious by at least one of the five scoring algorithms) [LoF, missense ($\geq 1/5$)]. For each of these categories, two different burden masks per gene were built using the maximum number of alternative alleles across sites based on the frequency of the alternative allele of the variants: $AAF < 0.01$ and singletons only. Furthermore, the SKATO-ACAT test, which combines the strengths of burden tests and sequence kernel association tests (SKAT), allowing for robust detection of genetic associations regardless of variant effect directions, was performed for those three mask categories using variants with $AAF < 0.01$. When performing gene burden tests on whole exome sequencing (WES) data, the conventional genome-wide significance threshold of $p < 5 \times 10^{-8}$ may be overly stringent, as the analysis is focused on a subset of the genome—the exome—and the tests are conducted on gene sets rather than individual variants. Therefore, we recalculated the significance threshold using a method based on the distribution of the minimum p-values across different phenotypes [37].

The number of independent tests was estimated by taking the median of the minimum inverted p-values across all phenotypes, and then dividing the nominal significance level of 0.05 by this value. This resulted in an adjusted P-value threshold of $p < 4.7 \times 10^{-07}$ for SKATO-ACAT.

Polygenic risk score calculation and disease prediction

Polygenic risk scores (PRS) for each latent were computed for 314,774 Caucasian subjects (the entire UKB Caucasian cohort with available baseline data and genotypes, excluding subjects with a kinship coefficient > 0.0625 (computed using PLINK version 2.0) and subjects used for the discovery GWAS) using LDpred2-auto [77, 78, 79]. Similar to previous approaches [78], variants with a missing call rate < 0.1 , $INFO > 0.4$ and $MAF > 0.01$ were pruned using PLINK [57] (version 2.0) using a window size of 250kb, a step size of 5, and an unphased hardcall R^2 threshold of 0.5, resulting in 1,281,751 variants. Subsequently, all conditionally independent genome-wide significant variants were included, leading to a final total of 1,281,790 variants that were used for computing the PRS from the summary statistics of the discovery GWAS.

Furthermore, multi-PRS [80] for ten cardiac (and related) diseases were computed by applying Lasso (stratified five-fold cross-validation using `LogisticRegressionCV` function of *scikit-learn* [81] with the L1 penalty and saga solver). Multi-PRS models were trained by supplying the individual PRS computed for each latent factor, along with baseline characteristics (age, sex, BMI, and smoking status) as input and whether the subject was a positive or negative control case with respect to the particular disease as the target.

To evaluate the predictive ability of individual latent-specific PRS and multi-PRS, binary disease cohorts were created for angina pectoris, atherosclerotic conditions, conduction block, coronary heart disease, heart failure, high cholesterol, hypertension, metabolic syndrome, myocardial infarction, and type 2 diabetes, by taking positive cases from the entire UKB Caucasian cohort (excluding the cohort used for the discovery GWAS) and an equal number of negative super controls (i.e., subjects without any cardiac or metabolic disease marked as positive) matched for sex and age. Two types of predictive task were evaluated: diagnostic and prognostic, with respect to the date of collection of the baseline characteristics. If the disease was reported on or before the date of baseline data collection, the subject was included in the diagnostic cohort, while if the disease was reported between one and ten years after the collection of the baseline data, the subject was included in the prognostic cohort. The number of subjects for each disease is reported in the **Supplementary Data 26**. Five-fold cross-validation was performed for each model to evaluate them without data selection bias. For the latent-specific PRS models, separate predictors were trained for each by applying linear regression on the latent-specific PRS and the set of baseline characteristics, and then the predictor demonstrating the best Area Under the Receiver Operating Characteristic Curve (ROC-AUC, or simply AUC) was selected, referred to here as max(Single PRS + Baseline). The performances of the max(Single PRS + Baseline) and Multi PRS + Baseline for those ten diseases were benchmarked against a linear regression predictor trained solely using the baseline characteristics. Finally, to contextualise the performance of our disease-agnostic approach, we extended our evaluation to include two additional benchmarks assessing diagnostic performance across four representative conditions: atrial fibrillation, cardiovascular disease, hypertension, and type 2 diabetes. First, we compared our latent multi-PRS against standard, established PRS available within the UK Biobank (Data-Field IDs: 26212, 26223, 26244, and 26285). Second, to validate the robustness of our latent representation against explicit phenotypic data, we derived a multi-PRS based on cardiac imaging-derived phenotypes (Cardiac IDPs; UK Biobank Category ID 157), constructed using the same methodology applied to our latent PRS.

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

Resource availability

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Craig A. Glastonbury (craig.glastonbury@fht.org).

Data availability The GWAS, EWAS and rare variant summary statistics generated in this study are currently available via the project page (https://glastonburygroup.github.io/CardiacDiffAE_GWAS/) and will be deposited in the GWAS Catalog (<https://www.ebi.ac.uk/gwas/>); accession codes will be provided on the project page upon release. The PRS scoring files generated in this study are likewise available via the project page and will be deposited in the PGS Catalog (<https://www.pgscatalog.org/>). The trained 3D DiffAE model weights are available on Hugging Face (<https://huggingface.co/GlastonburyGroup>). The individual-level phenotypic, imaging and genotypic data used in this study are available through the UK Biobank (<https://www.ukbiobank.ac.uk>) under application number 82779. These data are available under restricted access owing to participant consent and data protection requirements; to any UK biobank approved researcher <https://www.ukbiobank.ac.uk/enable-your-research/register>. Information and URLs of the publicly available GWAS summary statistics used for colocalisation analysis are detailed in **Supplementary Data 21**. The full *cis*-eQTL summary statistics from GTEx v8 were obtained from <https://www.gtexportal.org/home/protectedDataAccess> on 11 June 2024 under dbGaP application phs000424.v2.p1. Single-nucleus RNA sequencing data objects were downloaded via the Heart Cell Atlas web portal (<https://www.heartcellatlas.org/>). Source data for this manuscript are available in the Supplementary materials and in the Source Data files. Source Data files (one for all the main figures combined, one each for the supplementary figures) are provided with this paper.

Code availability For detailed information, including video presentations, see our project page at https://glastonburygroup.github.io/CardiacDiffAE_GWAS. The codebase of this project (except for the phenotyping using 3D DiffAE) is available at https://github.com/GlastonburyGroup/CardiacDiffAE_GWAS [50], whilst the pipeline for phenotyping using 3D DiffAE is available at <https://github.com/GlastonburyGroup/ImLatent> [49]. For the original source code of DiffAE (2D, for RGB images), see <https://github.com/phizaz/DiffAE>. The in-house GWAS pipeline based on REGENIE is available at <https://github.com/HTGenomeAnalysisUnit/nf-pipeline-regenie>. The rare variant analyses using WES were conducted on the UKB Research Analysis Platform (<https://ukbiobank.dnanexus.com>).

References

- [1] B. Gomes, A. Singh, J. W. O'Sullivan, T. M. Schnurr, P. C. Goddard, S. Loong, D. Amar, J. W. Hughes, M. Kostur, F. Haddad, *et al.*, "Genetic architecture of cardiac dynamic flow volumes," *Nature Genetics*, vol. 56, no. 2, pp. 245–257, 2024.
- [2] J. P. Pirruccello, M. D. Chaffin, E. L. Chou, S. J. Fleming, H. Lin, M. Nekoui, S. Khurshid, S. F. Friedman, A. G. Bick, A. Arduini, *et al.*, "Deep learning enables genetic analysis of the human thoracic aorta," *Nature genetics*, vol. 54, no. 1, pp. 40–51, 2022.
- [3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986," *Biometrika*, vol. 71, no. 599–607, p. 6, 1986.
- [4] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.
- [5] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *MICCAI workshop on deep generative models*, pp. 117–126, Springer, 2022.
- [6] K. Kreis, T. Dockhorn, Z. Li, and E. Zhong, "Latent space diffusion models of cryo-em structures," *arXiv preprint arXiv:2211.14169*, 2022.
- [7] M. Sadria and A. Layton, "scvaeder: integrating deep diffusion models and variational autoencoders for single-cell transcriptomics analysis," *Genome Biology*, vol. 26, no. 1, pp. 1–17, 2025.
- [8] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [9] M. Kirchler, S. Konigorski, M. Norden, C. Meltendorf, M. Kloft, C. Schurmann, and C. Lippert, "transfergwas: Gwas of images using deep transfer learning," *Bioinformatics*, vol. 38, no. 14, pp. 3621–3628, 2022.
- [10] R. Bonazzola, E. Ferrante, N. Ravikumar, Y. Xia, B. Keavney, S. Plein, T. Syeda-Mahmood, and A. F. Frangi, "Unsupervised ensemble-based phenotyping enhances discoverability of genes related to left-ventricular morphology," *Nature Machine Intelligence*, vol. 6, no. 3, pp. 291–306, 2024.
- [11] K. Patel, Z. Xie, H. Yuan, S. M. S. Islam, Y. Xie, W. He, W. Zhang, A. Gottlieb, H. Chen, L. Giancardo, *et al.*, "Unsupervised deep representation learning enables phenotype discovery for genetic association studies of brain imaging," *Communications Biology*, vol. 7, no. 1, p. 414, 2024.
- [12] C. A. Glastonbury, S. L. Pulit, J. Honecker, J. C. Censin, S. Laber, H. Yaghootkar, N. Rahmioglu, E. Pastel, K. Kos, A. Pitt, *et al.*, "Machine learning based histology phenotyping to investigate the epidemiologic and genetic basis of adipocyte morphology and cardiometabolic traits," *PLoS computational biology*, vol. 16, no. 8, p. e1008044, 2020.
- [13] Y. Liu, N. Bastý, B. Whitcher, J. D. Bell, E. P. Sorokin, N. van Bruggen, E. L. Thomas, and M. Cule, "Genetic architecture of 11 organ traits derived from abdominal mri using deep learning," *Elife*, vol. 10, p. e65554, 2021.
- [14] B. Alipanahi, F. Hormozdiari, B. Behsaz, J. Cosentino, Z. R. McCaw, E. Schorsch, D. Sculley, E. H. Dorfman, P. J. Foster, L. H. Peng, *et al.*, "Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology," *The American Journal of Human Genetics*, vol. 108, no. 7, pp. 1217–1230, 2021.
- [15] F. Cisternino, S. Ometto, S. Chatterjee, E. Giacomuzzi, A. P. Levine, and C. A. Glastonbury, "Self-supervised learning for characterising histomorphological diversity and spatial rna expression prediction across 23 human tissue types," *Nature Communications*, vol. 15, no. 1, p. 5906, 2024.
- [16] W. Bai, H. Suzuki, J. Huang, C. Francis, S. Wang, G. Tarroni, F. Guitton, N. Aung, K. Fung, S. E. Petersen, *et al.*, "A population-based phenome-wide association study of cardiac and aortic structure and function," *Nature medicine*, vol. 26, no. 10, pp. 1654–1662, 2020.

- [17] J. Li and L. Ji, "Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix," *Heredity*, vol. 95, no. 3, pp. 221–227, 2005.
- [18] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, *et al.*, "The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog)," *Nucleic acids research*, vol. 45, no. D1, pp. D896–D901, 2017.
- [19] A. Radhakrishnan, S. F. Friedman, S. Khurshid, K. Ng, P. Batra, S. A. Lubitz, A. A. Philippakis, and C. Uhler, "Cross-modal autoencoder framework learns holistic representations of cardiovascular state," *Nature Communications*, vol. 14, no. 1, p. 2436, 2023.
- [20] M. Ghoussaini, E. Mountjoy, M. Carmona, G. Peat, E. M. Schmidt, A. Hercules, L. Fumis, A. Miranda, D. Carvalho-Silva, A. Buniello, *et al.*, "Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics," *Nucleic acids research*, vol. 49, no. D1, pp. D1311–D1320, 2021.
- [21] T. Groza, F. L. Gomez, H. H. Mashhadi, V. Muñoz-Fuentes, O. Gunes, R. Wilson, P. Cacheiro, A. Frost, P. Kesivali-Bond, B. Vardal, *et al.*, "The international mouse phenotyping consortium: comprehensive knockout phenotyping underpinning the study of human disease," *Nucleic acids research*, vol. 51, no. D1, pp. D1038–D1045, 2023.
- [22] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, *et al.*, "The genotype-tissue expression (gtex) project," *Nature genetics*, vol. 45, no. 6, pp. 580–585, 2013.
- [23] Y. Zhou, X. Zhu, H. Cui, J. Shi, G. Yuan, S. Shi, and Y. Hu, "The role of the vegf family in coronary heart disease," *Frontiers in cardiovascular medicine*, vol. 8, p. 738325, 2021.
- [24] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, S. W. G. of the Psychiatric Genomics Consortium, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, "Ld score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nature genetics*, vol. 47, no. 3, pp. 291–295, 2015.
- [25] K. Watanabe, E. Taskesen, A. Van Bochoven, and D. Posthuma, "Functional mapping and annotation of genetic associations with fuma," *Nature communications*, vol. 8, no. 1, p. 1826, 2017.
- [26] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan, "Enrichr: interactive and collaborative html5 gene list enrichment analysis tool," *BMC bioinformatics*, vol. 14, pp. 1–14, 2013.
- [27] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma, "Magma: generalized gene-set analysis of gwas data," *PLoS computational biology*, vol. 11, no. 4, p. e1004219, 2015.
- [28] B. Soskic, E. Cano-Gamez, D. J. Smyth, W. C. Rowan, N. Nakic, J. Esparza-Gordillo, L. Bossini-Castillo, D. F. Tough, C. G. Larminie, P. G. Bronson, *et al.*, "Chromatin activity at gwas loci identifies t cell states driving complex immune diseases," *Nature genetics*, vol. 51, no. 10, pp. 1486–1493, 2019.
- [29] H. Sun, K. C. Olson, C. Gao, D. A. Prosdocimo, M. Zhou, Z. Wang, D. Jeyaraj, J.-Y. Youn, S. Ren, Y. Liu, *et al.*, "Catabolic defect of branched-chain amino acids promotes heart failure," *Circulation*, vol. 133, no. 21, pp. 2038–2049, 2016.
- [30] J.-y. Yu, N. Cao, C. D. Rau, R.-P. Lee, J. Yang, R. J. R. Flach, L. Petersen, C. Zhu, Y.-L. Pak, R. A. Miller, *et al.*, "Cell-autonomous effect of cardiomyocyte branched-chain amino acid catabolism in heart failure in mice," *Acta Pharmacologica Sinica*, vol. 44, no. 7, pp. 1380–1390, 2023.
- [31] H. Chen, A. Moreno-Moral, F. Pesce, N. Devapragash, M. Mancini, E. L. Heng, M. Rotival, P. K. Srivastava, N. Harmston, K. Shkura, *et al.*, "Wwp2 regulates pathological cardiac fibrosis by modulating smad2 signaling," *Nature communications*, vol. 10, no. 1, p. 3616, 2019.
- [32] H. Chen, G. Chew, N. Devapragash, J. Z. Loh, K. Y. Huang, J. Guo, S. Liu, E. L. S. Tan, S. Chen, N. G. Z. Tee, *et al.*, "The e3 ubiquitin ligase wwp2 regulates pro-fibrogenic monocyte infiltration and activity in heart fibrosis," *Nature Communications*, vol. 13, no. 1, p. 7375, 2022.
- [33] T. Zimmer and R. Surber, "Scn5a channelopathies—an update on mutations and mechanisms," *Progress in biophysics and molecular biology*, vol. 98, no. 2-3, pp. 120–136, 2008.

Ometto, Chatterjee et al.

- [34] M. Litviňuková, C. Talavera-López, H. Maatz, D. Reichart, C. L. Worth, E. L. Lindberg, M. Kanda, K. Polanski, M. Heinig, M. Lee, *et al.*, “Cells of the adult human heart,” *Nature*, vol. 588, no. 7838, pp. 466–472, 2020.
- [35] S. Sakaue, M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, S. Koshiba, A. Narita, T. Konuma, K. Yamamoto, M. Akiyama, *et al.*, “A cross-population atlas of genetic associations for 220 human phenotypes,” *Nature genetics*, vol. 53, no. 10, pp. 1415–1424, 2021.
- [36] S. Dong, N. Zhao, E. Spragins, M. S. Kagda, M. Li, P. Assis, O. Jolanki, Y. Luo, J. M. Cherry, A. P. Boyle, *et al.*, “Annotating and prioritizing human non-coding variants with regulomedb v. 2,” *Nature genetics*, vol. 55, no. 5, pp. 724–726, 2023.
- [37] K. J. Karczewski, M. Solomonson, K. R. Chao, J. K. Goodrich, G. Tiao, W. Lu, B. M. Riley-Gillis, E. A. Tsai, H. I. Kim, X. Zheng, *et al.*, “Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 uk biobank exomes,” *Cell Genomics*, vol. 2, no. 9, 2022.
- [38] R. Pekson, F. G. Liang, J. L. Axelrod, J. Lee, D. Qin, A. J. Wittig, V. M. Paulino, M. Zheng, P. M. Peixoto, and R. N. Kitsis, “The mitochondrial atp synthase is a negative regulator of the mitochondrial permeability transition pore,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 51, p. e2303713120, 2023.
- [39] M. A. Gargano, N. Matentzoglou, B. Coleman, E. B. Addo-Lartey, A. V. Anagnostopoulos, J. Anderton, P. Avilach, A. M. Bagley, E. Bakštein, J. P. Balhoff, G. Baynam, S. M. Bello, M. Berk, H. Bertram, S. Bishop, H. Blau, D. F. Bodenstein, P. Botas, K. Boztug, J. Čady, T. J. Callahan, R. Cameron, S. J. Carbon, F. Castellanos, J. H. Caufield, L. E. Chan, C. G. Chute, J. Cruz-Rojo, N. Dahan-Oliel, J. R. Davids, M. de Dieuleveult, V. de Souza, B. B. A. de Vries, E. de Vries, J. R. DePaulo, B. Derfalvi, F. Dhombres, C. Diaz-Byrd, A. J. M. Dingemans, B. Donadille, M. Duyzend, R. Elfeky, S. Essaid, C. Fabrizzi, G. Fico, H. V. Firth, Y. Freudenberg-Hua, J. M. Fullerton, D. L. Gabriel, K. Gilmour, J. Giordano, F. S. Goes, R. G. Moses, I. Green, M. Griesse, T. Groza, W. Gu, J. Guthrie, B. Gyori, A. Hamosh, M. Hanauer, K. Hanušová, Y. O. He, H. Hegde, I. Helbig, K. Holasová, C. T. Hoyt, S. Huang, E. Hurwitz, J. O. B. Jacobsen, X. Jiang, L. Joseph, K. Keramatian, B. King, K. Knoflach, D. A. Koolen, M. L. Kraus, C. Kroll, M. Kusters, M. S. Ladewig, D. Lagorce, M.-C. Lai, P. Lapunzina, B. Laraway, D. Lewis-Smith, X. Li, C. Lucano, M. Majd, M. L. Marazita, V. Martinez-Glez, T. H. McHenry, M. G. McInnis, J. A. McMurphy, M. Mihulová, C. E. Millett, P. B. Mitchell, V. Moslerová, K. Narutomi, S. Nematollahi, J. Nevado, A. A. Nierenberg, N. N. Čajbiková, J. Nurnberger, John I, S. Ogishima, D. Olson, A. Ortiz, H. Pachajoa, G. Perez de Nanclares, A. Peters, T. Putman, C. K. Rapp, A. Rath, J. Reese, L. Rekerle, A. M. Roberts, S. Roy, S. J. Sanders, C. Schuetz, E. C. Schulte, T. G. Schulze, M. Schwarz, K. Scott, D. Seelow, B. Seitz, Y. Shen, M. N. Similuk, E. S. Simon, B. Singh, D. Smedley, C. L. Smith, J. T. Smolinsky, S. Sperry, E. Stafford, R. Stefancsik, R. Steinhaus, R. Strawbridge, J. C. Sundaramurthi, P. Talapova, J. A. Tenorio Castano, P. Tesner, R. H. Thomas, A. Thurm, M. Turnovec, M. E. van Gijn, N. A. Vasilevsky, M. Vlčková, A. Walden, K. Wang, R. Wapner, J. S. Ware, A. A. Wiafe, S. A. Wiafe, L. D. Wiggins, A. E. Williams, C. Wu, M. J. Wyrwoll, H. Xiong, N. Yalin, Y. Yamamoto, L. N. Yatham, A. K. Yocum, A. H. Young, Z. Yüksel, P. P. Zandi, A. Zankl, I. Zarante, M. Zvolský, S. Toro, L. C. Carmody, N. L. Harris, M. C. Munoz-Torres, D. Danis, C. J. Mungall, S. Köhler, M. A. Haendel, and P. N. Robinson, “The human phenotype ontology in 2024: phenotypes around the world,” *Nucleic Acids Research*, vol. 52, pp. D1333–D1346, 11 2023.
- [40] D. J. Thompson, D. Wells, S. Selzam, I. Peneva, R. Moore, K. Sharp, W. A. Tarran, E. J. Beard, F. Riveros-Mckay, C. Giner-Delgado, *et al.*, “A systematic evaluation of the performance and properties of the uk biobank polygenic risk score (prs) release,” *Plos one*, vol. 19, no. 9, p. e0307270, 2024.
- [41] S. F. Friedman, G. E. Moran, M. Rakic, and A. Phillipakis, “Genetic architectures of medical images revealed by registration of multiple modalities,” *Bioinformatics and Biology Insights*, vol. 18, p. 11779322241282489, 2024.
- [42] S. Chaudhary, A. Voigts, M. Bereket, M. L. Albert, K. Schwamborn, E. Zeggini, and F. P. Casale, “Histogwas: An ai-enabled framework for automated genetic analysis of tissue phenotypes in histology cohorts,” *bioRxiv*, pp. 2024–06, 2024.
- [43] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, “Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age,” *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.
- [44] T. J. Littlejohns, C. Sudlow, N. E. Allen, and R. Collins, “Uk biobank: opportunities for cardiovascular research,” *European heart journal*, vol. 40, no. 14, pp. 1158–1166, 2019.

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

- [45] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, *et al.*, “Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches,” *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, pp. 1–10, 2013.
- [46] S. E. Petersen, P. M. Matthews, J. M. Francis, M. D. Robson, F. Zemrak, R. Boubertakh, A. A. Young, S. Hudson, P. Weale, S. Garratt, *et al.*, “Uk biobank’s cardiovascular magnetic resonance protocol,” *Journal of cardiovascular magnetic resonance*, vol. 18, no. 1, pp. 1–7, 2015.
- [47] P. J. Rousseeuw and K. V. Driessen, “A fast algorithm for the minimum covariance determinant estimator,” *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.
- [48] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O’Connell, *et al.*, “The uk biobank resource with deep phenotyping and genomic data,” *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.
- [49] S. Chatterjee, “GlastonburyGroup/ImLatent: Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture.” Zenodo, 2026. <https://doi.org/10.5281/zenodo.20204053>.
- [50] S. Chatterjee, “GlastonburyGroup/CardiacDiffAE_GWAS: Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture.” Zenodo, 2026. <https://doi.org/10.5281/zenodo.20281275>.
- [51] K. Godfrey and J. DiStefano III, “Identifiability of model parameter,” *IFAC Proceedings Volumes*, vol. 18, no. 5, pp. 89–114, 1985.
- [52] Y. Wang, D. Blei, and J. P. Cunningham, “Posterior collapse and latent variable non-identifiability,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 5443–5455, 2021.
- [53] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [54] J. Mbatchou, L. Barnard, J. Backman, A. Marcketta, J. A. Kosmicki, A. Ziyatdinov, C. Benner, C. O’Dushlaine, M. Barber, B. Boutkov, *et al.*, “Computationally efficient whole-genome regression for quantitative and binary traits,” *Nature genetics*, vol. 53, no. 7, pp. 1097–1103, 2021.
- [55] E. H. Livingston and S. Lee, “Body surface area prediction in normal-weight and obese patients,” *American Journal of Physiology-Endocrinology and Metabolism*, vol. 281, no. 3, pp. E586–E591, 2001.
- [56] B. B. Sun, J. Chiou, M. T aylor, C. Benner, Y.-H. Hsu, T. G. Richardson, P. Surendran, A. Mahajan, C. Robins, S. G. Vasquez-Grinnell, *et al.*, “Genetic regulation of the human plasma proteome in 54,306 uk biobank participants,” *BioRxiv*, pp. 2022–06, 2022.
- [57] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, “Second-generation plink: rising to the challenge of larger and richer datasets,” *Gigascience*, vol. 4, no. 1, pp. s13742–015, 2015.
- [58] T. Juliusdottir, “topr: an r package for viewing and annotating genetic association results,” *BMC bioinformatics*, vol. 24, no. 1, p. 268, 2023.
- [59] A. D. Grotzinger, M. Rhemtulla, R. de Vlaming, S. J. Ritchie, T. T. Mallard, W. D. Hill, H. F. Ip, R. E. Marioni, A. M. McIntosh, I. J. Deary, *et al.*, “Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits,” *Nature human behaviour*, vol. 3, no. 5, pp. 513–525, 2019.
- [60] A. R. Quinlan and I. M. Hall, “Bedtools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.
- [61] J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, G. I. of ANthropometric Traits (GIANT) Consortium, D. G. Replication, M. analysis (DIAGRAM) Consortium, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, *et al.*, “Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits,” *Nature genetics*, vol. 44, no. 4, pp. 369–375, 2012.
- [62] Y. He, M. Koido, Y. Shimmori, and Y. Kamatani, “Gwaslab: a python package for processing and visualizing gwas summary statistics,” *Jxiv*, 05 2023.

- [63] M. J. Machiela and S. J. Chanock, “Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants,” *Bioinformatics*, vol. 31, no. 21, pp. 3555–3557, 2015.
- [64] E. Mountjoy, E. M. Schmidt, M. Carmona, J. Schwartzentruber, G. Peat, A. Miranda, L. Fumis, J. Hayhurst, A. Buniello, M. A. Karim, *et al.*, “An open approach to systematically prioritize causal variants and genes at all published human gwas trait-associated loci,” *Nature genetics*, vol. 53, no. 11, pp. 1527–1533, 2021.
- [65] C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol, “Bayesian test for colocalisation between pairs of genetic association studies using summary statistics,” *PLoS genetics*, vol. 10, no. 5, p. e1004383, 2014.
- [66] W. Chen, Y. Wu, Z. Zheng, T. Qi, P. M. Visscher, Z. Zhu, and J. Yang, “Improved analyses of gwas summary statistics by reducing data heterogeneity and errors,” *Nature Communications*, vol. 12, no. 1, p. 7117, 2021.
- [67] G. Csardi and T. Nepusz, “The igraph software,” *Complex syst*, vol. 1695, pp. 1–9, 2006.
- [68] B. Liu, M. J. Gloudemans, A. S. Rao, E. Ingelsson, and S. B. Montgomery, “Abundant associations with gene expression complicate gwas follow-up,” *Nature genetics*, vol. 51, no. 5, pp. 768–769, 2019.
- [69] Z. Fang, X. Liu, and G. Peltz, “Gseapy: a comprehensive package for performing gene set enrichment analysis in python,” *Bioinformatics*, vol. 39, no. 1, p. btac757, 2023.
- [70] J. D. Hocker, O. B. Poirion, F. Zhu, J. Buchanan, K. Zhang, J. Chiou, T.-M. Wang, Q. Zhang, X. Hou, Y. E. Li, *et al.*, “Cardiac cell type-specific gene regulatory programs and disease risk association,” *Science advances*, vol. 7, no. 20, p. eabf1444, 2021.
- [71] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B. W. Penninx, R. Jansen, E. J. De Geus, D. I. Boomsma, F. A. Wright, *et al.*, “Integrative approaches for large-scale transcriptome-wide association studies,” *Nature genetics*, vol. 48, no. 3, pp. 245–252, 2016.
- [72] M. F. Lin, O. Rodeh, J. Penn, X. Bai, J. G. Reid, O. Krasheninina, and W. J. Salerno, “Glnexus: joint variant calling for large cohort sequencing,” *BioRxiv*, p. 343970, 2018.
- [73] J. D. Szustakowski, S. Balasubramanian, E. Kvikstad, S. Khalid, P. G. Bronson, A. Sasson, E. Wong, D. Liu, J. Wade Davis, C. Haefliger, *et al.*, “Advancing human genetics research and drug discovery through exome sequencing of the uk biobank,” *Nature genetics*, vol. 53, no. 7, pp. 942–948, 2021.
- [74] J. D. Backman, A. H. Li, A. Marcketta, D. Sun, J. Mbatchou, M. D. Kessler, C. Benner, D. Liu, A. E. Locke, S. Balasubramanian, *et al.*, “Exome sequencing and analysis of 454,787 uk biobank participants,” *Nature*, vol. 599, no. 7886, pp. 628–634, 2021.
- [75] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, *et al.*, “Twelve years of samtools and bcftools,” *Gigascience*, vol. 10, no. 2, p. giab008, 2021.
- [76] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden, “A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3,” *fly*, vol. 6, no. 2, pp. 80–92, 2012.
- [77] F. Privé, J. Arbel, and B. J. Vilhjálmsson, “Ldpred2: better, faster, stronger,” *Bioinformatics*, vol. 36, no. 22-23, pp. 5424–5431, 2020.
- [78] Y. Xu, D. Vuckovic, S. C. Ritchie, P. Akbari, T. Jiang, J. Grealey, A. S. Butterworth, W. H. Ouwehand, D. J. Roberts, E. Di Angelantonio, *et al.*, “Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease,” *Cell Genomics*, vol. 2, no. 1, 2022.
- [79] S. Zhang, H. Shu, J. Zhou, J. Rubin-Sigler, X. Yang, Y. Liu, J. Cooper-Knock, E. Monte, C. Zhu, S. Tu, *et al.*, “Deconvolution of polygenic risk score in single cells unravels cellular and molecular heterogeneity of complex human diseases,” *bioRxiv*, 2024.
- [80] C. Albiñana, Z. Zhu, A. J. Schork, A. Ingason, H. Aschard, I. Brikell, C. M. Bulik, L. V. Petersen, E. Agerbo, J. Grove, *et al.*, “Multi-pgs enhances polygenic prediction by combining 937 polygenic scores,” *Nature communications*, vol. 14, no. 1, p. 4702, 2023.

Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

- [81] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

End Notes

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 82779. We would like to thank all UKB participants for their invaluable data that make studies like these, and many more, possible.

A.M.V, E.S and F.I acknowledge the support by MUR, grant Department of Excellence 2023-2027 (Department of Mathematics, Politecnico di Milano). This work was carried out using core funding from Human Technopole.

Author Contributions Statement

C.A.G. conceived the study design. S.C. implemented 3DDiffAE, the image processing and result analysis pipelines, ran the GWAS, EWAS, and burden test pipelines, and performed polygenic risk score modelling and analysis. S.O. carried out all genetic analyses, including GWAS, finemapping, colocalisation, and heritability, and performed the post-GWAS enrichment analyses. A.M.V. performed the latent-phenotype association and latent manipulation experiments. E.G. implemented the Regenie GWAS pipeline. A.L., S.S. and N.P. implemented the multi-trait colocalisation pipeline. F.C. provided advice on image segmentation and latent selection and helped with the enrichment analysis. E.S. contributed to and suggested several main analyses, helped with additional requested experiments in revision and contributed to drafting the paper. F.S. provided advice on polygenic risk scores. E.B. and A.V. helped to QC the clinical data. C.A.P. quantified the changes in the manipulation analyses and assisted in the disease grouping criteria. C.A.G., S.O., S.C., A.M.V. and A.V. wrote and edited the manuscript. E.D.A. and F.I. read and reviewed the manuscript.

Competing Interests Statement

The authors declare no competing interests.

Figure legends

Figure 1: Study design and analytical framework. **A.** A 3D diffusion autoencoder (3D DiffAE) is trained on temporal cardiac MRIs from UK Biobank (UKB) to learn a vector of latent phenotypes (z). **B.** The learnt latent phenotypes are associated with specific cardiac outcomes and interpreted via latent space manipulation and visualisation of image reconstructions. **C.** The genetic architecture of latent phenotypes is investigated using genome-wide association studies (GWAS) of common variants, as well as rare-variant and burden-based exome-wide analyses. **D.** Fine-mapping, conditional analysis and colocalisation are employed to assess variant sharing across latent phenotypes, cardiac traits and cardiovascular diseases. **E.** Gene, tissue and single-cell enrichment analyses are carried out to understand underlying target genes. **F.** Polygenic risk scores (PRS) are built to demonstrate the prognostic predictive value of latent phenotypes. Created in BioRender. Ometto, S. (2026) <https://BioRender.com/mkdusos>.

Figure 2: Comparative analysis of a 3D Variational Autoencoder (VAE) and our 3D Diffusion Autoencoder (DiffAE). **A.** Cardiac MRI reconstructions for two example subjects are shown for the two models, illustrating qualitative differences. **B.** Quantitative reconstruction performance for $n=10,000$ randomly chosen subjects from the test set, comparing Structural Similarity Index (SSIM) of DiffAE ($t_{step} = 20$) and VAE models. Each data point represents one subject. The violin shape shows the kernel density estimate of the distribution. The embedded box spans the interquartile range (IQR; 25th–75th percentile); whiskers extend to the minimum and maximum observed values. The red dot and annotated value (μ) denote the mean SSIM. Source data are provided as a Source Data file. **C–D.** Associations between 128 latent phenotypes from single-seeded VAE (**C**) versus single-seeded DiffAE (**D**) and diseases, tested with a two-sided t test: volcano plots of effect sizes vs statistical significance. The dashed horizontal line indicates the Bonferroni-adjusted P-value threshold: $p < 2.8 \times 10^{-5}$ ($n=47849$ individuals). Source data are provided as Supplementary Data 1.

Figure 3: Latent phenotypes’ associations with diseases and continuous traits. Latent manipulation of left ventricle end diastolic volume and myocardial wall thickness. **A–B** Associations between latent

phenotypes and diseases (**A**) and continuous traits (**B**), tested with a two-sided t test: volcano plot of effect sizes vs statistical significance. The dashed horizontal line indicates the Bonferroni-adjusted P -value threshold: $p < 1.96 \times 10^{-5}$ for diseases ($n=47849$ individuals), $p < 9.8 \times 10^{-6}$ for continuous traits ($n=27446$ individuals). The top three continuous traits for each category showing the most significant associations with latent phenotypes are labelled. Source data are provided as Supplementary Data 2 and 3. LV: left ventricle, RV: right ventricle, LA: left atrium, RA: right atrium. **C-D** Effects of latent manipulation on left ventricle end diastolic volume (LVEDV) (**C**) and myocardial wall thickness (**D**) for two subjects, with increasing manipulation magnitude from left to right. Quantitative measurement of the anatomical changes was performed on the manipulated reconstructions.

Figure 4: GWAS for 182 DiffAE latent phenotypes. **A.** Manhattan plot summarising genome-wide associations across 182 latent phenotypes in the GWAS discovery cohort ($n=47740$ individuals). Associations were tested using REGENIE under an additive genetic model (two-sided). Blue horizontal line: genome-wide significance (5×10^{-8}). Red horizontal line: study-wide significance (2.27×10^{-9} ; with Bonferroni correction for 22 effective independent tests). The 44 lead SNPs are annotated with the nearest or overlapping gene and highlighted according to the most frequent category of previously reported associations. Source data are provided as summary statistics (see Data Availability). **B.** Dotplot showing the categories of traits previously associated with our lead variants, according to the GWAS Catalog. "MRI latents" refers to the unsupervised GWAS of cross-modal representations from the study by Radhakrishnan et al. [19] Previously unreported loci are labelled as novel. Source data are provided as a Source Data file.

Figure 5: Gene set and tissue enrichments, Transcriptome-Wide Association Study (TWAS). **A.** EnrichR over-representation analysis of MAGMA-prioritised genes using a two-sided Fisher's exact test; bars show significant terms with Benjamini-Hochberg adjusted P -values (log scale). Source data are provided as Supplementary Data 16. **B.** MAGMA gene-property analysis testing association between gene-level signals and GTEx tissue expression ($n=54$ tissues) using one-sided linear regression model. Only tissues with $-\log_{10}(P) > 0.01$ are shown. The dashed line represents the Bonferroni-corrected threshold: $-\log_{10}(0.05/54)$. Source data are provided as Supplementary Data 18. **C.** TWAS heatmap of 82 genes prioritised by FUMA, colored by Z-scores. Associations were tested using two-sided Z-tests, with significance defined by Bonferroni-adjusted P -values; asterisks denote significant gene-tissue pairs. Source data are provided as Supplementary Data 19. SCAT: subcutaneous adipose tissue, VAT: visceral adipose tissue, AOR: aorta, COR: coronary artery, TIB: tibial artery, AA: heart atrial appendage, LV: heart left ventricle, LIV: liver, LUN: lung, SM: skeletal muscle, PAN: pancreas, BLD: blood.

Figure 6: Gene expression enrichment in heart scRNAseq. **A.** Uniform manifold approximation and projection (UMAP) visualisation of human heart single-nucleus RNA sequencing (snRNA-seq) data, illustrating 12 distinct cellular populations ($n=704296$ cells). **B.** Differential expression of MAGMA-prioritised genes related to latent phenotypes across cell types, comparing atrial and ventricular cardiomyocytes to other cell types using a two-sided t -test with overestimated variance. Source data are provided as Supplementary Data 20. **C.** Detailed UMAP visualisations of human heart snRNA-seq data, coloured by gene expression of *MYH6*, *MLF1*, *CCDC141*.

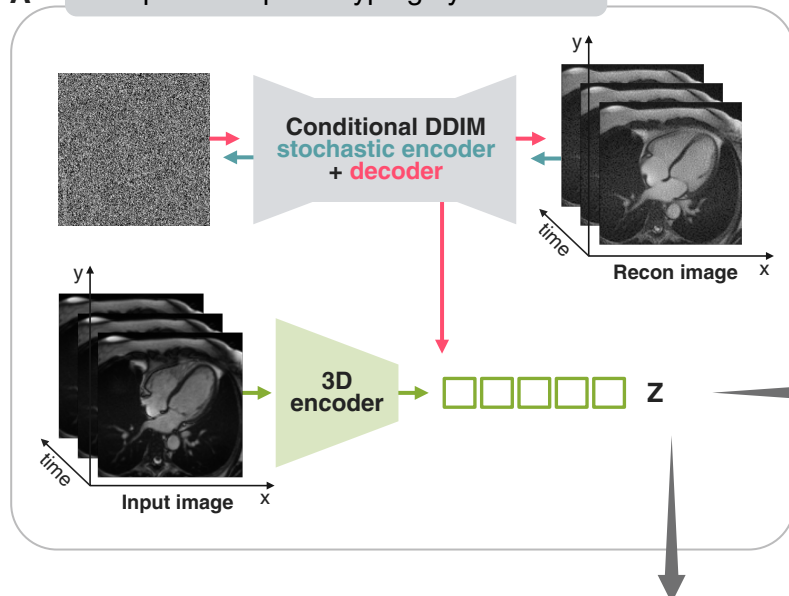
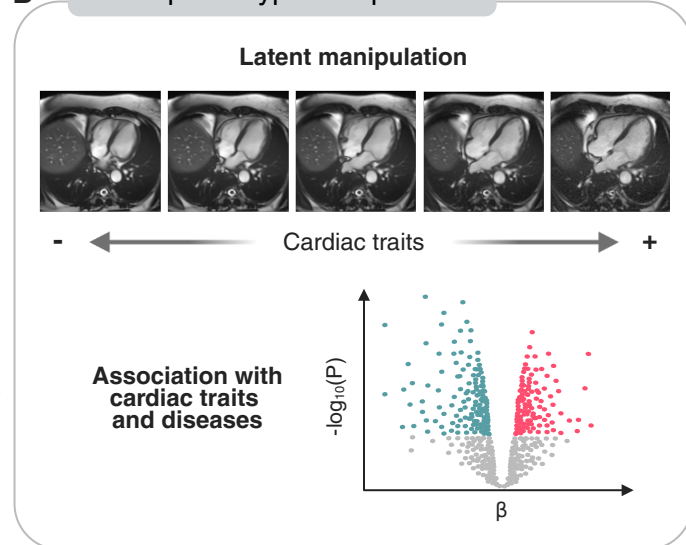
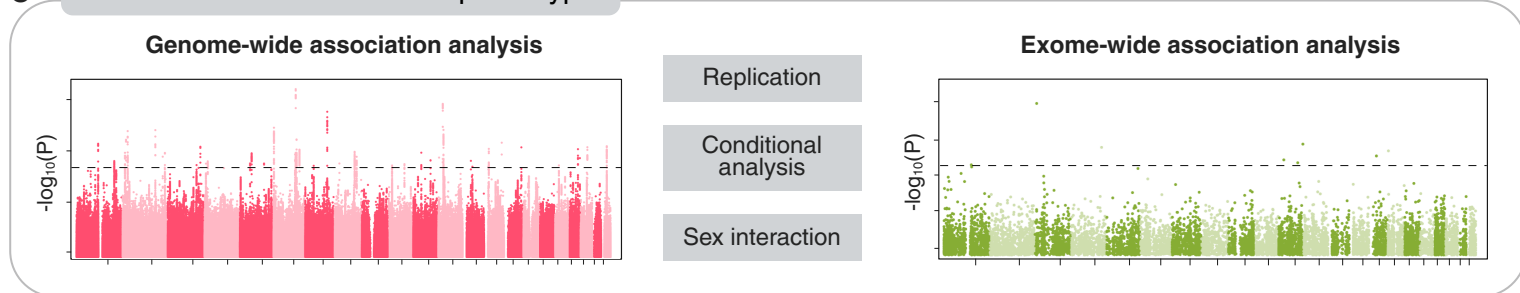
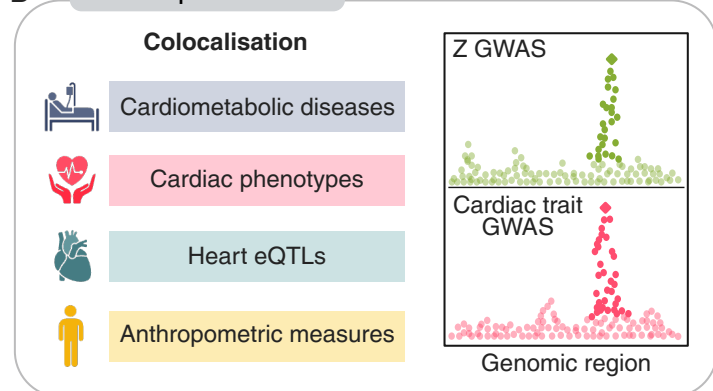
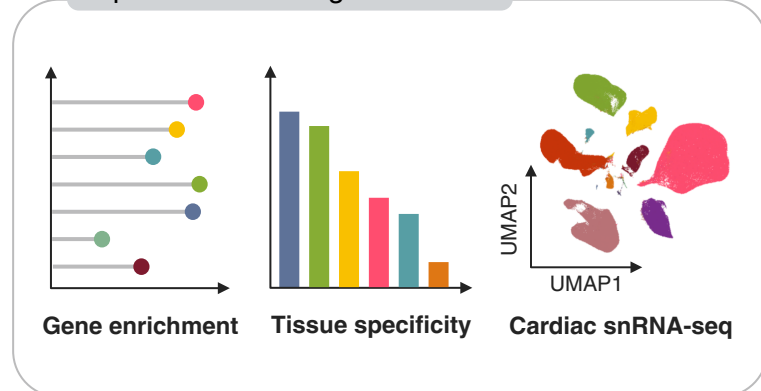
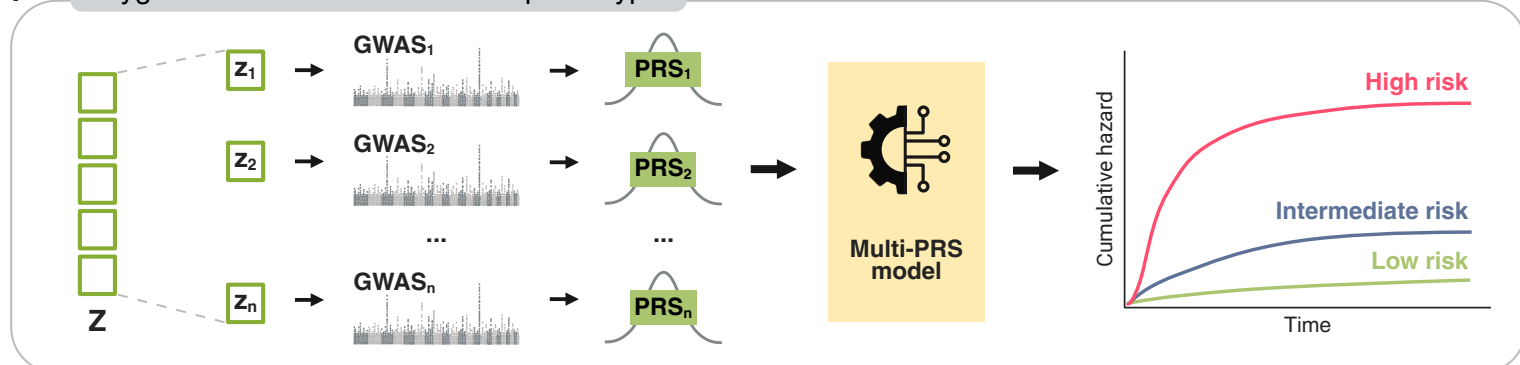
Figure 7: Colocalisation analysis for the 44 lead SNPs from latent phenotypes GWAS across 54 GWAS of relevant traits and 12 GTEx tissues. **A.** Upset plot showing overlap of latent phenotypes lead variants with associations across cardiac eQTLs, non-cardiac eQTLs, cardiac disease GWAS, cardiac trait GWAS (ECG and blood pressure traits, MRI-derived heart and aorta phenotypes), and non-cardiac trait GWAS. Colocalisation was assessed using a Bayesian framework, with significant pairs having posterior probability $PP.H_4 > 0.8$. Source data are provided as a Source Data file. **B.** Heatmap of colocalisation results, coloured by posterior probability of a shared causal variant (H_4). Latent phenotypes lead SNPs (x-axis) are annotated by closest gene and grouped by chromosome; GWAS traits and eQTL tissues (y-axis) include only those with at least one significant colocalisation. Source data are provided as Supplementary Data 22. **C.** Locuszoom plots for the *HSPA4* locus on chromosome 5, showing GWAS association signals (log scale) across genomic position for colocalising traits. The lead SNP is annotated, and the variants are coloured by linkage disequilibrium (LD). Sample sizes for each GWAS are provided in Supplementary Data 21. Source data availability: latent phenotypes GWAS are provided as summary statistics (see Data Availability). GWAS summary statistics are publicly available via the sources listed in Supplementary Data 21. eQTL summary statistics are available from the GTEx database upon request.

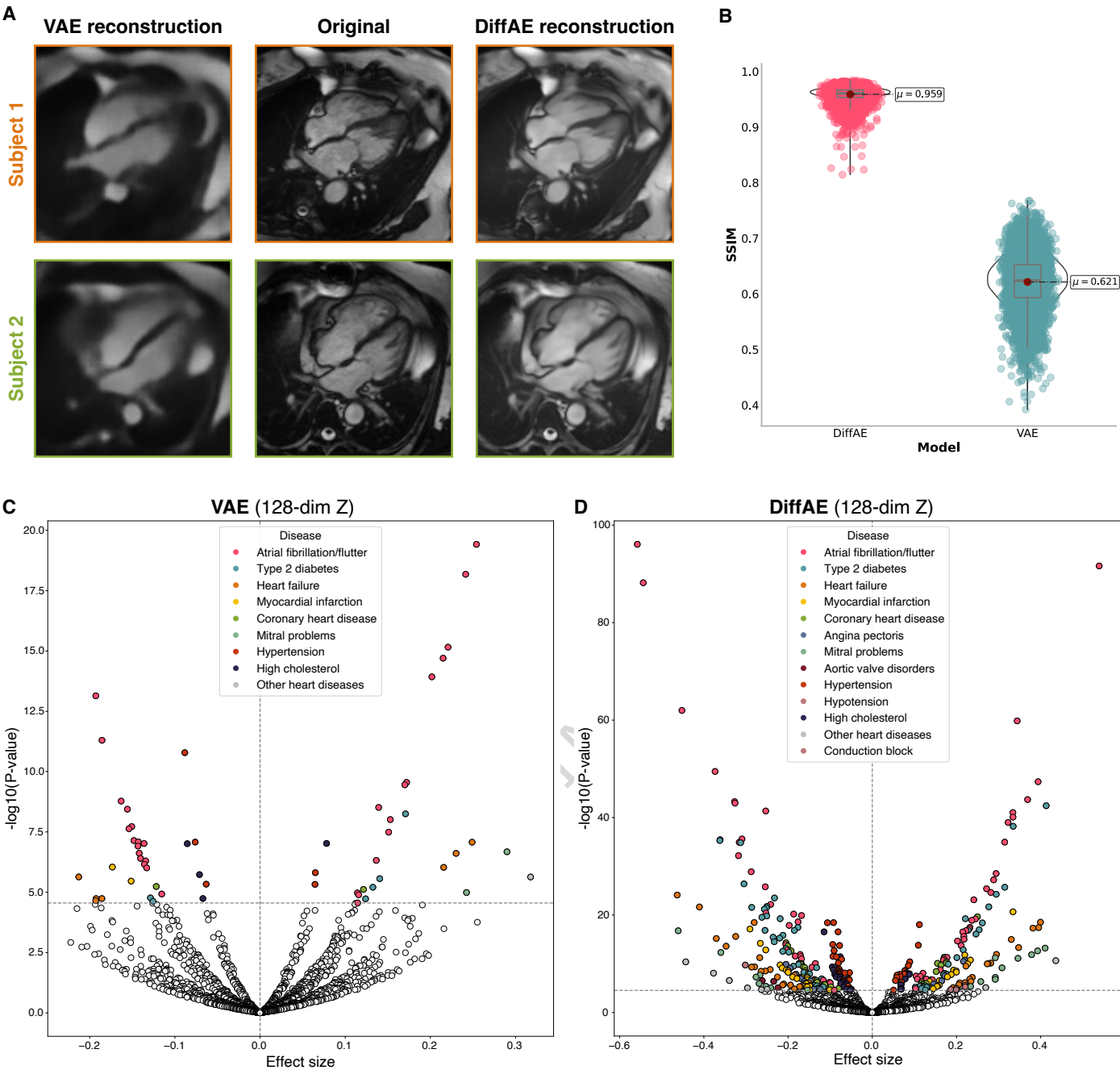
Figure 8: Rare variant gene-based association analysis using SKATO-ACAT. Circular Manhattan plot showing gene-based association results in the discovery cohort ($n=46134$ individuals) using SKATO-ACAT two-sided test across three variant masks (see Methods): M1 loss-of-function (LoF) variants only, M2 LoF and likely deleterious missense variants (5/5 prediction score), M3 LoF and possibly deleterious missense variants ($\geq 1/5$ prediction score). Radial position corresponds to statistical significance ($-\log_{10}(P\text{-value})$). Genes surpassing the significance threshold

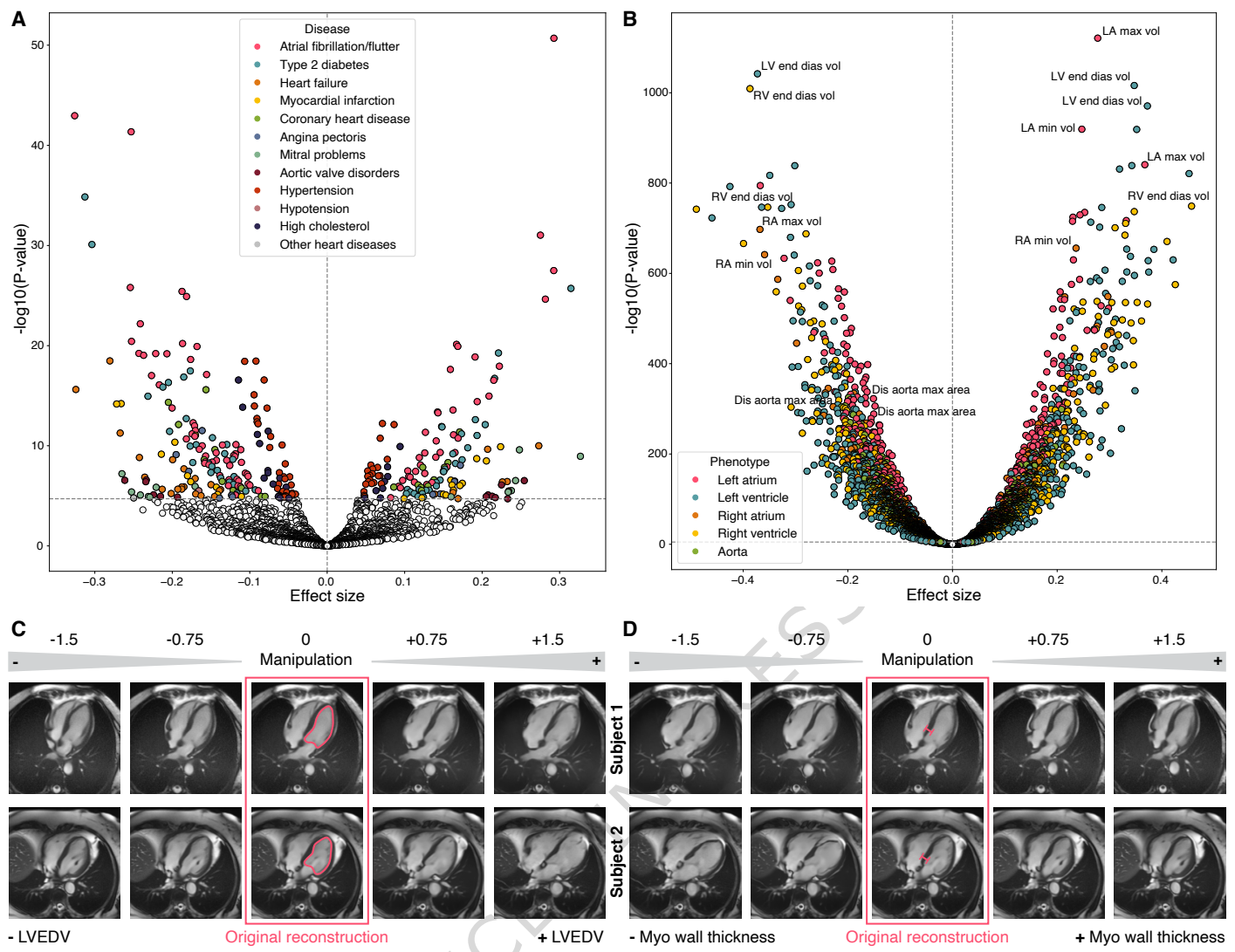
Hundreds of cardiac MRI traits derived using 3D diffusion autoencoders share a common genetic architecture

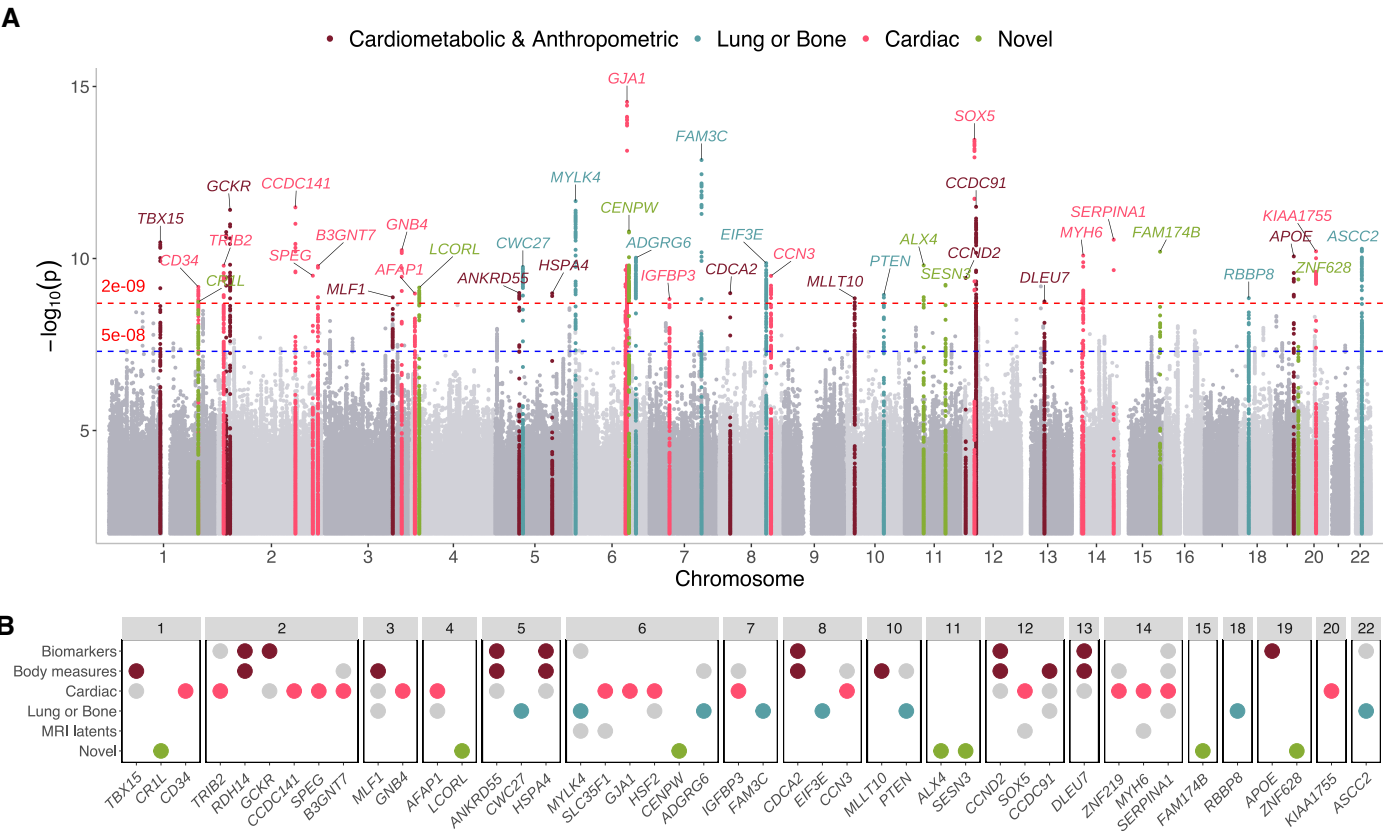
of 4.7×10^{-7} (Bonferroni correction, dashed line) are annotated and highlighted with larger markers. Source data are provided as summary statistics (see Data Availability).

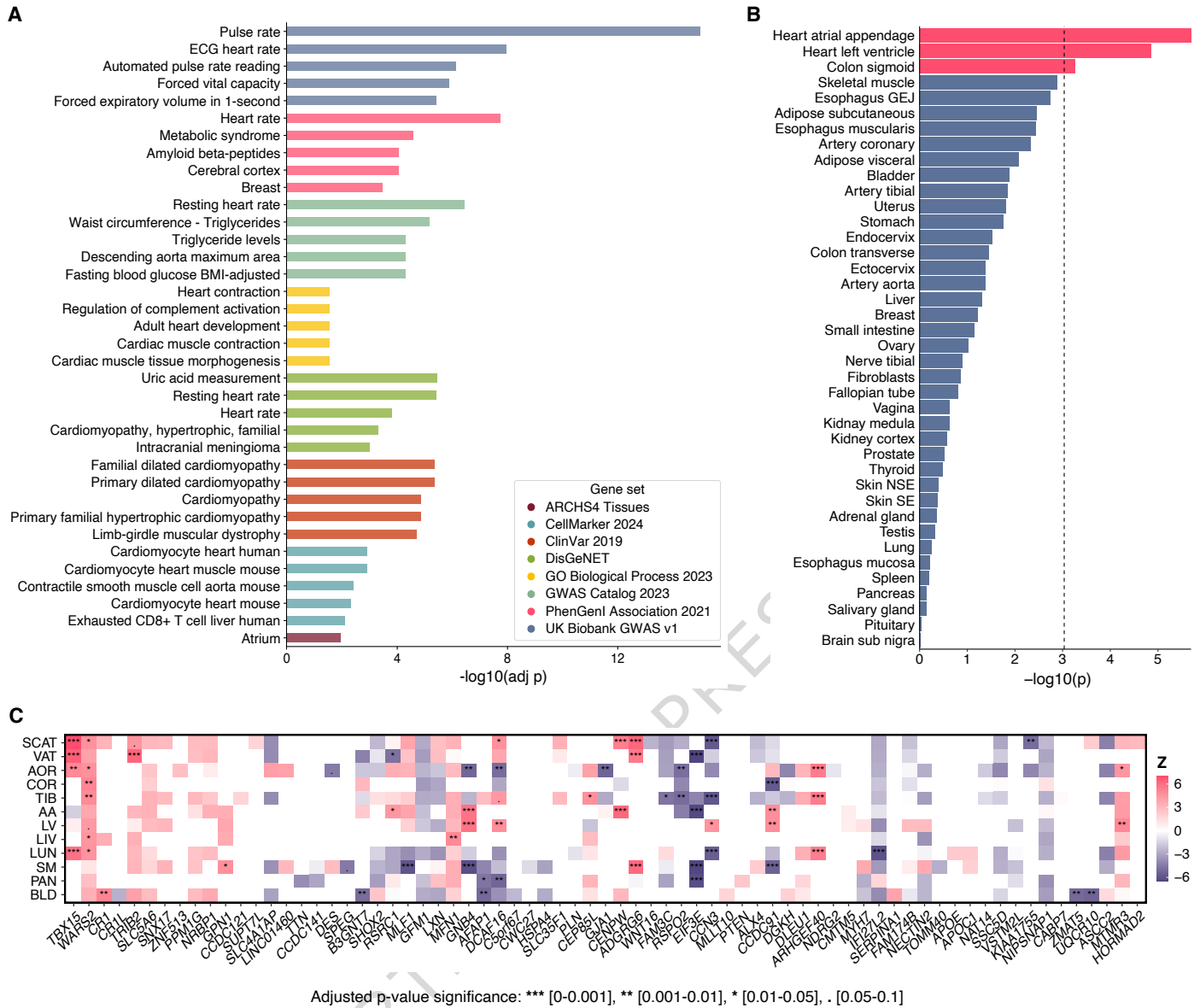
Figure 9: Stratification of disease risk by imaging-derived PRS. **A, D:** Prevalence of atherosclerotic conditions and metabolic syndrome, stratified by sex across PRS percentiles ($n = 314,774$, where $n_{female} = 171,312$ and $n_{male} = 143,462$). Data points show the observed disease prevalence (proportion of cases) within each multi-PRS percentile group. Error bars represent 95% confidence intervals calculated using the normal approximation of the binomial proportion. The centre of each error bar is the observed proportion $\hat{p} = n_{cases}/n_{total}$ within the respective percentile stratum. **B, E** Estimated cumulative number of new diagnoses over ten years for top, mid, and low-risk PRS groups, derived from disease-specific generalised linear models ($n_{atherosclerotic} = 189,705$ and $n_{metabolic} = 220,497$). **C, F** Nelson–Aalen cumulative hazard curves up to ten years post baseline for the same cohorts as B and E, respectively. Shaded bands represent 95% confidence intervals derived from the Nelson–Aalen estimator (pointwise, on the cumulative hazard scale). The number of individuals at risk and total events per group are provided in the source data and are provided as a Source Data file.

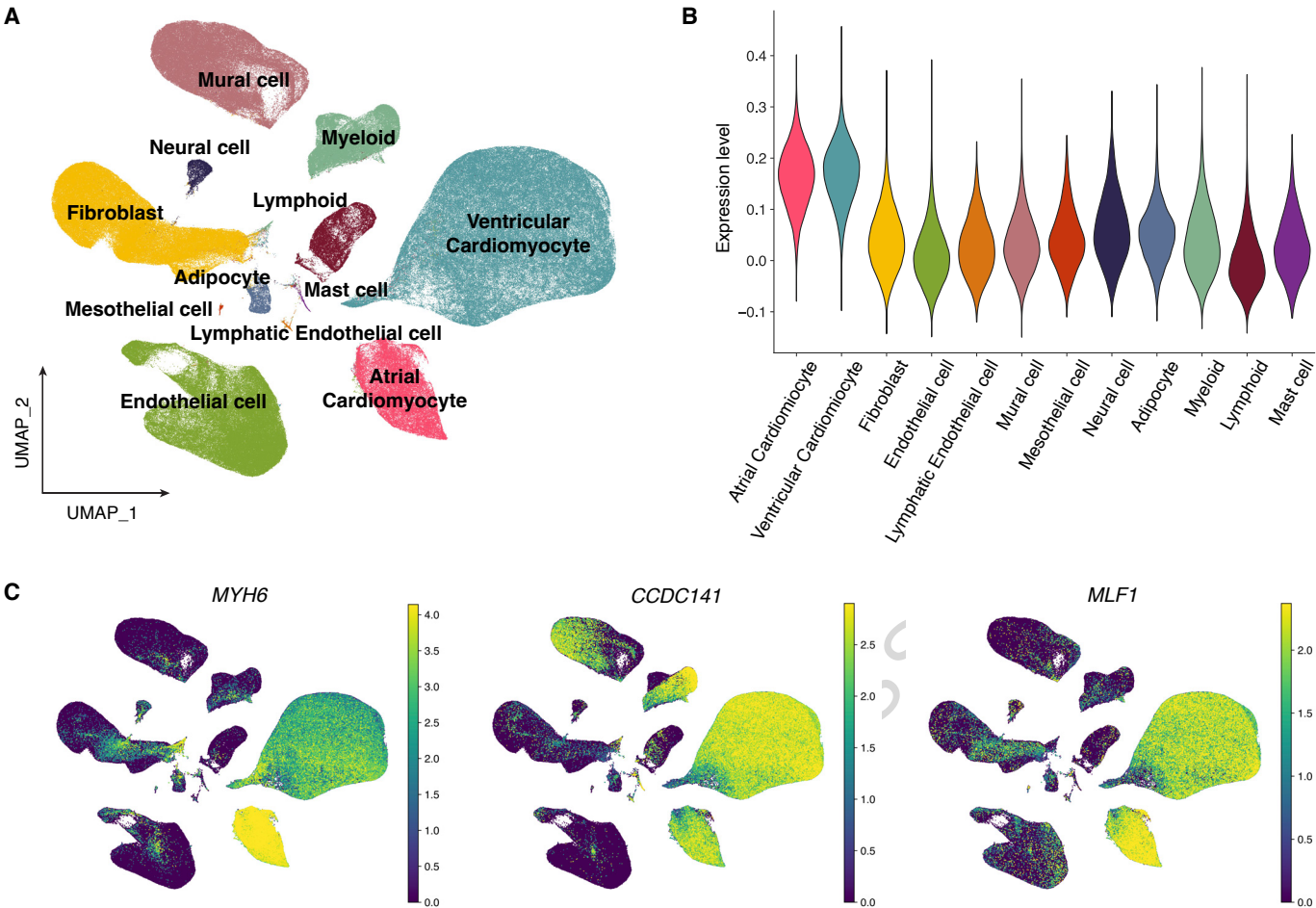
A Unsupervised phenotyping by 3D DiffAE**B** Latent phenotype interpretation**C** Genetic architecture of latent phenotypes**D** Variant prioritisation**E** Exploration of biological function**F** Polygenic Risk Scores from latent phenotypes

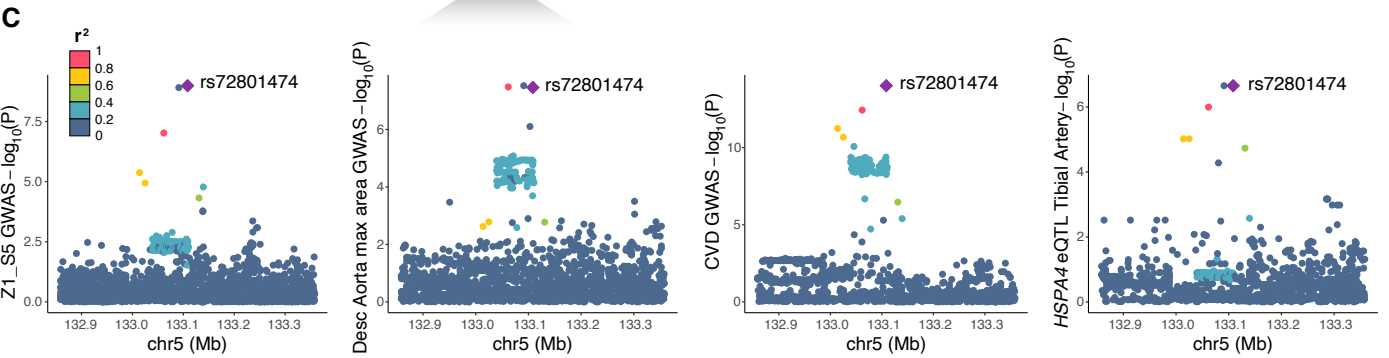
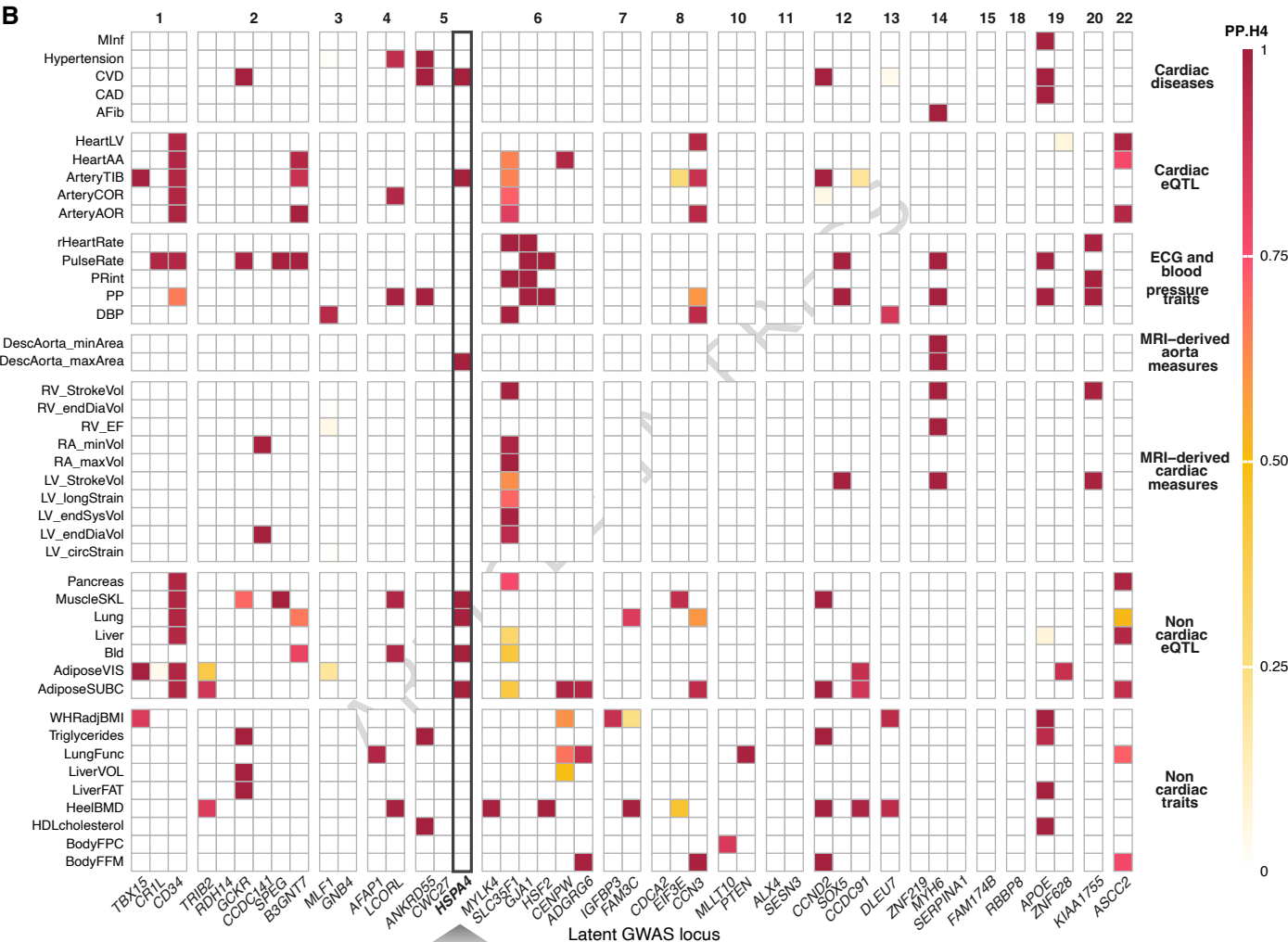
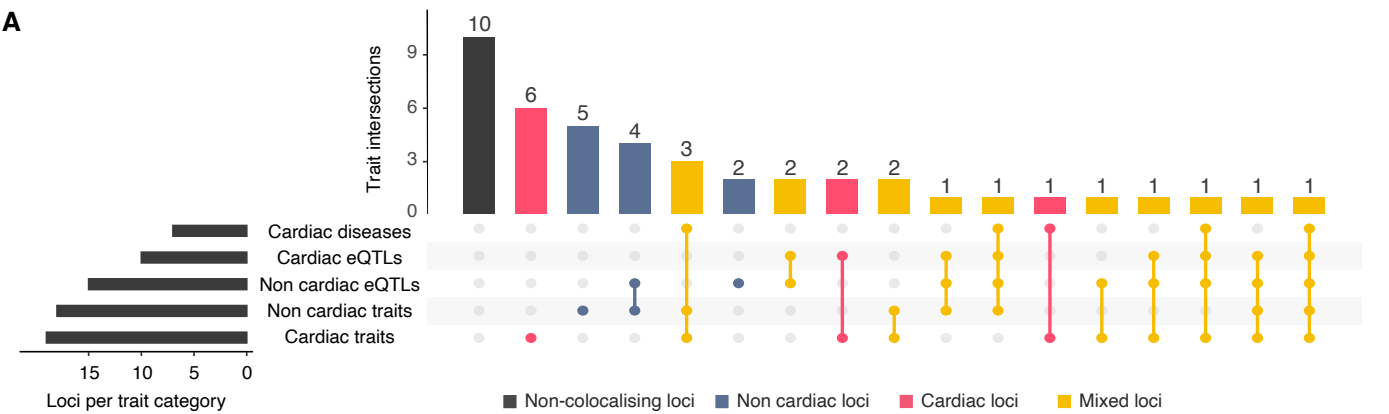


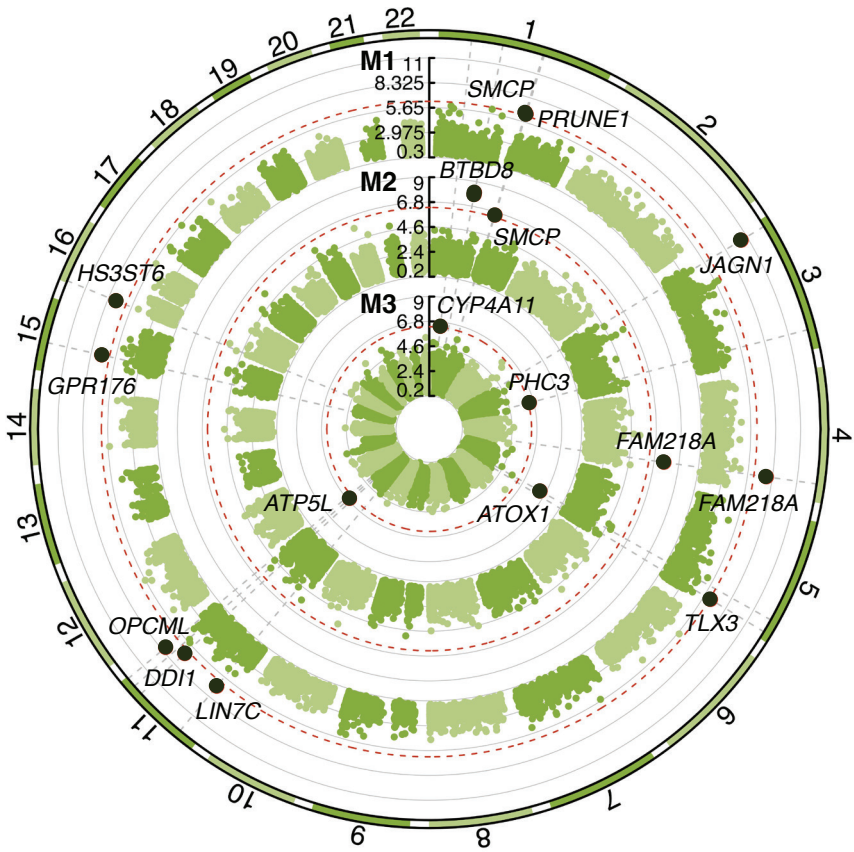




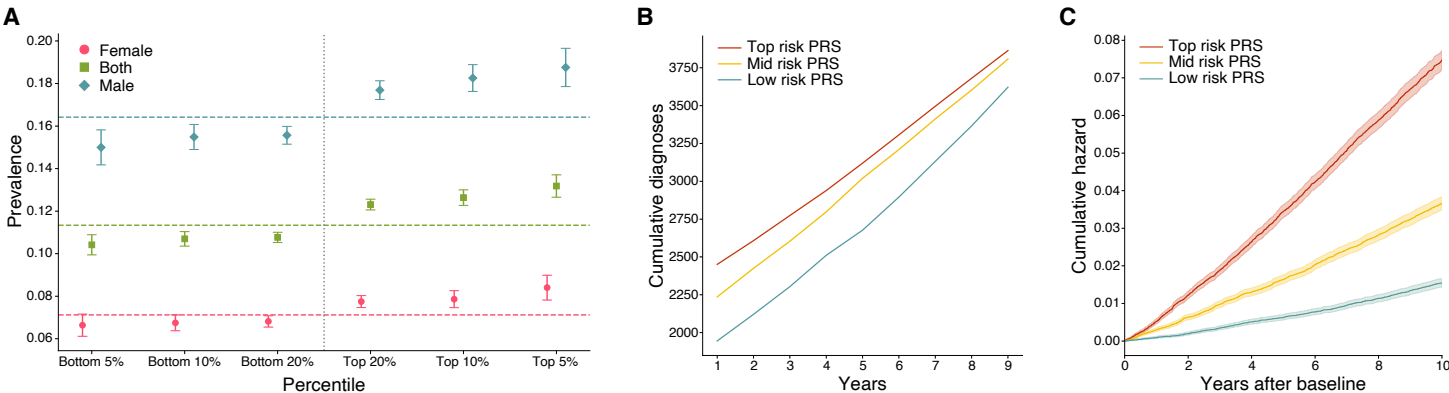








Atherosclerotic PRS: Angina pectoris, Myocardial infarction, Coronary heart disease



Metabolic syndrome PRS: Hypertension, High cholesterol, Type 2 diabetes

