UNIVERSITY OF LEEDS

Deep cardiac phenotyping with applications in imaging genetics

Rodrigo Bonazzola

Submitted in accordance with the requirements for the degree of PhD in Computer Science

The University of Leeds

Faculty of Engineering and Physical Sciences

School of Computer Science

March, 2024

Contents

Abstract

The emergence of large prospective biobanks with paired imaging and genetic data, such as the UK Biobank (UKB), has enabled the investigation of image-derived phenotypes in genetic association studies, aiming to identify genetic variants that drive phenotypic differences observable in medical images. Traditionally, these studies have focused on handcrafted phenotypes—traits known to be clinically relevant. However, the unprecedented sample sizes now available facilitate data-driven phenotyping, offering the potential to uncover more subtle phenotypic variations that can enhance genetic discovery.

In this work, we focus on cardiovascular cine magnetic resonance (CMR) imaging and apply state-of-the-art image processing techniques to generate 3D meshes representing the myocardium from over 50,000 UKB participants. We then leverage geometric deep learning to learn unsupervised representations of these shapes while explicitly preserving their topology.

First, we phenotype static left-ventricular meshes at end-diastole (ED) in an unsupervised manner using convolutional mesh autoencoders (CoMA), a well-established shape analysis technique. We then introduce a novel methodology to capture dynamic patterns across the full cardiac cycle, inherently incorporating the periodicity of cardiac motion. These learned representations are subsequently analyzed in genome-wide association studies (GWAS) to identify genetic variants linked to both static and dynamic phenotypes. Additionally, we demonstrate the effectiveness of an ensemble-based phenotyping framework in improving discovery power. Finally, we perform an in-depth genetic analysis to interpret our findings in the context of existing biological evidence and identify potential candidate genes underlying these associations.

Chapter 1

Introduction

This thesis investigates the genetic basis of structural and functional cardiac phenotypes, derived from cardiovascular magnetic resonance (CMR) imaging, by leveraging recent advances in automatic segmentation techniques and representation learning. In this chapter, we aim to provide an overview of the background and motivation behind this study, our aims and objectives and our contributions. Finally, we summarise the structure of this dissertation.

1.1 Background and rationale

To advance therapeutic development for diverse pathological conditions, it is indispensable to understand the underlying biological mechanisms that lead to these conditions, so that we can effectively act on them. In this sense, genome-wide association studies (GWAS), which aim to identify associations of genotypes with phenotypes, have enabled remarkable advances in the understanding of a great diversity of traits, for instance anthropometric, behavioural and disease traits [102, 103, 1]. Genetic associations constitute a unique piece of information due to their incontrovertibly causal nature, unlike many statistical associations one may encounter in epidemiology. In spite of this remarkable fact, i.e. the causality linked to genetic associations, they constitute a double-edged sword: *also* by virtue of being the most fundamental layer of biology, it is usually difficult to make sense of the mechanisms that lead to observable traits due to the numerous chain links present in the causal cascade, making it challenging to transform this information into useful knowledge.

Endophenotypes, or *intermediate phenotypes*, are biological traits that bridge the gap between

genetic variation (for example, single-nucleotide polymorphisms) and high-level phenotypes, such as diagnoses or clinical symptoms [60]. Endophenotypes are closer to genetics in the causal cascade, potentially increasing detection power of genetic factors. Examples of endophenotypes are biochemical traits such as metabolite concentrations or, of special focus in this dissertation, anatomical or functional image-derived phenotypes (IDPs).

In general terms, phenotypes of higher level are less costly, since they rely on clinical observations, interviews or simple measurements. However, these traits are also more genetically heterogeneous and hence, the knowledge that can be extracted from genetic associations studies is more limited. Complex diseases are caused by multiple intermediate phenotypes involved in their pathogenesis and, very often, each one of these intermediate phenotypes has a component of quantitative inheritance: the closer we get to the "underlying biology", the more likely we are to find more specific associations.

The best examples of endophenotypes to date come from the area of psychiatric genetics, for which neuroimaging has played an important role in unveiling the genetic basis of a number of conditions, such as schizophrenia. However, the concept is applicable to other areas, for example cardiology, as we will study in this thesis.

It is thus of great interest to unravel the genetic basis of phenotypes and, in particular, that of IDPs. However, unless extremely large genotype-phenotype effects were to be expected, an association study of IDPs would require the recruitment, genotyping and image scanning of tens of thousands of subjects, which represents an immense undertaking. Fortunately, the advent of large-scale biobanks including imaging data in the last decade, has come to satisfy this need. In particular, UK Biobank (UKB), which has recruited over 500,000 subjects between 2006 and 2009 and keeps collecting data at present, is hitherto the largest of the biobanks with linked genetic and imaging data. UKB is planning to conduct whole-body MRI imaging on around 20% of the total biobank population, with almost 70,000 (14%) already scanned at the time of this writing, and over 5,000 with a second imaging visit.

Another key challenge is of *methodological* nature, and refers to the extraction of relevant quantitative phenotypic information from the image. This problem lies at the core of this thesis. An image, being a collection of pixel (or voxel) intensities arranged in a 2D (or 3D) grid, is evidently not suitable as a phenotype to be tested in its raw form. Indeed, it normally contains significant pose and appearance variations which hold no anatomical value. On the

other hand, anatomical correspondence between two images is generally not preserved at the level of pixel or voxel locations.

Coincidentally with the advent of large biobanks with linked genetic and imaging data, automatic segmentation approaches have enabled the efficient processing and quantification of large amounts of imaging data, allowing to readily delineate structures of interest in an image. These advances have made possible the study of the genetic basis of IDPs, the study of which constitutes the field of *imaging genetics*.

Among the imaging modalities present in the UKB, cine cardiovascular magnetic resonance (cine-CMR) imaging provides detailed information about the morphological and functional characteristics of the heart. Since 2019, several studies have been published leveraging these data to unveil a great number of novel genetic associations with cardiac phenotypes of structure and function. However, these phenotypes have been limited to handcrafted quantities, typically chamber volumes and volume-derived quantities such as ejection fraction. We argue that the unprecedented sample size allows for a different approach: leveraging recent machine learning techniques for representation learning, phenotypes can be derived in an unsupervised, data-driven manner. This approach, which we study in this dissertation, would in turn unveil previously unstudied patterns of variation, hence permitting the discovery of additional genetic variation linked to cardiac morphology and function.

1.2 Aims and objectives of this thesis

The general aim of this thesis is to leverage the unprecedented linked genotypic and phenotypic data from UKB to answer the question of how genetic variation induces changes in cardiac morphology and function across the population.

We aim to advance present knowledge on this topic by proposing novel data-driven phenotyping techniques that provide a deeper characterization of the full breadth of phenotypic variability, when compared with traditional handcrafted phenotype extraction based on clinical function indices.

More specifically, our objectives are to use representation learning techniques on a 3D mesh representation of the cardiac chambers (obtained by segmentation of the CMR with state-of-the-art methods) to then test the quantitative phenotypes thus obtained for association with

genetic variants. For this, we leverage state-of-the-art automatic segmentation approaches, that we apply at scale to UKB data, as well as recent geometric deep learning techniques for shape representation learning. In a first stage, we study static cardiac meshes (3D-manifold-like objects), to then investigate sequences of cardiac meshes across the cardiac cycle (4D-manifold-like objects).

1.3 Contributions

Broadly, this thesis makes contributions both on the methodological side of phenotyping of 3D- and 4D-manifold-like objects derived from images, and more specifically on the understanding of cardiac genetics. Among the methodological contributions are:

- A framework based on deep learning for performing phenotyping and GWAS of static anatomical structures, which leverages to a fuller extent the morphological information contained in the images as compared to previously applied methods.
- A deep-learning-based method for extracting disentangled static and dynamic phenotypes from temporal sequences of cardiac 3D meshes across the cardiac cycle, explicitly leveraging the periodic nature of cardiac motion.

For the particular application of cardiac imaging genetics, we have advanced knowledge of the genetic basis of cardiac image-derived phenotypes. In particular:

- The number of genetic loci implicated in cardiac anatomy and motion has been increased substantially.
- A more refined knowledge on the effect of different genetic variants on left-ventricular morphology has been obtained.
- Distinct loci have been identified for static and dynamic cardiac phenotypes.

In addition, side contributions that were auxiliary steps for the work in this thesis:

- The development of a GWAS pipeline encompassing the data pre-processing, GWAS execution and downstream analysis steps.
- CardioMesh: an open-source Python library for cardiac mesh processing.
- A web application open to the scientific community, to explore the findings from our work.

1.4 Statement on reproducibility

Special attention was given to guaranteeing that the work presented in this thesis is reproducible, as well as to give transparency to the whole process by allowing the interested researcher to examine the full set of results, including the intermediate ones.

To achieve this goal, we employed state-of-the-art tools in the machine learning and deep learning ecosystem. Code for published work is publicly available on GitHub. Also, we built a web application that allows to navigate the data that was produced during this work. We encourage the readers of this thesis to make use of this application.

1.5 Structure of the dissertation

This thesis is composed of eight chapters.

- **Chapter** 2 provides the necessary biological background for the rest of the thesis as well as a literature review, covering the fundamental concepts of genetics and genetic association studies, an introduction to cardiac structure and physiology, and recent studies on the genetics of image-derived cardiac phenotypes. Finally, we overview the idea of applying deep learning methods to imaging genetics, both for segmentation and for phenotyping.
- **Chapter** 3 provides an overview of the imaging data used in this work, both the raw CMR data from UKB as well as the intermediate 3D mesh data generated therefrom via automatic segmentation.
- **Chapter** 4 provides details on the methodologies for shape representation learning used in this thesis, in particular, shape principal component analysis (PCA) and convolutional mesh autoencoders (CoMA).
- **Chapter** 5 contains the results of my study of the genetic basis of traditional CMR-derived phenotypes, as well as their relation to demographic variables.
- **Chapter** 6 and **Chapter** 7 contain most of the novel work in this dissertation. Chapter 6 focuses on the unsupervised phenotyping of static left-ventricular meshes, whereas the second proposes a novel methodology for unsupervised dynamic phenotyping of the heart, which we implement across the four chambers of the heart.
- Finally, **Chapter** 8 summarises the key findings in this thesis and discusses some limita-

tions of our work, as well as ideas for future research.

Chapter 2

Biological background

In this chapter, I provide the general biological background for this work. It is composed of four sections:

- In **Section 2.1**, the basic concepts from genetics in general, and statistical genetics in particular, that support our work are covered. The latter part includes genome-wide association studies (GWAS) and several downstream analysis strategies for making sense of GWAS results.
- **Section 2.2** provides an overview of the relevant aspects of the structure and functioning of the heart. While not exhaustive, this section aims to provide enough background to the reader with little or no biological training, to follow the discussion regarding the genes associated with cardiac phenotypes in later chapters.
- **Section 2.3** reviews succinctly what is known about the genetic basis of cardiac structure and function in health and disease, with a special focus on recent studies using cardiac imaging data.
- **Section 2.4** describes past uses, challenges and future opportunities of deep learning for the purpose of genetic discovery, introducing the idea of using neural networks as *phenotypers* as an overarching theme in this thesis.

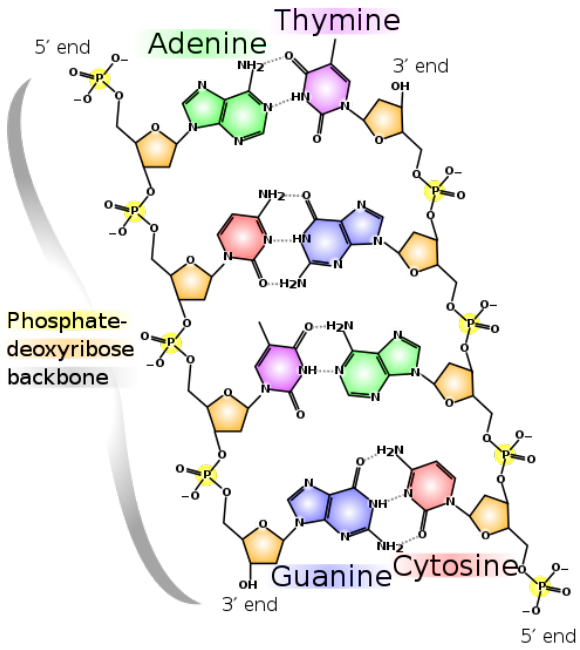

Figure 2.1: Structure of the double chain of DNA (from Wikipedia)

2.1 Fundamental concepts of genetics

2.1.1 Inheritance, DNA structure and the genetic code

The phenomenon of inheritance has been recognized since ancient times, but it was Gregor Mendel who first investigated it systematically, formulating the basic laws of heredity through controlled experiments on pea plants. Many decades later, a major milestone in the molecular understanding of inheritance was achieved in 1944, with the identification of DNA as the genetic material and the characterization of its composition from four nucleotide bases: adenine (A), thymine (T), cytosine (C), and guanine (G). Less than a decade later, in 1953, the double helix structure of DNA was elucidated by James Watson and Francis Crick, building on X-ray diffraction data produced by Rosalind Franklin and Maurice Wilkins.

DNA is composed of two long chains of nucleotides twisted into a double helix and held together by hydrogen bonds between complementary base pairs. Each nucleotide consists of three components: a sugar molecule (deoxyribose), a phosphate group, and a nitrogenous base (see Figure 2.1). Adenine pairs up with thymine (A-T), while guanine pairs up with cytosine (G-C), forming the complementary base pairs. This complementary base pairing is essential for DNA replication and maintaining the genetic code.

Shortly after the discovery of the double-helical structure of DNA in 1953, efforts began to be made to understand how information flows from genes to functional elements. This is known as the *genetic code*: the set of rules by which information encoded in DNA sequences is translated into proteins, which are the functional molecules in cells. The genetic code is universal across all organisms, with a few minor exceptions.

The process of protein synthesis, or translation, involves several key players, including messenger RNA (mRNA), transfer RNA (tRNA), and ribosomes. During translation, mRNA carries the genetic information from DNA to the ribosome, where tRNA molecules deliver the corresponding amino acids according to the sequence of codons on the mRNA. These amino acids are then linked together to form a protein.

The genetic code is read in triplets of nucleotides, known as *codons*. Each codon specifies a particular amino acid, the building blocks of proteins. There are 64 ($= 4^3$) possible codons, but they encode only 20 amino acids, meaning that most amino acids are specified by more than one codon. In addition, three codons function as stop signals, marking the termination of protein synthesis.

2.1.2 Single-nucleotide polymorphisms and indels

After the completion of the Human Genome Project in 2003 [26], studies such as 1000 Genomes [27] and HapMap [43] have committed themselves to characterising genetic variation in humans, across different populations. At the vast majority (around 99%) of genomic sites, every human carries the same pair of nucleotides on both chromosomal homologs. The remainder encodes much of the diversity among humans, including differences in disease susceptibility.

Single-nucleotide polymorphisms (SNPs) are genome positions at which there are at least two alleles[1] where each allele appears in a significant portion of the human population. For practical purposes, the smallest of these fractions, the minor allele frequency (MAF), can be taken to be 1–5%, with recent large sample sizes allowing one to consider even lower frequencies, down to 0.1%. It is estimated that there are around 10 million SNPs in the human genome with a MAF threshold of 1%.

On the other hand, indels, short for insertions and deletions, represent another type of common

[1]Typically, only two different nucleotides can be present at a given site, since they originate through rare mutation events. These are called bi-allelic SNPs and are by far the most common. Note, however, that multi-allelic SNPs do exist as well.

genetic variation characterised by the insertion or deletion of bases in an organism's genome. These variations typically involve only a few base pairs, although larger events may also occur. Note that both types of events represent the same underlying form of variation, and declaring a specific variant as an insertion or a deletion depends on the choice of reference genome. Namely, on whether the reference genome contains the inserted or deleted sequence; if it does, then the variation will be called a deletion regardless of the actual mechanism that took place.

2.1.3 SNP microarrays

SNP microarrays are a tool used in genetic analysis to detect variations in the DNA sequence at the single nucleotide level [55].

Despite differences in the specific technologies, SNP arrays rely on the principle according to which nucleotide bases bind to their complementary partners: A binds to T, and C binds to G. Specifically, array protocols call on the hybridisation of single-stranded DNA to arrays containing hundreds of thousands of unique nucleotide probe sequences. A probe is designed to be complementary, or nearly complementary, to a portion of the DNA sequence harboring the SNP site. By measuring the intensity of the signal produced by the hybridization of these probes with the target DNA, SNP microarrays can identify genetic variations such as SNPs, indels and copy number variations.

SNP arrays offer a cost-effective, scalable technology that remains widely used despite the increasing popularity of sequencing-based approaches.

SNP microarrays used in the UK Biobank. Here, we describe the genotypic data available on the UKB, which will be the primary dataset used in this thesis. Over 480,000 UK Biobank subjects were assayed using one of two very similar SNP arrays [23]. A subset of 49,550 were genotyped using an array by Affymetrix covering 807,411 markers, called the UK BiLEVE Axiom array. Following this, 438,427 subjects were genotyped using a closely related SNP array covering 825,927 markers, called UK Biobank Axiom (also by Affymetrix). Both arrays were designed specifically for the UKB, sharing 95% of their marker content. The selection criteria for genetic variants were the following: markers relevant to specific phenotypes (around 45,000), markers in genomic regions of interest (around 47,000), coding variation (around 125,000) and markers that provide good genome-wide coverage for imputation in European populations in the common (greater than 5%) and low-frequency (1-5%) MAF ranges.

Genotype imputation in the UK Biobank. *Genotype imputation* is the process of estimating genotypes that are not directly measured in a sample of individuals: a reference panel with haplotypes at a dense set of genetic variants is used to impute unassayed genotypes in a sample with a subset of those markers. This is beneficial for several reasons: it increases the power of genetic associations studies, as well as the ability to resolve causal variants via genetic fine-mapping, and facilitates integration of GWAS results with data coming from external panels using different genotyped SNPs.

Imputation was carried out by UKB itself and is available for download. Two reference panels were used: the first was built as a merge of the UK10K [24] and 1000 Genomes Phase 3 [27] reference panels, consisting of 12,570 haplotypes with over 87 million bi-allelic markers. Another reference panel, the Haplotype Reference Consortium (HRC) [65], contained many more haplotypes (nearly 65,000) but fewer genetic variants (nearly 40 million). For this reason, SNPs were imputed using both panels, with HRC imputations being preferred for overlapping SNPs. The HRC panel was favoured due to its larger sample size and its higher accuracy for common variants in European populations.

The imputation was carried out using a modified version of the IMPUTE2 tool, using a hidden Markov model. The result of the imputation was a dataset with 92,693,895 autosomal SNPs, short indels and large structural variants in 487,442 individuals. dbSNP Reference SNP (rs) IDs were assigned to as many markers as possible using rs ID lists available from the UCSC genome annotation database for the GRCh37 assembly of the human genome.

2.2 Genome-wide association studies (GWAS)

Genome-wide association studies (GWAS) are observational studies involving thousands or tens of thousands of subjects whose genotypes are available. A phenotype, which can be a binary trait (like a disease) or a continuous trait, is measured, and subjects are recruited accordingly. Statistical tests then assess each genetic variant for association with the phenotype.

GWAS gained popularity in the early 2000s due to factors such as the completion of the human genome sequence, advancements in SNP genotyping technology, and the expansion of genetic variation databases. This method has become a common approach for identifying genetic factors influencing susceptibility to common and complex diseases. By testing individual genotyped

SNPs across a study sample, GWAS provides an unbiased and comprehensive survey of the genetic architecture of a trait without requiring prior knowledge of the location or function of causal genes.

2.2.1 Single-variant models

In a single variant model, we posit the following model for a quantitative phenotype Y:

$$Y = \beta_\ell X_\ell + \alpha + \varepsilon_\ell, \tag{2.1}$$

where $\mathbb{E}[\varepsilon_\ell] = 0$ and $\mathrm{Var}[\varepsilon_\ell] = \sigma_Y^2$, and X_ℓ is a function of the genotype G_ℓ at locus ℓ.

Different single-variant models exist to account for the effect of genetic variants on phenotypes. These depend on the inheritance pattern that is assumed: recessive, dominant or additive. In the following, we will examine autosomic inheritance, i.e. that which corresponds to non-sexual chromosomes. If we call **A** and **a** to the reference and alternative alleles at a given genomic locus, respectively:

- The **recessive model** of inheritance maps the genotypes in the following way: **AA** $\mapsto 0$, **Aa** $\mapsto 0$, **aa** $\mapsto 1$: it is necessary to carry two copies of the alternative allele to observe a difference in the phenotype.
- For the **dominant model** of inheritance, the mapping is **AA** $\mapsto 0$, **Aa** $\mapsto 1$, **aa** $\mapsto 1$: a single copy of the alternative allele is enough to affect the phenotype.
- Finally, in the **additive model** we have: **AA** $\mapsto 0$, **Aa** $\mapsto 1$, **aa** $\mapsto 2$, i.e. each additional copy of the alternative allele increases the expected value of the phenotype by a constant amount.

Throughout this work, we will always assume an additive model. Additive models are more flexible as they are able to detect part of the signal even if the true underlying process is actually recessive or dominant, albeit with less power and worse estimates for the effect size. Moreover, this is the most common practice in GWAS, where the whole genome is tested in an agnostic way without assuming prior knowledge.

Statistical inference. Here, we will derive the formulas for estimating the effect sizes β_ℓ, and a p-value for rejecting the null hypothesis that $\beta_\ell = 0$. Rejecting the null hypothesis $\beta_\ell = 0$

indicates that the genetic variant is statistically associated with variation in the phenotype. This statistical association suggests a potential biological influence. In the following, we will remove the ℓ subindex for the locus, to avoid cluttered notation.

In particular, let us find the maximum likelihood estimators (MLE) for α and β, which we call $\hat{\alpha}$ and $\hat{\beta}$ respectively. For this, we need to write the likelihood which, in turn, requires assuming a probabilistic distribution for ε: namely, we assume they are i.i.d. $\mathcal{N}(0, \sigma_Y^2)$, where σ_Y is not known. This allows us to arrive at a closed form solution for the statistical inference. Under this model, the sampling distribution of $\hat{\alpha}$ and $\hat{\beta}$ are:

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma_Y^2}{N S_{xx}} \sum_{i=1}^{N} x_i^2\right), \quad \hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma_Y^2}{S_{xx}}\right). \tag{2.2}$$

where $S_{xx} = \sum_i^N (x_i - \bar{x}) x_i$. These estimators are centered at the real values, i.e. they are unbiased estimators of α and β. These expressions, however, do not depend purely on the available data: we need to take into account that the real σ_Y^2 is unknown. It can be estimated from the data; indeed, an unbiased estimator of σ_Y^2 is

$$S^2 = \left(\frac{N}{N-2}\right)\hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 \tag{2.3}$$

where the $\hat{\varepsilon}_i^2$ are the squared residuals under the least square regression line. Furthermore, the distribution of S^2 can be characterised explicitly:

$$\frac{(N-2)S^2}{\sigma_Y^2} \sim \chi_{n-2}^2, \tag{2.4}$$

To perform statistical inference, we need to find a p-value for rejecting the null hypothesis $\mathcal{H}_0$ that $\beta = 0$. This inference is normally based on the following Student's t distribution:

$$\frac{\hat{\beta} - \beta}{\sqrt{S/S_{xx}}} \sim t_{n-2} \tag{2.5}$$

We reject $\mathcal{H}_0$ at level α if:

$$\left| \frac{\hat{\beta} - 0}{S/\sqrt{S_{xx}}} \right| > t_{N-2,\alpha/2}. \tag{2.6}$$

GWAS summary statistics normally provide, in addition to the p-value of the association, the estimates $\hat{\beta}$ and its standard error, given by $S/\sqrt{S_{xx}}$.

2.2.2 Family-wise error rate (FWER)

The FWER is defined as the probability of making one or more false discoveries (rejecting null hypotheses incorrectly) among a set of hypotheses or tests. It represents the overall probability of committing a type I error in the entire family of tests [2]. The "family" refers to the collection of tests being conducted simultaneously, such as comparing multiple groups or examining multiple variables. In the case of GWAS, each test of the family corresponds to a different genetic variant.

To control the FWER, researchers often apply a correction method, such as the Bonferroni correction, to adjust the p-values or significance thresholds of individual tests. These corrections help maintain the desired overall level of significance while accounting for the increased chance of making false discoveries in multiple testing scenarios. In other words, it reduces the probability of making a type I error, at the expense of committing more type II errors. In GWAS, this leads to more trustworthy positive associations, which is desirable since these associations would open the door to further resource-consuming research.

In the case of GWAS of a single quantitative phenotype, there is one test per genetic variant tested. However, given that linkage disequilibrium implies that nearby genetic variants tend to be highly correlated, the number of effective tests is lower than the number of genetic variants tested. Within the GWAS community, 10^6 is normally accepted as the effective number of tests performed on GWAS, which leads to a genome-wide p-value threshold of 5×10^{-8} to control the FWER at 5%.

2.2.3 Adjustment for covariates and inverse rank normalisation (IRN)

Before GWAS, the phenotypes were adjusted for a set of covariates. To make this adjustment in this work, multivariate linear regression on these covariates was performed, and then the

[2] A **type I error**, also known as a "false positive" or "alpha error", occurs when a statistical test incorrectly rejects a true null hypothesis; in this case, suggesting that there is a real genetic association when, in fact, there is not. On the other hand, a **type II error**, or "false negative" or "beta error", refers to those real associations for which the null hypothesis is not rejected.

residues of this regression were inverse-rank-normalised. In other words, if C is the data matrix for covariates, then $\mathbf{y}_{\text{adj}} = \text{IRN}(M\mathbf{y})$ where $M = \mathbf{1} - C(C^tC)^{-1}C$ and $\text{IRN}(\cdot)$ is the function performing inverse-rank normalisation. The latter step is justified if we remember that our statistical inference framework requires for p-values to be well calibrated[3]. For this to be the case, we require the data to follow a normal distribution under the null hypothesis ($\beta = 0$). These inverse-normalised residues, $\mathbf{y}_{\text{adj}}$, are the phenotypic scores to be tested in the GWAS.

Rationale behind covariate adjustment

We briefly discuss the rationale behind the procedure of adjustment by covariates. We are normally interested in adjusting specifically for those variables which *precede causally* to our phenotype of interest. By doing so, we would be focusing on the fraction of phenotypic variation that remains unexplained, thereby increasing the normalized genetic effect sizes and the chances of a true-positive association in the GWAS. Additionally, we would be preventing the inflation of p-values caused by genetic loci influencing the trait of interest *through* the covariate (if existing), on which we are *not* interested.

If it were known that variation in a given measure, say systolic blood pressure (SBP) to fix ideas, always *follows* variation in some structural parameter, say LVEDV, then adjusting LVEDV for SBP would not be desirable when trying to find genetic associations. Indeed, the effect of this procedure would be simply to reduce extent of the variation in the phenotype of interest and hence the magnitude of the estimated SNP effect size, thereby limiting the number of genetic discoveries. On the other hand, if changes in SBP were *followed* by changes in LVEDV (the causal arrow is reversed with respect to the previous case), then it *would* be desirable to adjust LVEDV for SBP, since SBP acts as a confounder .

Reality is usually more complex than the hypothetical examples presented, and the causal structure of the problem will be more intricate and not fully known *a priori*. For instance, in the particular case of blood pressure and chamber volumes, the relation is rather bidirectional: both variables might influence each other in a feedback loop. However, there are some other measures for which it is obvious that that the variable precedes the trait of interest, i.e. is a cause and not a consequence; examples of such variables are age or sex. For the ambiguous

[3] A p-value being well calibrated means that they will distribute uniformly between 0 and 1 when the null hypothesis is true. If they are miscalibrated, we will tend to systematically underestimate or overestimate the probability of type I error, thus making misleading conclusions.

cases, like the aforementioned SBP, we decide to include them as covariates, which would be a conservative decision; albeit at the cost of, presumably, lowering our ability to detect implicated loci.

2.3 Downstream analysis

In order to interpret the GWAS results, a series of downstream analysis are usually performed. Critically, note that GWAS provides SNP-level associations, as opposed to gene-level associations. It is not always clear which is the gene that mediates these associations: while in some cases, the lead SNP in a locus lies within the body of a gene, this is not always the case. We describe two common approaches used in this work: proximity analysis and transcriptome-wide association studies (TWAS).

2.3.1 Proximity analysis

The first step, after identifying the lead variant in each locus, is to determine which genes lie in its vicinity. For this, databases such as Ensembl can be utilised [29]. In this database, the transcription start sites (TSS) and transcription end sites (TES) of all the genes of the genome can be queried.

Coding variation corresponds to variants that are localised between TSS and TES, however a variant can act on a (typically) nearby gene by regulating its expression. These variants are called *expression quantitative trait loci* (eQTL), where the prefix *cis* is sometimes added to emphasise that the variants are included in the same chromosome as the target genes. eQTL can live up to 2 Mb away from their target gene, however they are typically not further than 100 kb away.

2.3.2 Transcriptome-wide association studies

GWAS have been very successful in identifying numerous associations between genetic variants and a wealth of diverse phenotypes. However, the biological mechanisms behind most of these associations is not well understood. Moreover, it is often the case that genetic associations from GWAS lie on intergenic regions, that is, stretches of DNA located between genes. We will see examples of this in our own results in later chapters. Interpreting these associations biologically

is particularly challenging, since it is not clear which is the mediating gene [4]. Unlike the case of a variant lying within a coding region, where the mechanism is likely to be a modification in the structure of the molecule encoded for by the gene, variants lying outside of genes typically regulate the expression of nearby genes, i.e. differential gene expression is observed in the presence of one or the other allele. For instance, SNPs in regulatory regions like promoters and enhancers can create, disrupt or alter the affinity of transcription factor binding sites. This is the main mechanism of action of these intergenic SNPs.

A more direct way to assess whether gene expression is the true mediating factor would be to directly measure gene expression, and test statistically for association between this and the phenotype of interest. In the case of protein-coding genes, this would entail measuring messenger RNA (mRNA) or protein levels. However, this approach has a number of problems. First, gene expression changes greatly across cell types, which means that we need to extract samples from the relevant tissues, which may not be easily accessible. Furthermore, and related to the previous, we need to know which is the relevant tissue beforehand; this knowledge is not always available. On the other hand, unlike genetic variation which can only be the cause of the phenotype since it is unalterable, there is no guarantee of a causal role for gene expression: it could be the case that a systemic state linked to the tested phenotype leads to changes in gene expression (reverse-causation). Finally, techniques for measuring mRNA or protein levels are more costly.

The above reasons motivate a different framework, known as *transcriptome-wide association studies* (TWAS). TWAS aim to integrate GWAS results with studies on gene expression carried out on an independent cohort [41]. They require a cohort on which both the genotype and gene expression levels (called the *transcriptome*, when measuring mRNA levels) are measured. From this panel, the effect on gene expression of variations in the genotype can be estimated. This allows the estimation of the *genetic component of gene expression* T_g for each gene g on an independent cohort (in this case, the GWAS cohort), by means of the SNP-mRNA associations found. By correlating T_g with the phenotype that is being tested, we thus obtain a gene-level association as opposed to a SNP-level association, the former being more interpretable. Moreover, the described TWAS framework does not face the same challenges related to reverse

[4] Recall that the genetic variation must always produce a change in some biologically active element, that is, a *gene*, for the change to have an effect on the phenotype. For this reason, we can think that there is always a mediating gene for any given (true) genetic association.

causation, since it leverages genetic data that is inherently causal rather than expression data that may be influenced by the phenotype itself, nor does it require tissue-specific gene expression measurements on the GWAS cohort (it can be done once on an independent cohort and used across many different GWAS). To date, the largest such transcriptomic database with a broad range of sampled tissues is the Genotype Tissue Expression (GTEx) project, which has utilised post-mortem samples of over 800 subjects (in its version 8) and over 50 tissues [28], many of which would have been impossible to access on living subjects. In this work, we will utilise the S-PrediXcan tool to perform TWAS [14], which requires only GWAS summary statistics to obtain gene-level associations, as well as SNP-based gene expression prediction models which were trained on GTEx data and made available by the authors of the tool.

TWAS can be thought of as a particular case of a Mendelian randomization (MR) study, where the *exposure* (in MR terminology) is gene expression in a specific tissue or cell type, and the *outcome* is the phenotype of interest. As with other MR studies, results need to be interpreted with care since the assumptions of MR may not hold [106]. Of particular relevance, the "independence" assumption, according to which there are no common causes of the genetic variants and the outcome is usually broken because of linkage disequilibrium. Therefore, TWAS provides a way of identifying candidate genes and candidate mediating mechanisms (i.e. gene expression), without guaranteeing causality. In this thesis, we will employ TWAS as a way to provide evidence towards possible candidate genes responsible for a genetic association.

2.4 Cardiac physiology

Since the heart is the organ that will occupy us in this work, its structure and function are reviewed in this section. The purpose is to provide a succinct overview that the reader can use as a reference throughout this dissertation. For instance, among our genetic findings in later chapters we will often encounter the themes of calcium handling in the myocyte, of ion channels, of sarcomere banding or of early cardiac development, to name only a few concepts that may result unfamiliar to a reader with no biological training. This section will hopefully help maintain the thesis self-contained, however I encourage the interested reader to consult the cited references.

2.4.1 Overview of cardiac anatomy and function

The heart is a fist-sized muscle in charge of supplying the blood containing oxygen (O_2) and other nutrients to the cells of an organism's body. It consists of two halves, left and right (Figure 2.2 can be consulted while reading the text). Each half, in turn, is composed of two compartments, called atrium and ventricle, that are located in the upper and lower parts, respectively.

Blood is pumped in a cyclic fashion. The right atrium (RA) receives deoxygenated blood from the body, which flows through the tricuspid valve into the right ventricle (RV). Upon contraction, the RV pumps blood into the pulmonary arteries, which transport it to the lung capillaries, closely associated with the alveoli. Gas exchange occurs there: carbon dioxide (CO_2), a metabolic byproduct, is released, and oxygen (O_2) is absorbed. The oxygen-rich blood returns to the heart via the left atrium (LA), flows through the mitral valve into the left ventricle (LV), and is then pumped to the systemic circulation.

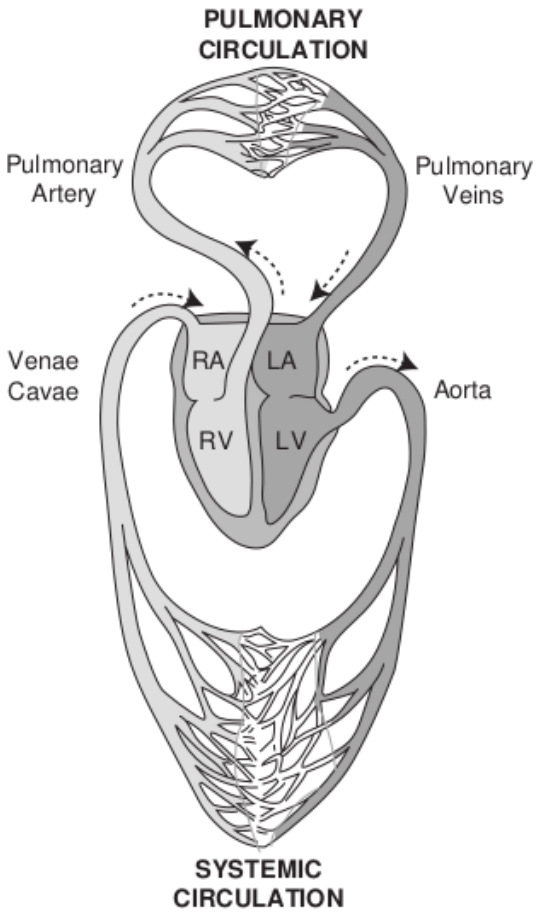

Figure 2.2: Overview of blood circulation (from [47]).

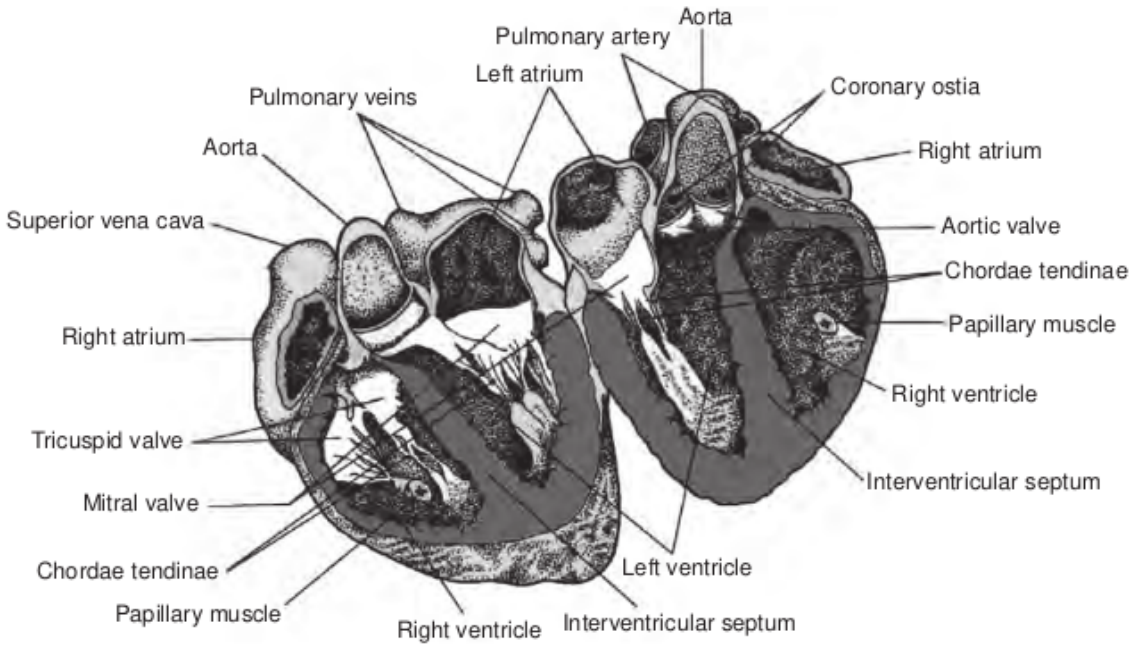

Figure 2.3: A detail of the different structures that compose the heart (from [47]).

2.4.2 The cardiac cycle

In the following, we examine the cardiac cycle in more detail. The Wiggers diagram (see Figure 2.4) depicts the different events that take place in the heart throught the cardiac cycle. It superimposes the pressure curves for ventricles, atria and aorta, and the ventricular volume curves, along with the electrocardiogram (ECG).

The contraction and subsequent relaxation of the two ventricles occur in phase. The first is called *systole*, whereas the expansion part is known as *diastole*.

The electrical events in the heart are observed in the ECG, which provides the global electrical state of the heart. The three main events that can be observed are:

- The P wave: where the atria are stimulated to pump blood into the ventricles.
- The QRS complex: where the ventricles are triggered to pump blood into the respective outlet vessels.
- The T wave: which marks the recovery period of the ventricles.

2.4.3 Cardiac histology

There are three types of muscle tissue in the human body: skeletal, cardiac and smooth. Both skeletal muscle and cardiac muscle are striated, but the latter differs from the former in that it

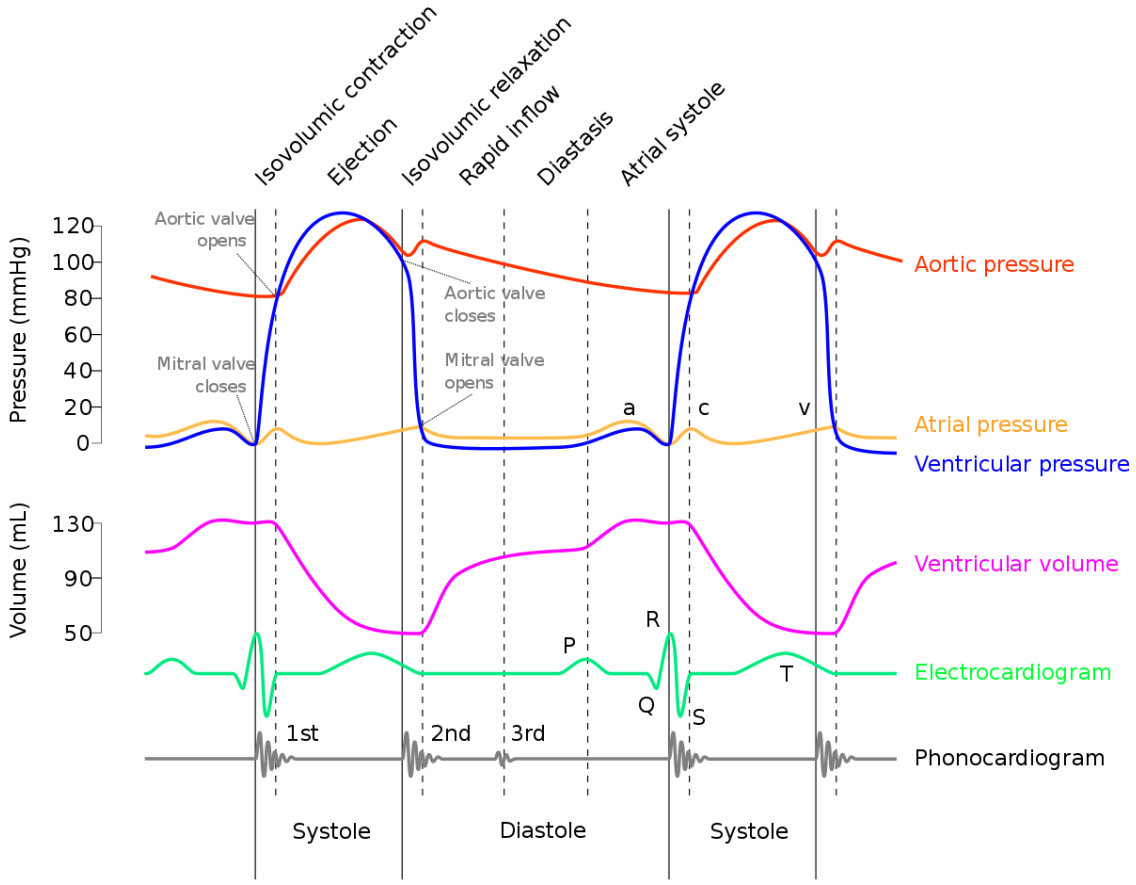

Figure 2.4: Wiggers diagram showing the concomitant events that take place in the heart during one complete heartbeat (from `https://en.wikipedia.org/wiki/Cardiac_cycle`).

is not under voluntary control and exhibits rhythmic contractions.

The individual cardiac muscle cell, or **cardiomyocyte**, is a tubular structure composed of chains of myofibrils, which are rod-like units within the cell (see Figure 2.5). The myofibrils consist of repeating sections of sarcomeres, which are the fundamental contractile units of the muscle cells. Sarcomeres are composed of long proteins that organize into thick and thin filaments, called myofilaments. Thin myofilaments contain the protein actin, and thick myofilaments contain the protein myosin.

2.4.4 Quantitative measures of cardiac structure and function

Here, I provide an overview of some parameters that are useful for assessing cardiac function [38]:

- **End-diastolic volume (EDV) (ml):** the volume of blood in the LV or RV before contraction. This is the highest ventricular volume of blood throughout the cardiac cycle.

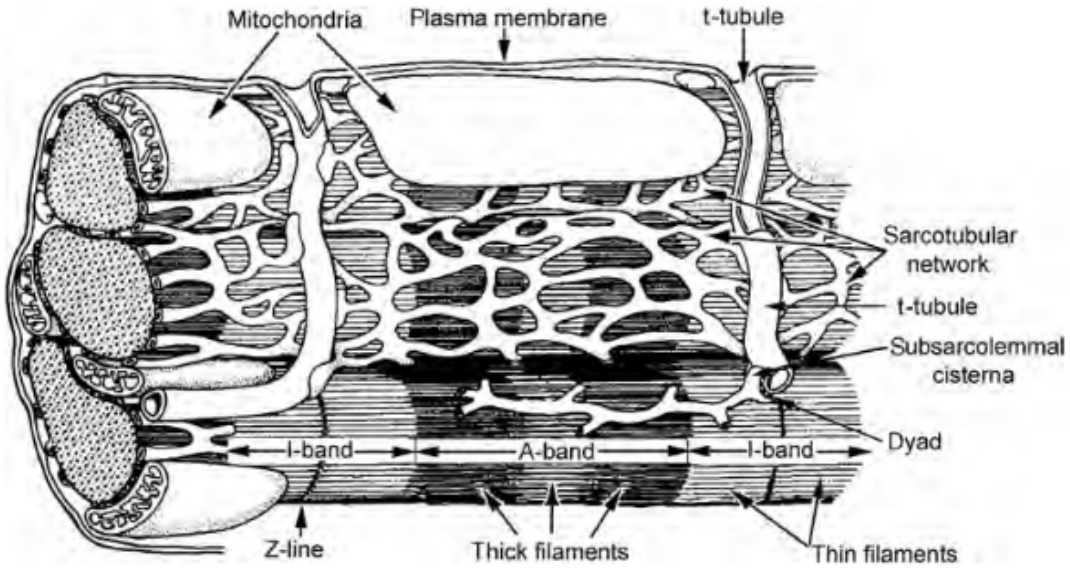

Figure 2.5: Ultrastructure of a cardiomyocyte (from [47]).

- **End-systolic volume (ESV) (ml):** the volume of blood in the LV or RV at the end of contraction. This is the lowest ventricular blood volume throughout the cardiac cycle.
- **Stroke volume (SV) (ml):** the volume of blood pumped from the ventricle per beat obtained by subtracting the ESV from the EDV for a given ventricle. This term can be applied to either of the two ventricles[5].
- **Ejection fraction (EF) (%):** fraction of blood ejected from a ventricle during each heartbeat. This measure shows the pumping efficiency of the heart and is calculated by dividing the SV by the EDV. Note that the left ventricular EF (LVEF) is a measure of the efficiency of pumping blood into the body's systemic circulation, whereas the right ventricular EF (RVEF) is a measure of the efficiency of pumping blood into pulmonary circulation (i.e. the lungs).
- **Myocardial mass (M) (g):** this quantity is calculated by multiplying the volume by the density of the muscle tissue (1.05 g/cm^3). In this work it will always refer to the mass of the left-ventricular wall (LVM).

[5]Since the chambers are part of a series circuit, in steady state the stroke volume is expected to be approximately the same regardless of the chamber that is used to compute it.

2.5 Genetics of cardiac IDPs

As reviewed by De Marvao *et al.* [30], numerous studies have investigated the genetic basis of cardiac image-derived phenotypes. Up to 2019, however, most of them had utilized echocardiography as imaging technique. From this year onwards, several articles published novel findings using CMR data from the UK Biobank. In this section, we summarise the findings from these studies.

Together, these studies have established a robust set of loci associated with LV phenotypes, forming a strong foundation for further genetic discovery based on advanced phenotyping approaches, such as those pursued in this thesis.

2.5.1 Studies on the left ventricle

The first GWAS of cardiac structure performed on the general population (as opposed to case-control studies) were carried out on echocardiogram-derived phenotypes. [99] investigated left-ventricular phenotypes extracted from echocardiograms. They performed a meta-analysis using a set of five cohorts.

Biffi *et al.* [16] investigated the association of LV myocardial wall thickness against a set of 6 exonic SNPs known to have effects on cardiac phenotypes (thus it does not categorize as GWAS as it is not genome-wide). They performed an association test for each SNP in a vertex-by-vertex fashion (termed *mass univariate regression*) across an LV 3D mesh in end-diastole. Using these associations, the authors generate heat maps where the color represents the strength of the association at that particular mesh location. This methodology aims to provide further insights into the effects of known loci, but not to perform genetic discovery.

The article by Aung *et al.* on LV CMR-derived phenotypes [9], published in September 2019 (coincidentally with the start of this PhD), is the first one on cardiac imaging genetics using UK Biobank data, and also the first performing GWAS on CMR-derived phenotypes. The authors investigated LV structure and function parameters described in **Section 2.4** using a set of ∼16,000 European subjects and performing replication analysis on an independent cohort of ∼5000 individuals (*Multi-Ethnic Study of Atherosclerosis*, MESA). After performing GWAS, they utilized additional in-silico evidence in order to establish which was the gene responsible for the SNP association. Their findings are shown in figure 2.6, where the candidate genes are

classified according to the extent to which they are supported by evidence.

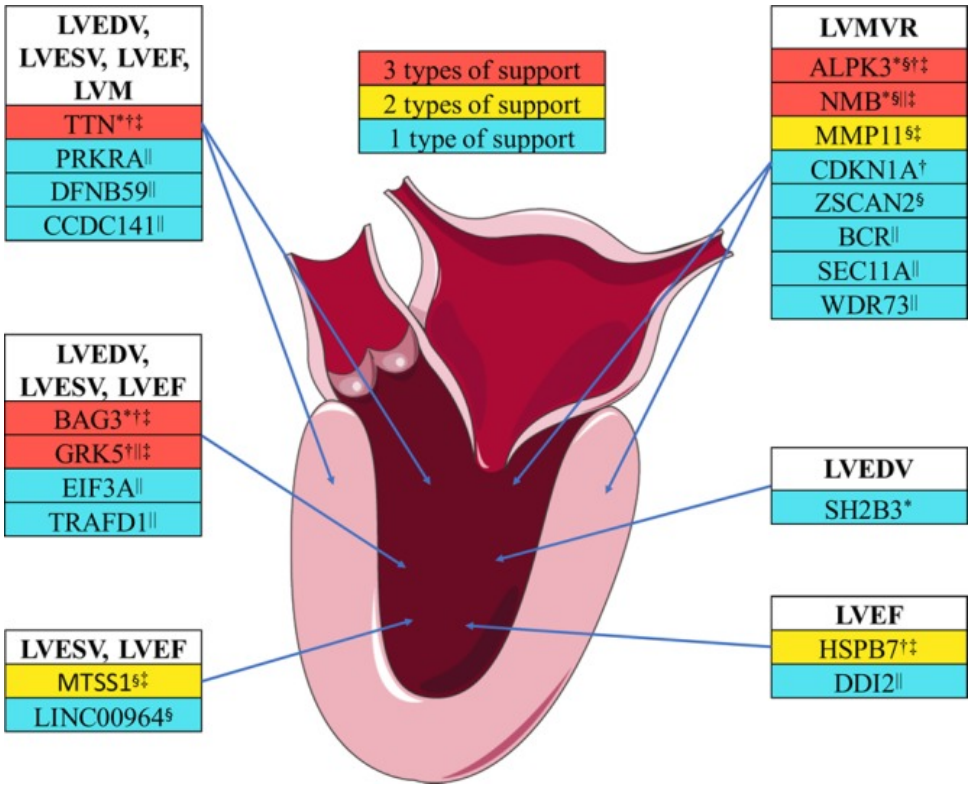

Figure 2.6: Candidate genes responsible for the GWAS signals found in [9] (figure from [9]).

The paper [78] by Pirruccello *et al.* was published the following year (2020) and overlaps with the previous in that it also focused on LV using UKB CMR data. The sample size is 36,014 and is not restricted to subjects with British ancestry. The authors investigated LVEDV, LVESV, LVSV and LVEF and did not perform automatic segmentation themselves but instead utilised values produced by the CMR scanner via its inline quantification software. These values are available from UKB data fields (e.g., field IDs 22421–22423 and 22420 for LVEDV, LVESV, SV, and LVEF, respectively). Furthermore, the body-surface-area-indexed versions of these phenotypes (except LVEF), were studied. These represent the values of the different measurements divided by the body surface area (BSA) of the individual, estimated from their height and weight. They are denoted by appending an "i" in the end of the phenotype name (e.g. LVEDVi). This article reports a great number of genetic associations: 57 independent genome-wide significant loci, 45 of which were novel. Discriminated by phenotype, 22 loci were associated with LVEDV, 14 with LVEDVi, 32 with LVESV, 28 with LVESVi, 22 with LVEF, 12 with SV, and eight with SVi.

The study [105] by Vukadinovic *et al.* investigated left-ventricular sphericity index. Some of the conclusions of this work, especially regarding the effect of variants in the *PLN* locus,

are consistent with our own findings, as reported in our MICCAI 2021 conference paper and further developed in our Nature Machine Intelligence publication, which are discussed in detail in Chapters 6 and 7.

2.6 Deep learning and genetic discovery in imaging genetics

The use of deep learning in GWAS has been hitherto limited. Here, we briefly discuss the opportunities that this new technology has opened, some of which we explore in this thesis.

2.6.1 Neural networks as "segmenters"

Most of the previously cited works on cardiac imaging genetics have employed state-of-the-art automatic segmentation techniques, in particular, deep-learning-based ones. The output of these approaches typically is a *segmentation mask* for the structures of interest, which can then be processed to derive quantitative phenotypes using predefined computational routines. In the context of cardiac imaging genetics, the quantities usually studied are the previously discussed volumes and derived quantities, or other relevant dimensions such as wall thickness.

2.6.2 Neural networks as "phenotypers"

In this work, we center the discussion around the use of artificial neural networks as phenotypers—that is, models whose input is a complex biological object and whose output is a vector of quantitative phenotypes that can be used for various downstream tasks, particularly genetic association studies. The input objects may include, for instance: images, meshes, curves, tabular data, or combinations thereof.

With this approach, neural networks can be trained in a *supervised manner* to predict a predefined quantitative phenotype of interest, and then applied at scale to estimate this parameter across a larger population; for instance, a network trained to estimate LVEDV from CMR scans using ground truth labels for a subset of the cohort. Alternatively, they can be trained in an *unsupervised way*, where the phenotypes are purely data-driven—rather than handcrafted—and correspond to hidden, or *latent*, neural representations of the input data.

This thesis focuses on the latter, unsupervised approaches, leveraging the unprecedented sample size of UK Biobank to automatically learn quantitative phenotypes. We hypothesize that this strategy allows for better coverage of the phenotypic variability and improved genetic discovery.

Concurrently with this work, similar analytic approaches have been proposed and applied to other imaging modalities: [51] and [110] applied autoencoder networks to extract unsupervised phenotypes from fundus photographs, whereas [90] conducted a similar study on retinal thickness maps derived from optical coherence tomography (OCT), a technique that allows for the visualization of the internal structure of the tissue with high resolution. Our work focuses on a different modality, cardiovascular magnetic resonance imaging, for which, to our knowledge, no other similar studies have been published to date.

Chapter 3

Cardiac imaging and segmentation

This chapter describes the cardiac imaging data used in this work and the procedures conducted to extract 3D cardiac meshes from it. It is organised as follows:

- **Section 3.1** introduces the fundamentals of cardiovascular magnetic resonance (CMR), including basic MRI physics and the specific challenges of imaging a moving organ like the heart.
- **Section 3.2** then describes the CMR acquisition protocol of the UK Biobank and the nature of the datasets employed in this study.
- In **Section 3.3**, subsequently, we discuss the segmentation approaches used to delineate cardiac structures, with a particular focus on landmark-based mesh generation methods.
- In **Section 3.4** we present the full segmentation pipeline developed and executed to generate the registered cardiac meshes from raw CMR data.
- Finally, **Section 3.5** includes a note on the code availability.

By the end of this chapter, the reader should have a detailed understanding of the characteristics of the imaging data and the steps required to produce standardised 3D cardiac meshes, which constitute the primary input for the subsequent work in this dissertation.

3.1 Basics of Cardiovascular Magnetic Resonance (CMR)

Assessment of cardiac function encompasses various baseline tests, including blood tests or electrocardiograms (ECGs), as well as a range of imaging techniques available in clinical practice. These techniques, such as echocardiography (echo), chest X-rays, computed tomography (CT), nuclear imaging, and magnetic resonance imaging (MRI), offer non-invasive qualitative and quantitative evaluation of cardiac function and structure, facilitating diagnosis, disease monitoring, treatment planning, and prognosis.

Cardiovascular magnetic resonance (CMR) imaging stands out among cardiac imaging modalities due to its ability to provide precise morphological detail and high-quality soft-tissue contrast without the use of contrast agents or ionising radiation. It has become the non-invasive gold standard for assessing cardiac chambers across a broad range of cardiovascular diseases (CVDs), offering accurate and reproducible tomographic, static, or cine images with high spatial and temporal resolution in any desired plane.

3.1.1 Physics of MRI

The technique of MRI is based on the fact that nuclear spin interacts with magnetic fields. In the hydrogen atom, the nucleus is composed of a single proton in the majority isotope, ^{1}H. The energy levels of the nucleus, degenerate in absence of magnetic field, become split into two levels where the energy difference ΔE is given by:

$$\Delta E = \gamma B \tag{3.1}$$

where $\gamma = 42.58\,\mathrm{MHz/T}$ is the gyromagnetic constant and B is the intensity of the magnetic field. The population of the two states, $E_\downarrow$ and $E_\uparrow$ is given by the Boltzmann statistics. If $f_\downarrow$ and $f_\uparrow$ are the occupation fractions for each state, they are given by:

$$(f_\downarrow, f_\uparrow) = \left(\frac{\alpha}{1+\alpha}, \frac{1}{1+\alpha}\right) \tag{3.2}$$

where $\alpha = \exp(\frac{-\gamma B}{kT})$, where $k = 1.381 \times 10^{-23}\mathrm{J/K}$ is the Boltzmann constant and T is the temperature (normally, room temperature). With $B = 1.5\mathrm{T}$ (the field strength used in the UK Biobank CMR protocol), $\alpha \approx 0.990$.

We will first analyse the case where a uniform field $\mathbf{B}_0 = B_0\hat{\mathbf{z}}$ is applied along the z direction ($\hat{\mathbf{z}}$ represents a unit vector along the z direction). As discussed above, the nuclear spins are aligned along this direction, generating a magnetisation $\mathbf{M} = M\hat{z}$. As a result of the interaction between $\mathbf{B}_0$ and the positively charged ^{1}H nucleus, the individual protons begin to rotate (or *precess*) around the z axis, at a frequency called the *Larmor frequency*, $\omega = \gamma B/2\pi$. This frequency is fundamental in MRI imaging: to generate an image, magnetic field will be varied for different spatial locations, with each tissue location consequently precessing at different frequencies.

Radiofrequency pulses

MRI relies on an energy transfer process, by which ^{1}H are first excited with a pulse of energy at the Larmor frequency. This energy is absorbed and then re-emitted, where this re-emission is the detected signal. In MRI, this energy is provided by a radiofrequency (RF) magnetic pulse.

For simplicity of exposition, we will still assume a spatially uniform magnetic field $\mathbf{B}_0$ and hence the associated Larmor frequency is $\gamma B_0/2\pi$ (the Larmor frequency associated to B_0). The pulse is of amplitude B_1, orthogonal to (and significantly smaller than) $\mathbf{B}_0$. B_1 induces a rotation of the magnetisation vector $\mathbf{M}$ towards the $x-y$ plane. Furthermore, the intensity, duration and shape of this RF pulse can be adjusted so that the direction of the spins becomes orthogonal to z. Pulses are normally termed based on the rotation they produce, for example a $\pi/2$ pulse produces a rotation of 90 degrees.

Receiver coils

By Faraday's law of induction, if a loop of wire is placed perpendicular to the transverse plane, the oscillating net magnetic flux from the excited spin system will induce a voltage. This induced voltage, called the free induction decay (FID), in turn generates an electric current which is then amplified, digitized, and filtered to extract frequency and phase information. These wire loops are called *receiver coils*.

T_1 and T_2 decoherence processes

Right after the application of a $\pi/2$ RF pulse, all the magnetisation is on the transverse plane. Once the pulse is turned off, the spins will return to the equilibrium state, in which they are aligned to $\mathbf{B}_0$, in an exponential growth process, given by $M_z(t) = M_0(1 - e^{t/T_1})$, where t is the time after the RF pulse. T_1 is thus the characteristic time for this relaxation process, which

is termed spin-lattice relaxation because the energy transfer is from the excited nuclear spins to the lattice. T_1 depends on the particular tissue and on the strength of the applied magnetic field, B_0. At lower field strengths, the spin-lattice energy transfer happens more readily because the probability of molecular motion matching the resonant frequency is is higher, and thus T_1 is shorter. For example, with $B_0 = 1.5$T (the field strength used in the UK Biobank CMR protocol), T_1 for the myocardium is around 1100 ms, whereas for blood it is of around 1600 ms.

Another process, denominated *transverse* or *spin-spin relaxation*, characterises the decay of transverse magnetisation, M_{xy}. This is an exponential decay process given by $M_{xy}(t) = M_0 e^{-t/T_2}$, where T_2 is the characteristic time for this phenomenon. The dependence of T_2 on the field strength is not as readily described as with T_1.

Magnetic field gradients encode position

It is possible to encode position by applying a magnetic field gradient in a certain direction. Protons at different spatial locations will thus resonate at different frequencies. In the following, we details the techniques for slice selection, frequency encoding and phase encoding, which constitute the building blocks for 2-dimensional MRI imaging. The discussion is kept simple and important technical details are omitted in the interest of brevity.

Slice selection. Firstly, the MRI signal is localised in one of the directions, which we call z. To do this, a gradient $\nabla B_0 \parallel \hat{\mathbf{z}}$ is applied along this direction causing the nuclei to precess at different frequencies along the z axis. A RF pulse is applied to excite spins within a certain range of frequencies, given by its central frequency (which specifies the location) and its bandwidth (which specifies the slice thickness). This is usually achieved by applying a sinc-shape RF pulse, $B_1(t) = B_1 \text{sinc}(\pi t/T) = B_1 \frac{\sin(\pi t/T)}{\pi t/T}$, where T is chosen to excite a certain range of frequencies (i.e. a certain range of z).

Frequency encoding. By generating another gradient, say in the x direction during the readout, the ^{1}H nuclei will precess at different frequencies based on the x coordinate. The signal being detected becomes thus a superposition of these frequencies. By applying the Fourier transform to this signal, the number of precessing nuclei at each location can be determined.

Phase encoding. For the remaining y direction, spatial encoding is carried out by looking at phase information in the signal, taking advantage of the periodicity of nuclei precession: after

slice selection, and prior to application of the gradient for frequency encoding, another gradient G_y is applied along the y direction. When G_y is turned on, protons start precessing at different frequencies depending on their y coordinate. Once G_y is turned off, spins go back to their original precession frequency, but the frequency differences during the period in which $G_y \neq 0$ are reflected as a gradient in phase. The phase shift depends on the magnitude of G_z and its duration.

3.1.2 Particularities of CMR imaging

CMR presents particular challenges since it aims to scan an organ that moves faster than the MRI acquisition time. For this reason, the measurement is performed across several heartbeats and by synchronising MRI acquisition with certain features of the ECG signal. About thirty phases can be obtained during one cardiac cycle with the currently available equipment, yielding a temporal resolution of about 30 ms. For example, in the UK Biobank, 33 phases were acquired and then interpolated to the final 50 phases.

The standard imaging plane, called the *short-axis* plane, is perpendicular to the long axis (joining the apex with base of the left ventricle). CMR imaging covers the whole organ with about 8–15 short-axis slices, with distance between adjacent slices ranging from 10 to 20 mm.

3.2 CMR data in UK Biobank

3.2.1 The acquisition protocol

CMR imaging in the UK Biobank was performed across four imaging centres. They employed a clinical wide bore 1.5 Tesla scanner (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany). The scanners are endowed with 48 receptor coils, a 45 mT/m and 200 T/m/s gradient system, an 18-channels anterior body surface coil used in combination with a 12 elements of an integrated 32 element spine coil and electrocardiogram (ECG) gating for cardiac synchronization. The Shortened Modified Look-Locker Inversion recovery technique (ShMOLLI, WIP780B) was implemented on the scanner in order to perform non-contrast myocardial T_1 mapping.

3.2.2 Available data

Each subject's data comprises three long-axis (LAX) cines, which we call horizontal long axis (HLA), vertical long axis (VLA), and left ventricular outflow tract (LVOT) cines, both sagittal and coronal and a complete short-axis (SAX) stack of balanced steady state free precession (bSSFP) cines. Depending on heart size, 8 to 15 SA slices were acquired at the specified out-of-plane resolution.

We downloaded cardiac MRI image sequences, where each sequence is given as a $W \times D \times 50$ arrays in DICOM format (with W and D varying across subjects with values of around 200), with in-plane resolution $1.8 \times 1.8\,\mathrm{mm}^2$ and out-of-plane resolution of 8 mm for SAX and 6 mm for LAX slices. The temporal resolution was of 32 ms, and the sequences were interpolated to 50 time frames.

The dataset was downloaded using the `ukbfetch` tool from the UK Biobank, with the field code `20208` for long-axis views, and `20209` for short-axis views.

3.3 From images to meshes: segmentation approaches

The word *segmentation*, in the context of medical image analysis, refers to the process of delineating the structures of interest in an image. In the case of cardiac imaging, these structures are myocardial surfaces of the different cardiac chambers —LV, RV, LA and RA— as well as the great vessels —aortic root, the inferior vena cava, the superior vena cava, the pulmonary artery and the pulmonary veins.

While it is possible to carry out this process manually (with the help of dedicated software), the scale of the dataset that needs to be analysed for this works makes it prohibitive. In the present section, we provide an overview of some automatic methods, as well as a detailed explanation of the deep-learning-based method utilised for this work, MCSI-Net.

3.3.1 Overview of automatic segmentation methods

We briefly review automatic segmentation methods, by classifying them in two ways: with respect to the form of the segmentation output, and with respect to the nature of the methodology.

Semantic segmentation vs. landmark-based methods. *Semantic segmentation* refers to the process of assigning labels to the pixels of voxels, corresponding to the structure they belong to, usually called *segmentation masks* in this context. For example, [11] has proposed to utilise a fully-convolutional network to produce segmentation masks.

On the other hand, *landmark-based segmentation methods* rely on specific anatomical reference points within an image (the *landmarks*), to delineate the structures of interest. In the case of the heart, there are a number of such representative locations, e.g. the apex. The rest of the surface is filled with so-called *pseudo-landmarks*, which are not linked to specific anatomical locations. Since the topology of normal cardiac shape does not change across subjects, landmark-based methods are suitable and, we argue, desirable, since they produce a standardised object regardless of the characteristics of the acquisition or of the size of the subject's heart.

Traditional versus deep-learning-based methods. The emergence of deep learning in recent years has revolutionised the field of image segmentation by providing more accurate and efficient solutions compared to traditional methods relying on handcrafted feature definition, hence displacing them to a large extent in the present day.

3.3.2 Automatic segmentation approaches used in this work

In this work, cine-cardiovascular magnetic resonance (cine-CMR) images from UK Biobank were processed through a segmentation pipeline which produced registered meshes, i.e. meshes with the same number of vertices and the same connectivity between them.

Two segmentation methods were utilised, both of them developed by close collaborators.

- **SpASM:** the first one is a classic method based on sparse active shape models (SpASM). In this case, the meshes were generated from the images by a colleague of mine, Dr. Rahman Attar, then doctoral candidate, and the method is described in detail in [6]. The method was applied by the author on a set of 31,000 subjects and constitutes the first dataset I had access to during my PhD.
- **MCSI-Net:** the second method was deep-learning-based, and its concept was developed also by Dr. Rahman Attar, and continued by colleagues Avan Suinesiaputra, Yan Xia, Nishant Ravikumar and Xiang Chen after Rahman's leaving CISTIB. The initial concept

was developed and preliminary results were presented in a conference paper [5], while the final implementation was applied to UK Biobank data and published in a journal paper [108].

In addition to these two methods, I collaborated in the development of a third deep-learning-based method [40]. However, this method was not applied at scale for this dissertation due to time constraints and hence it is not discussed in this chapter. It is brought back into discussion in the last chapter, Conclusions and future directions, in regards to the possibility of producing volumetric cardiac meshes.

In the following, I focus particularly on the MCSI-Net algorithm, as this is the one on which the integrity of the presented results are based, with the exception of those in section 6.6, which are chronologically previous to the rest. While I was not involved in the development of this automatic segmentation method, it helps maintain this dissertation more self-contained. Furthermore, the application of the whole pipeline from the raw images to the 3D surface meshes was conducted at scale by myself with the assistance of the authors.

3.3.3 The MCSI-Net Algorithm

The architecture of MCSI-Net is depicted in Figure 3.3. The network contains two branches, which are called MMF-Net and Loc-Net. MMF-Net outputs a PCA-based shape representation of the heart, whereas Loc-Net produces a set of transformation parameters to bring the shape back to the image space, with the original scale. Details on how to produce the training PCA labels are provided later in this subsection.

The two tasks are learned independently by these two sub-networks. The network is trained using the following loss functions:

$$L_{\text{MMF-Net}} = \sum_{j=1}^{K} f(\hat{b}_j, b_j), \tag{3.3}$$

and

$$L_{\text{Loc-Net}} = \sum_{l=1}^{8} f(\hat{t}_l, t_l), \tag{3.4}$$

where K represents the number of shape parameters (PCs), θ denotes the trainable parameters

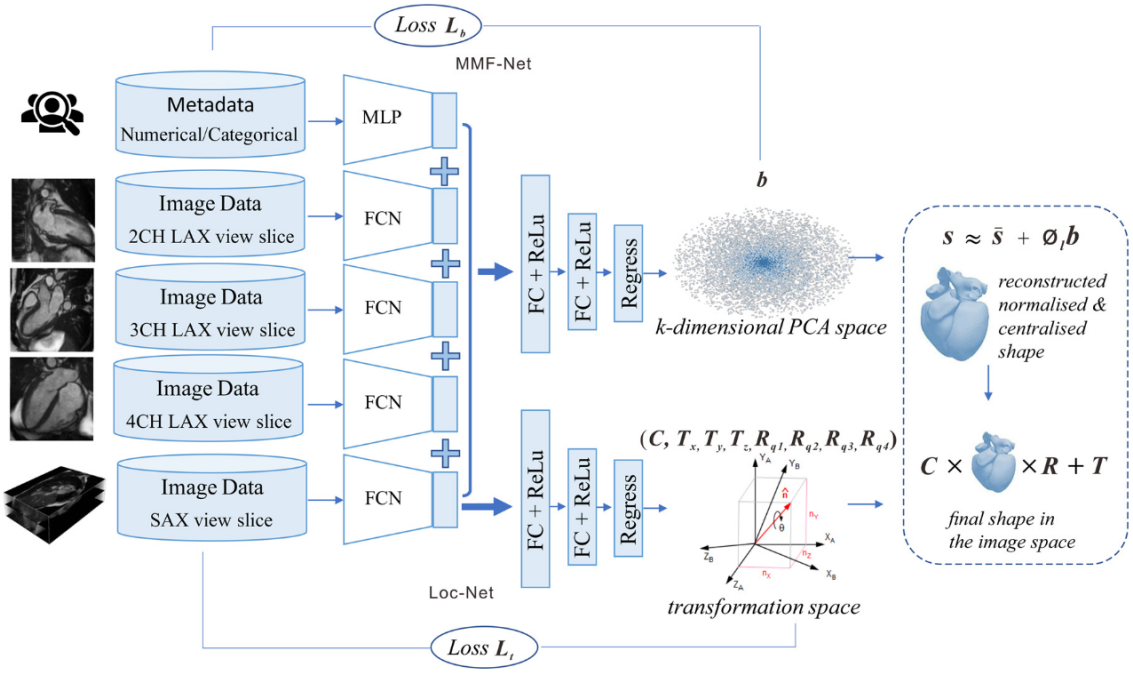

Figure 3.1: Scheme of the segmentor network MCSI-Net (from[108]).

of the network, and $f(\cdot)$ denotes the loss functions used to measure the discrepancy between the predicted values, $\hat{b}_j$ or $\hat{t}_l$, and the reference ones, b_j and t_l. For $L_{\text{Loc-Net}}$, an L_2 norm is utilised for the translation and scaling parameters whereas, for the rotation parameters, the geodesic distance is employed (considering the associated spherical topology). For the term $L_{\text{MMF-Net}}$, the L_2 norm is employed in all cases.

As shown in Figure 3.1, a multi-layer perceptron (MLP) is used to learn features from patient data and integrate it with the extracted image-based features. However, for this work, a different version of the network, relying solely on images as input, was employed.

The appearance information from the different views is concatenated into a single vector and fed into a fully connected layer, with ReLU activation functions, so that, by minimising the loss function, they produce the first parameters in PCA space, which describe the 3D shape of the cardiac chambers. To capture 99.7% of shape variability in the training dataset, K is set to 70 and only those parameters are regressed. Similarly, in the Loc-Net, three different fully connected layers are applied to predict 8 parameters: the scale (C), four rotation parameters in a quaternion representation (R_{q_1}, R_{q_2}, R_{q_3}, R_{q_4}) and three translation parameters (T_x, T_y, T_z).

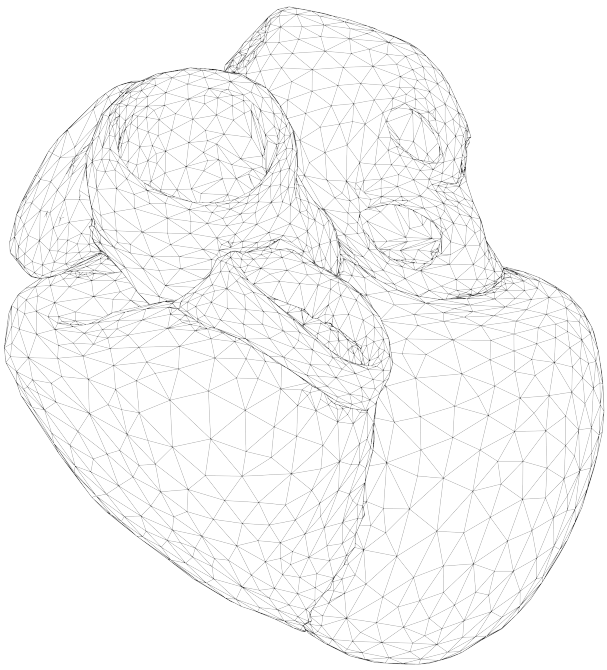

Figure 3.2: Anterior view of the cardiac atlas used in this work (decimated to 10% of the original vertices).

Training data generation

Given that the training labels for the MCSI-Net algorithms are the b components from a point distribution model (PDM), and this model must be fit using registered point clouds, contours had to be registered to a cardiac atlas.

Ground truth: manual contours. Cardiologists Steffen Petersen and Steffan Piechnik conducted manual segmentation on a set of 4,700 subjects, at two timeframes: end-diastole and end-systole [75]. To perform this task, the software tool `cvi42` was utilised. Segmentation was carried out on a slice-by-slice manner, and it consisted of delineating the contours of the myocardium of different cardiac chambers: for the left ventricle, both the endo and epicardial surfaces were obtained, whereas for the rest of the chambers, only the endocardial surface was delineated. Only atrial contours were delineated from LAX views. Note that the ascending aorta was *not* delineated in any view.

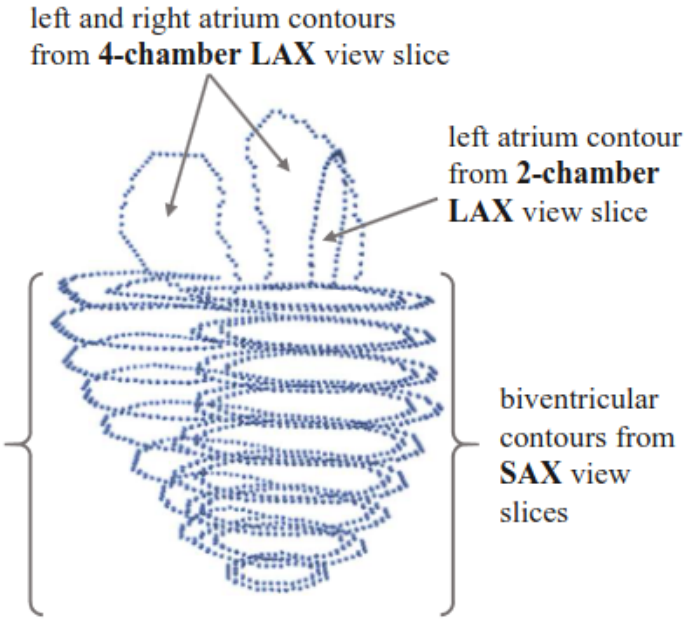

Figure 3.3: Exemplar of manual contours produced from a stack of SAX slices and 3 and 2 LAX slices (from [108]).

Mesh-to-contour registration. This part of the work was conducted by Dr. Nishant Ravikumar. It consisted on registering the chosen cardiac atlas to the stack of 2D contours. For this, the gCPD algorithm was utilised, which performs affine and non-rigid registration, and where shapes are represented by 6-dimensional vectors created by concatenating the spatial coordinates of the vertices with the associated surface normal vector, i.e., $[x, y, z, n_x, n_y, n_z]$ [83].

The cardiac atlas utilised for this work was produced by Rodero et al. [85]. It consists of 194,541 vertices encompassing the four cardiac chambers, as well as the ascending portion of the aortic artery. It is depicted in 3.2 using a decimation of 10%.

Point distribution model (PDM). Shape PCA is performed on the meshes obtained for the 9,400 volumes. The principal components explaining 99.7% of the variability (measured as the mean-squared vertex-wise deviation) were kept, which amounts to 70 principal components.

3.4 Segmentation pipeline

I executed the segmentation pipeline on the ARC4 HPC cluster of the University of Leeds. This pipeline consists of four steps:

Organ	**Count** (full resolution)	**Count** (decimated to 10%)
LV	52,193	4,396
RV	44,556	3,981
RA	29,928	3,059
LA	28,143	3,099
Aorta	19,608	2252
PA	6,209	711
TVP	3,668	456
AVP	2,513	340
MVP	2,447	365
PVP	2,372	301
PV1	664	99
PV2	371	57
PV3	265	43
PV4	396	60
PV5	372	71
PV6	414	86
PV7	422	79
Total	**194,541**	19,455

Table 3.1: Point counts by partition. PA represents the pulmonary artery PVN partitions correspond to the 7 pulmonary veins. Finally, TVP, AVP, MVP and PVP denote the tricuspid, aortic, mitral and pulmonary valves, respectively.

- **Data pre-processing**: this part consists of processing the raw data files from the UKB (in DICOM format) and establishing which view, slice and timeframe they correspond to. This part was carried out through custom Python scripts which have used code in this GitHub repository, by Wenjia Bai, as a useful reference.

- **ROI cropping**: the MCSI-Net model used for this work was trained with SAX slices cropped around the location of the heart, hence removing irrelevant information from the image and assisting with the network's process of learning. Therefore, during inference the same procedure had to be applied. This was carried out by a deep learning model developed in TensorFlow 1 and trained by Dr. Rahman Attar.

- **Inference of b-values and transformation parameters**: the PDM coefficients (b-values) are estimated, as well as the rotation, translation and scaling parameters that are necessary to bring the mesh back to the image space.

- **Mesh generation and downsampling**: the point clouds are generated from the b-values and the transformation parameters. In order to keep the file sizes manageable, the meshes had to be decimated to 10% of their original vertices, before saving them into permanent storage.

The dataset used in this study spans the full UK Biobank cine-CMR database (with over 62,000 subjects) available by June 2023 and hence constitutes the largest study on this dataset conducted hitherto. The whole process, from start to finish, can be carried out in approximately 5 days by parallel execution on 100-150 CPU nodes of ARC4, for the full dataset.

3.4.1 Mesh downsampling

A simple calculation shows that the size of the whole population in the original resolution is somewhat inconvenient: indeed, taking into account that each vertex has three spatial coordinates, and assuming that each coordinate is stored as a simple precision float (4 bytes), the file size for the point cloud for a single subject and time frame would be $3 \times 194,541 \times 4\,\mathrm{B} = 2.226\,\mathrm{MB}$. This amounts to almost 7 TB when considering the full population of over 60,000 subjects and the full cardiac cycle, composed of 50 time frames. While not entirely prohibitive, such storage demands significantly slow down data loading during training. For this reason, I chose to decimate the meshes, to keep only about 10% of the original vertices. The algorithm used to perform this decimation is described in **Section 4.6** but, briefly, it takes a template mesh and produces a decimation matrix and a new surface topology (i.e. triangular faces). Qualitative assessment of the meshes prior and after the decimation reveals that this procedure does not remove important information.

3.4.2 Removal of poor quality segmentations

Subjects were removed based on the quality of the segmentations. For this, features of the volume-time curves were analysed to assess their plausibility. The following exclusion rules were applied:

- LV volume being 10% higher than LV at time 0 (assumed to correspond to LVEDV) in more than 5 time frames.
- In addition, estimated t_{ES} being below 0.2 and above 0.7 of the cardiac cycle length (with the average being 0.38).
- Finally, LVV below 10 ml for any time frame.

This process led to the exclusion of 1,015 subjects.

3.4.3 Procrustes analysis

Since for this work we are interested in purely anatomical features, pose parameters need to be removed. This is done via minimisation of the deviations between each shape and a template shape, over the set of all possible rigid-body transformations (translations and rotations). We argue that scaling transformations are not desired since the scale of the shapes contains biologically relevant information. This process of removal of pose parameters via rigid-body transformations is called *partial Procrustes* analysis, to distinguish it from plain Procrustes analysis which does include scaling transformations.

To perform this operation, the Python package `SciPy` is utilised, in particular its submodule `linalg` which implements partial Procrustes via the `orthogonal_procrustes` function.

3.5 Code availability

As part of this thesis, a codebase for manipulating cardiac meshes using this atlas has been developed and made publicly available on GitHub. I have named this package `CardioMesh`. The repository is available at `www.github.com/rbonazzola/CardioMesh` and is used extensively in the rest of the work in this dissertation.

Chapter 4

Representation learning on shapes

This chapter describes the methods used in this work to perform dimensionality reduction on populations of cardiac shapes:

- **Section 4.1** provides an introduction to the concepts of dimensionality reduction and representation learning.
- **Section 4.2** describes the traditional PCA method, and how it can be applied to shapes (represented as point clouds),
- **Section 4.3** describes autoencoders and variational autoencoders.
- **Section 4.4** covers traditional convolutional neural networks (CNNs).
- **Section 4.5** introduces graph neural networks and, in particular, spectral graph convolutions based on Chebyshev polynomials.
- **Section 4.6** details the specific type of autoencoder used in this work, convolutional mesh autoencoders (CoMA), which leverage the topological relationships between the vertices of the mesh to produce an efficient low-dimensional representation.

4.1 Introduction

Representation learning and dimensionality reduction are fundamental concepts in machine learning and play crucial roles in data analysis.

Dimensionality reduction (DR) aims to obtain a compressed representation of original data

through feature selection (removing irrelevant features) and feature extraction (transforming data into suitable features). DR involves finding a function $f : \Omega \subset \mathbb{R}^M \to \mathbb{R}^K$, where $K < M$ (and, typically, also $K \ll M$), such that the relevant information contained in a given population of vectors $\{\mathbf{X}_i\}_{i=1}^N$ is efficiently condensed.

Representation learning can be viewed as an unsupervised DR problem, where the goal is to learn general representations of data without the need for associated labels. For this, the encoder-decoder framework is normally used. In such paradigm, there is a pair of encoding and decoding functions, $E_\theta : \mathbb{R}^M \to \mathbb{R}^K$ and $D_\phi : \mathbb{R}^K \to \mathbb{R}^M$ that are parameterised by a set of learnable coefficients θ and ϕ, respectively. $K \in \mathbb{N}$ is the size of the latent space, and is usually chosen so that $K \ll M$ (hence the reduction in dimension).

Linear methods find a linear transform of the input features such that the resulting features efficiently condense the information (e.g. maximize the variance explained). We will start by examining one example of these methods, principal component analysis, to then move on to non-linear neural-based methods.

4.2 Principal component analysis (PCA)

PCA is a standard linear technique for DR [73]. In terms of the encoder-decoder framework detailed above, it can be obtained by requiring D and E to be linear transformations and using the norm L_2, in addition to imposing an orthogonality constraint on the latent vectors [44].

The idea is to find a basis of vectors $\mathcal{B}_K = \{\mathbf{v}_i\}_{i=1}^K \subset \mathbb{R}^{3M}$ for a fixed $K < 3M$. The K-dimensional linear subspace generated by $\mathcal{B}_K$ captures as much data variability as possible. It can be shown that this basis corresponds to the K eigenvectors of $\mathbf{C}$ with the largest eigenvalues; that is, if $\mathbf{C} = U^t \Lambda U$ where $\Lambda_{ij} = \delta_{ij}\lambda_i$ and $\lambda_i \geq \lambda_j$ if $i \leq j$, then $\mathcal{B}_K = \{\mathbf{u}_i\}_{i=1}^K$. δ_{ij} is the Kronecker delta, which equals 1 if $i = j$ and 0 otherwise, and $\mathbf{u}_i$ is the i-th column of U.

Shape PCA. Let $X \in \mathbb{R}^{M\times 3}$ denote the 3D coordinate matrix for an M-point point cloud. We define, for convenience, the vectorised form of the shapes, $\mathbf{s}_i = \big(x_{i1}, y_{i1}, z_{i1}, ..., x_{iM}, y_{iM}, z_{iM}\big) \in \mathbb{R}^{3M}$. We refer to this approach as *shape PCA* throughout the text. Given a set of 3D shapes $\mathbb{S} = \{\mathbf{s}_i\}_{i=1}^N$, we derive the mean shape $\bar{\mathbf{s}}$ and the shape covariance matrix $\mathbf{C}$:

$$\bar{\mathbf{s}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i, \tag{4.1}$$

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^t. \tag{4.2}$$

Shape PCA can then be performed by finding the top K eigenvectors of the shape covariance matrix, $\mathbf{C}$. In this work, we implemented shape PCA using the Python `scikit-learn` package, which actually performs singular value decomposition of the data matrix (composed of the $\mathbf{s}_i$ vectors).

We approximate a given instance of a shape, $\mathbf{s}_i$ by projecting the point cloud onto the first t eigenvectors to produce the coefficients $\mathbf{b}_i \in \mathbb{R}^K$:

$$\mathbf{s}_i \approx \bar{\mathbf{s}} + \sum_{k=1}^{K} b_i^{(k)}\mathbf{u}_k \tag{4.3}$$

4.3 Autoencoders

In an *autoencoder* network, or encoder-decoder network, both the encoding and decoding functions are feedforward neural networks. Autoencoders have been proposed by Rumelhart *et al.* in [87] as early as 1986, and they have been successful in many applications, like: representation learning, image denoising, feature extraction, anomaly detection and missing data imputation [101, 74, 97].

Relation between PCA and autoencoders. PCA can be thought of as an autoencoder with only one layer in the encoder and one in the decoder (without non-linear activation function), with an extra imposition of orthogonality between the latent vectors.

4.3.1 Variational autoencoders (VAE)

In 2013, Kingma and Welling [50] proposed auto-encoding variational Bayes, later called *variational autoencoders* (VAEs), a method that combines autoencoders with variational Bayesian inference to learn parameters in probabilistic models with continuous latent variables and intractable posterior distributions. This approach involves re-parameterising the variational lower

bound to create a simple differentiable and unbiased estimator that can be optimised via common deep learning algorithms (such as stochastic gradient descent or Adam).

The variational lower bound for a VAE, also called evidence lower bound or ELBO, is given by:

$$\mathcal{L}_{\text{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\big) \tag{4.4}$$

where $p(\mathbf{z})$ is the prior distribution over the latent factors $\mathbf{z}$, and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ represent the encoder and decoder networks, respectively, now taken to be probabilistic.

Here, $\boldsymbol{\theta}$ represents the parameters of the generative model, $\boldsymbol{\phi}$ represents the parameters of the inference model, $\mathbf{x}$ is the input data, $\mathbf{z}$ is a latent variable, $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is the generative distribution, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is the variational distribution, and KL represents the Kullback-Leibler divergence.

In the VAE framework, during the training phase the encoder maps the input into a probability distribution instead of a fixed vector. During training, for the j-th latent variable (with $1 \leq z_j \leq n_z$) two quantities are learned, $\mu_j(\mathbf{x})$ and $\sigma_j(\mathbf{x})$, and a realisation z_j of the random variable $Z_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ is produced and passed through the decoder to generate the output object. The aforementioned KL-divergence term is then used to encourage the variational approximate posterior to be a multivariate Gaussian with a diagonal covariance structure. The regularisation term is computed as:

$$\begin{aligned} \Omega(\mathbb{X}_{\text{train}}|\theta, \phi) &= \mathbb{E}_{\mathbf{s}\sim\hat{p}_{\text{train}}}\ D_{\text{KL}}\Big(q_\theta(\mathbf{z}|\mathbf{x})||\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbb{1}_{n_z})\Big) \\ &= \mathbb{E}_{\mathbf{s}\sim\hat{p}_{\text{train}}} \frac{-1}{2n_z} \sum_{j=1}^{n_z} \Big(\log \sigma_j^2 - \sigma_j^2 - \mu_j^2 + 1\Big) \end{aligned} \tag{4.5}$$

where $\mathbb{1}_n$ is the $n \times n$ identity matrix, $D_{\text{KL}}(p||q)$ is the KL divergence between the probability distributions p and q, and $\hat{p}_{\text{train}}$ is the empirical probability distribution associated with $\mathbb{S}_{\text{train}}$. $D_{\text{KL}}(p||q) := \int p(x) \ln \frac{p(x)}{q(x)} dp(x)$. The last equality in Equation 4.5 arises from the formula for the KL divergence between two normal distributions, where the second is also standardised. During inference, the mode of the latent distribution, $\boldsymbol{\mu}(\mathbf{x})$, is the latent representation of the object $\mathbf{x}$.

4.4 Convolutional neural networks

Before introducing mesh-based convolutional networks, we briefly review standard convolutional neural networks (CNNs) to motivate their generalisation to irregular domains.

This section describes *convolutional layers*, which are mainly used for processing image data. Although not used in this work, where we will work with 3D mesh data (except for the segmentation part previously discussed), it provides a convenient starting point to discuss mesh-convolutional layers, which we do below, in **Section 4.5**.

Following the discussion in [80], we begin by noticing that images have three properties that highlight the limitations of fully connected networks and motivate alternative architectures. Firstly, the usually high dimensionality of images would imply an impractically large network, in terms of memory, execution time and data requirements for training.

Secondly, nearby pixels are highly correlated. However, fully connected networks treat input data features equally and have no notion of locality.

Finally, the interpretation (or semantic content) of an image is not altered by certain geometric transformations. For example, a small shift or rotation in an image should not affect a neural classifier that aims to detect whether an object, say a tree, is present in an image.

Convolutional layers solve these issues by processing each region of the image independently, through so-called *filters* or *kernels* that are shared across the image. Hence, these layers use fewer parameters (by virtue of the small size of kernels) than their fully-connected counterparts, exploit spatial relationships (the pixels cannot be permuted randomly without affecting the output) and the "interpretation" of pixels needs not be re-learned at every different position (because the kernels are shared). Networks composed predominantly of convolutional layers are referred to as *convolutional neural networks* (CNNs).

To build intuition, we detail the application of a 2D convolutional filter to a 2D signal as an example. Let us assume we have a 2D signal, x_{ij}, and a 3×3 kernel $K \in \mathbb{R}^{3\times 3}$ (kernels typically have odd dimensions to allow centering at a pixel). The output of 2D convolution, h_{ij}, is given by

$$h_{ij} = \mathrm{a}\left(\beta + \sum_{m=1}^{3}\sum_{n=1}^{3} \Omega_{mn} x_{i+m-2,j+n-2}\right) \tag{4.6}$$

where a$(\cdot)$ is an activation function, β is a bias. The indices m and n displace the kernel over a region (a 3×3 square) centered at x_{ij}. A convolution is, thus, simply a weighted sum over that region, which produces an output h_{ij} for each position (i, j) in the input.

In order to produce rich intermediate representations that account for diverse patterns in the input data, a battery of several kernels is normally utilised at each layer. The output of the battery of kernels are usually called *feature maps*, highlighting their image-like nature.

4.5 Graph neural networks

In this section, we introduce the concept of graph neural networks which will be essential to our work.

Graphs and meshes

A *graph* can be defined as the tuple $G = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ where $\mathcal{V}$ are a set of vertices or nodes, $\mathcal{E}$ are the edges or links, and $\mathcal{W}$ the strengths of these links. $v_i \in \mathcal{V}$ is a node, $e_{ij} \in \mathcal{E}$ denotes an edge connecting vertices v_i and v_j. In this work, we will always take the strength of the edges to be equal for all the edges.

A graph is a very flexible structure that can be used to represent a wealth of objects: molecules, different types of networks (e.g. social networks, gene expression networks, etc.). A *3D mesh* is a particular kind of graph, which represents the discretised version of a surface. In this case, vertices are endowed with positional information, as a tuple of (x, y, z) coordinates. The set of vertices, without the connectivity information, constitutes a *point cloud*. In a triangular surface mesh, edges define triangular faces, and every edge belongs to either one or two faces. For closed (also called *watertight*) meshes, each edge belongs to exactly two faces. In this dissertation, we will also utilise the term *registered meshes* to denote population of meshes for which the vertex cardinality and connectivity is the same and only the vertex coordinates change.

4.5.1 Convolutional graph neural networks

Bronstein et al. [21] give a comprehensive overview of generalizations of CNNs on non-Euclidean domains, including meshes and graphs. These methods can be classified in two broad groups: spatial or spectral. The approach proposed in this work belongs to the latter category, which relies on expressing the features in the Fourier basis of the graph, as explained below.

Among the spatial family of graph convolutions, *spiral convolutions* [20] can also be readily applied to registered meshes. They have not been employed in this study due to time constraints, however we wonder whether including the type of graph convolution as an additional hyperparameter could be effective in enhancing the diversity of CoMA ensembles (see Chapter 6).

4.5.2 Spectral graph convolutions

Bruna et al. [22] exploit the connection between the graph Laplacian and the Fourier basis, which leads to spectral kernels that generalise convolutions to arbitrary graphs. Defferrard et al. [33] approximate the spectral filters by truncated Chebyshev polynomials, which avoids explicitly computing the Laplacian eigenvectors, as we will see later in this subsection.

Laplace-Beltrami operator. The Laplace-Beltrami operator $\mathcal{L}$ (or, more simply, the Laplacian) of a graph with adjacency matrix A is defined as $\mathcal{L} := D - A$, where D is the degree matrix, i.e. a diagonal matrix with $D_{ii} := \sum_j A_{ij}$ being the number of edges that connect to vertex i.

The Fourier basis of the graph can be obtained by diagonalising the Laplace operator, $\mathcal{L} = U^t \Lambda U$. The columns of U constitute the Fourier basis, and the operation of convolution $\star$ for a graph can be defined in the following manner:

$$x \star y := U(U^t x \odot U^t y), \tag{4.7}$$

where $\odot$ is the element-wise product (also known as Hadamard product), and x and y are arbitrary functions defined over the vertices of the graph. All spectral methods rely on this definition of convolution and differ from one another in the specific filter used.

Polynomial filters In polynomial filters, the filter is given by a polynomial function applied to the Laplacian $\mathcal{L}$:

$$g_\theta(\mathcal{L}) = \sum_{k=0}^{K} \theta_k \mathcal{L}^k, \tag{4.8}$$

For functions of this type,

$$y = g_\theta(\mathcal{L})x = g_\theta(U\Lambda U^T)x = U g_\theta(\Lambda)U^T x, \tag{4.9}$$

Polynomial filters, despite their spectral formulation, have the characteristic of being local. To see this, let's consider a δ-like signal centered at vertex i, δ_i (i.e. the signal's value is 1 at vertex i and 0 elsewhere on the graph). If we apply the filter on this signal, we get:

$$(g_\theta(\mathcal{L})\delta_i)_j = (g_\theta(\mathcal{L}))_{i,j} = \sum_k \theta_k(\mathcal{L}^k)_{i,j} \tag{4.10}$$

Given that k-th powers of the Laplacian connect vertices being at most k edges apart, we conclude that the filter defined by k-degree polynomials are local.

Chebyshev convolutions Note that the computational complexity of the polynomial convolutions described earlier is $\mathcal{O}(n^2)$, due to the change of basis which requires matrix multiplications (with U and U^T).

In this work, a parameterisation proposed in [32] will be used. The said method is based on the Chebyshev family of polynomials $\{T_i\}$. The kernel g_ξ is defined as:

$$g_\xi(\mathcal{L}) = \sum_{i=1}^{K} \xi_i T_i(\mathcal{L}). \tag{4.11}$$

K is the highest degree of the polynomials considered (in this work $K = 6$). Chebyshev polynomials have the advantage of being computable recursively through the relation $T_i(x) = xT_{i-1}(x) - T_{i-2}(x)$ and the base cases $T_1(x) = 1$ and $T_2(x) = x$.

Other similar formulations have been proposed after this one, which employ different families of recursive polynomials, such as Laguerre polynomials. In this work, we have not explored these other formulations since no advantage should be expected *a priori*, nor has any been demonstrated empirically by previous work.

4.6 Convolutional mesh autoencoders (CoMA)

Inspired by work on unsupervised geometric deep learning for facial meshes [81], we propose constructing a convolutional mesh-autoencoder which uses spectral convolutions [32] to learn non-linear and low-dimensional representations of cardiac mesh structures. Here, each layer of the encoder and decoder implements convolution operations parameterised by the graph Laplacian, to leverage information about the local context of each vertex (similarly to the discussion

in the section regarding traditional CNNs). In order to learn global features, a hierarchical approach is used where each layer of the encoder and decoder implements downsampling and upsampling operations, respectively. Figure 4.1 depicts the full CoMA network and its application to left-ventricular meshes.

Note that this methodology was originally proposed in its entirety in [81] and is not a methodological contribution of this thesis. As a novelty, in Chapter 6 we apply it for the purpose of phenotyping cardiac meshes, using a deep ensemble framework, to discover new genetic associations via GWAS; whereas in Chapter 7 we extend it to consider sequences of meshes with periodic motion.

In the remainder of this section, we detail the missing component to build a mesh-convolutional autoencoder: the mesh downsampling and upsampling operations.

4.6.1 Mesh downsampling and upsampling

In order to build a mesh autoencoder, we also need a way to both downsample and upsample the meshes, in the encoder and the decoder, respectively. Following [81], we utilise the quadric decimation algorithm, as well as an upsampling operation proposed in that work.

Quadric decimation This method was proposed in [42]. It is an iterative procedure in which vertices are removed one at a time, and a new connectivity is computed between the removed vertex's neighbors. At each iteration, the vertex is selected so that a quantity, called the *quadric error*, is minimised. The procedure does not alter the coordinates of the remaining vertices.

A template mesh is used to generate a downsampling matrix $Q_d \in \{0,1\}^{M_2 \times M_1}$, where M_1 is the original number vertices, and $M_1 - M_2$ is the number of vertices to be removed. The matrix Q_d is determined by specifying the template mesh and $M_1 - M_2$. In CoMA, each of the Chebyshev convolutions is followed by such decimation procedure.

As mentioned in **Section 3.2**, this same procedure was conducted to obtain a 1:10 downsampling matrix to decimate the original atlas. The generation of this matrix took approximately 25 hours on an AMD EPYC 7742 CPU. This matrix is provided in the CardioMesh repository (see Code Availability section).

Upsampling. During the generation of the downsampling matrix, the upsampling matrix $Q_u \in [0,1]^{M_1 \times M_2}$ is simultaneously generated. This is a matrix that receives as input a point cloud represented as a matrix of coordinates of size $M_2 \times 3$ and outputs another, denser, point cloud as a matrix of size $M_1 \times 3$ (recall that $M_1 > M_2$). The M_2 vertices will keep their coordinates while $M_1 - M_2$ will be added.

The generation of Q_u takes place according to the following algorithm. As explained before, a set of $M_1 - M_2$ vertices are discarded by Q_d, and for each such vertex the closest triangle (v_i, v_j, v_k) in the downsampled mesh is identified, and the coordinates of the removed vertex is approximated as a convex combination of the coordinates of the vertices of such triangle, that is $\tilde{v}_p = w_{pi}v_i + w_{pj}v_j + w_{pk}v_k$ with $w_{pi} + w_{pj} + w_{pk} = 1$. The upsampling matrix Q_u is then given by: $Q_u(p,i) = w_{pi}$, $Q_u(p,j) = w_{pj}$ and $Q_u(p,k) = w_{pk}$ and $Q_u(p,\cdot) = 0$ otherwise.

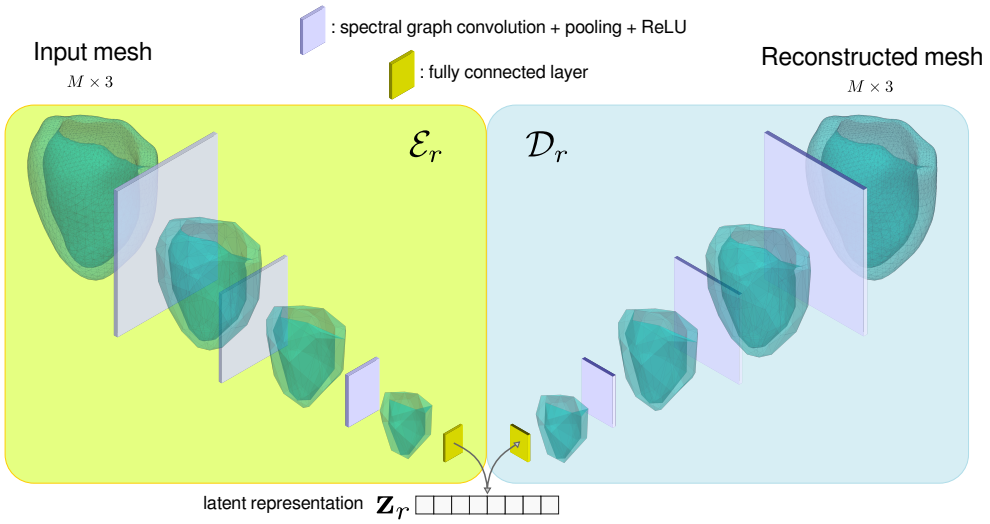

Figure 4.1: A graph-convolutional autoencoder is trained and applied to our set of cardiac meshes (in the figure, left-ventricular meshes are depicted) to produce low-dimensional representations of these shapes. In each layer, a representation with fewer vertices is obtained.

In the next chapter, we leverage these learned representations to extract phenotypes and investigate their genetic basis.

4.7 Deep ensembles

In machine learning, ensemble methods create multiple models and then combine them to produce improved results, typically in the context of predictive models (either classifiers or

regressors). It may involve fitting many models on different samples of the same dataset and aggregating the predictions in some way (*bagging*), or fitting different model types on the same data and, again, combining the predictions (*stacking*). They are widely used in machine learning due to their ability to produce more accurate solutions than a single model.

To create a deep ensemble, multiple deep neural networks are trained independently on the same dataset. [37] has studied in detail the characteristics of such ensembles, reaching the conclusion that often a simple variation in weight initialisation can lead to different local minima in the loss function, that correspond to different, non-trivially connected, local optima in the loss landscape. In Chapter 6, we exploit this insight to build diverse ensembles of CoMA models which exhaust the phenotypic variability more effectively than individual runs, allowing to discover more genetic associations.

4.8 Summary and outlook

In this chapter, we presented a range of methods for dimensionality reduction and representation learning, focusing on their application to shapes represented as point clouds and meshes. Starting from linear techniques like PCA and moving towards non-linear neural-based approaches, we discussed how convolutional mesh autoencoders (CoMA) can effectively capture shape variability by leveraging mesh topology and spectral graph convolutions. We also introduced deep ensembles as a way to enhance the expressivity and diversity of the learned representations.

In the next chapter, we make use of these latent representations to define phenotypes and explore their genetic basis through genome-wide association studies.

Chapter 5

Traditional CMR-derived phenotypes and their genetic associations

In this chapter, we leverage our cardiac segmentations to derive traditional cardiac phenotypes, similarly to previous work. These phenotypes are studied in the context of genome-wide association studies, and we find new genetic associations with respect to recent publications by virtue of our greater sample size.

Part of the results in this chapter, in particular those concerning genetic associations of LV phenotypes at end-diastole, were published on and adapted from a paper in the journal *Nature Machine Intelligence* [17], where they served as a baseline for our ensemble-based approach that we presented there and which we discuss at length in Chapter 6. The rest of the findings, including those concerning LV wall thickness and thickening, have not been published elsewhere to date.

5.1 Quantification of the cardiac chambers

In this section, I detail the procedure undertaken to obtain the different mesh-derived values. Note that, due to the characteristics of the atlas used, a different approach was conducted for ventricles and for atria.

5.1.1 Ventricular volumes

Mesh voxelisation This approach relies on the Python library `Trimesh` for 3D triangular mesh manipulation, to create a voxelised version of the mesh, more specifically an object belonging to the `trimesh.voxel.VoxelGrid` class. This is a grid of $W \times D \times H$ voxels with a boolean value: 1 for voxels that overlap the myocardium and 0 otherwise. By counting the the 1-valued voxels, the volume can be estimated.

As described in Training data generation, the atlas contains surfaces that represent the different valves of the ventricular chambers and allow us to "close" these chambers. In this way, after voxelising the mesh, a set of voxels enclosing the blood pools are obtained. By using the method `fill` provided by the `trimesh.voxel.VoxelGrid` class, it is possible to identify the voxels belonging to the interior of the chambers (the blood pool). Finally, by simply counting the numbers of 1-valued voxels contained in the interior, the volume of the blood pool for each ventricle can be estimated by difference.

5.1.2 Atrial volumes

Given the characteristics of our atlas, it was not possible to obtain the volumes of the atria by means of the above method. In particular, the issue with these chambers is that they do not count with surfaces closing the cavity for different pulmonary veins and arteries, hence the voxelisation does not provide a closed arrangement of voxels that allow to compute the volume of the blood pool by the above procedure consisting of "filling" the interior.

For this reason, we use an alternative method for estimating the volume, which is obtaining the convex hull of the LA and RA shapes, to then find the volume of the convex hull. Since these shapes are nearly convex, this procedure only slightly overestimates the volume.

We refer to the maximum and minimum value for the volumes across the cardiac cycle as $\mathrm{LAV}_{\mathrm{min}}$ and $\mathrm{LAV}_{\mathrm{max}}$, respectively, and likewise for the right atrium, $\mathrm{RAV}_{\mathrm{min}}$ and $\mathrm{RAV}_{\mathrm{max}}$.

These phenotypes are not analysed in this chapter, however they are studied in Chapter 7 in relation with dynamic atrial phenotypes derived in an unsupervised way.

5.1.3 Volume-derived quantities

Stroke volumes (computed for LV and RV) were computed as the difference between EDV and ESV. The ES timeframe, t_{ES}, was estimated as $t_{\mathrm{ES}} \approx \mathrm{argmin}_t\big(\mathrm{LVV}(t)+\mathrm{RVV}(t)\big)$. We note that stroke volumes are expected to be the same for LV and RV, since the chambers form part of a circuit series; we add a prefix LV or RV, however, to indicate that they were calculated with left-ventricular or right-ventricular volumes, respectively.

As detailed in the previous chapter, for LV the segmentation approach yields endocardial and epicardial surfaces, from which the myocardial mass of this chamber was calculated. This quantity is abbreviated as LVM. Also, the left-ventricular mass-to-volume ratio (LVMVR) was computed as a the ratio between LVM and LVEDV (following [9]).

5.1.4 Left-ventricular sphericity

The importance of left-ventricular sphericity index has been known for a long time, however this phenotype has only recently been studied in the context of population imaging studies.

As discussed in **section** 2.5 on related work, Vukadinovic *et al.* in [105] calculate the sphericity index based on the long-axis view, as the ratio between the LV short axis and LV long axis. Instead, we propose a different mesh-based method to compute the sphericity index. It relies on the traditional definition of the sphericity index of a body with surface area A and volume V, given by the ratio of the area of a sphere of volume V, $A_{\mathrm{sph}}(V) = (36\pi V)^{2/3}$, and the actual surface area of the body, A. Since a sphere is the solid with minimal surface area given a fixed volume, this number lies between 0 and 1.

Recall that the left ventricle is represented by a cup-shaped mesh. Instead, we are rather interested in the sphericity index of the smallest convex envelope, or *convex hull* (CH), of this shape. The sphericity index for the LV at end-diastole (LVEDSph) was then estimated as follows: first, the convex hull of each LVED mesh was obtained and, from it, its surface area A_{CH} and volume V_{CH}. The sphericity index was obtained as $\mathrm{LVEDSph} = (36\pi V_{\mathrm{CH}})^{2/3}/A_{\mathrm{CH}}$. To obtain the CH the submodule `Spatial` from the `SciPy` Python library was used (version 1.9.3), which provides both the vertices and the faces of the CH mesh; whereas for the associated areas and volumes, the `Vedo` library was employed. (For reference, the average LVEDSph is around 0.95 as we will see later.)

5.1.5 LV myocardial thickness and thickness-derived values

Myocardial thickness was computed in a vertex-wise manner, as the distance between a given vertex in the LV epicardial (respectively, endocardial) surface and the closest vertex on the endocardial (respectively, epicardial surface). This vertex-wise thickness was averaged across all the vertices in an AHA segment to obtain the segment-wise average. Per-segment wall thicknesses were computed at ED and ES, and wall thickening was obtained therefrom as the difference between both values, magnitude that which we term *absolute wall thickening. Relative wall thickening* was obtained as the ratio between absolute wall thickening and wall thickness at ED. We also compute a *dimensionless* wall thickness value, which we term *relative wall thickness*, by dividing the segment-wise wall thicknesses by the cubic root of the volume of the LV blood pool (either LVEDV or LVESV depending on the cardiac phase being considered).

LV AHA parcellation. In this paragraph, we detail the procedure followed to parcellate the LV atlas into the 17 AHA segments. The procedure is undertaken on a template mesh, obtained as the average of all the shapes, and the vertex-wise AHA segment labels (1 through 17) are then propagated to all the subjects by using the vertex correspondence.

The inertia tensor I_{LV} for LV was computed for the reference shape:

$$I_{\mathrm{LV}} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{x}_j \mathbf{x}_j^T \tag{5.1}$$

By diagonalisation of the inertia tensor, the three principal axes are obtained: $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$. The first principal axis, $\mathbf{v}_1$, is in the longitudinal direction and, along with the second principal axis $\mathbf{v}_2$, defines the plane that bisects the root of the ascending aortic artery.

Parcellation of the LV according to the 17 AHA segments was carried out by splitting the mesh along the longitudinal axis into five regions according to the following list of relative coordinates [20%, 40%, 55%, 70%]. The top region, including the aortic root, was removed, while the remaining four slices corresponded to: the apex (0-20%), the apical slice (20-40%), the mid slice (40-55%) and the basal slice (55-70%). The sectors corresponding to each segment in each slice were obtained by splitting the slice into sixths (mid and basal slices) or fourths (apical slice).

Figure 5.1: The 17 left-ventricular AHA segments.

5.2 Traditional indices: their relation to demographic data

Male proportion	48.29%
Age (years)	59.3 ± 14.9
Height in males (cm)	176.4 ± 13.1
Height females (cm)	163.1 ± 12.2
BMI in males (kg/m^2)	27.1 ± 7.5
BMI in females (kg/m^2)	26.1 ± 9.0
SBP in males (mmHg)	140.2 ± 33.9
SBP in females (mmHg)	133.5 ± 30.0
DBP in males (mmHg)	82.4 ± 17.1
DBP in females (mmHg)	78.4 ± 17.6

Table 5.1: Summary statistics of the demographic variables of the subsample of 50,000 unrelated British individuals used in this work. Continuous variables are expressed as mean $\pm$ 2 s.d.

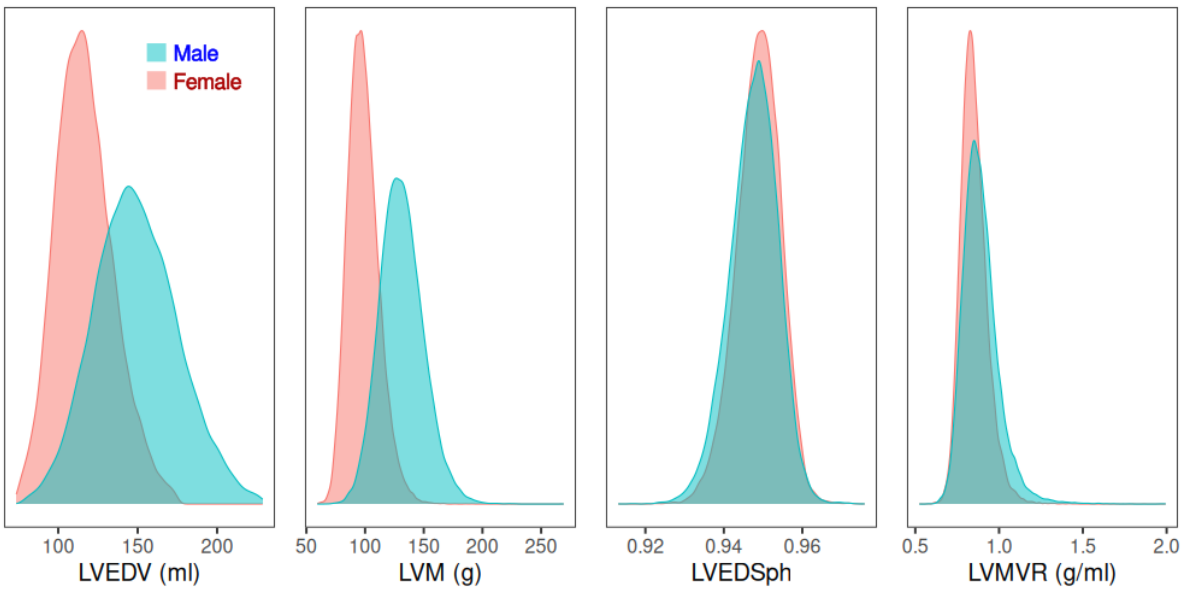

Figure 5.2: Distribution of four LV indices: LVEDV, LVM, LVEDSph, LVMVR distinguished by genetic sex.

Firstly, we investigated the relation between traditional indices and some demographic and anthropometric variables. The variables studied were: genetic sex, age, height and BMI. Figure 5.2 shows the sex-specific distributions of four LV indices via density plots: LVEDV, LVM,

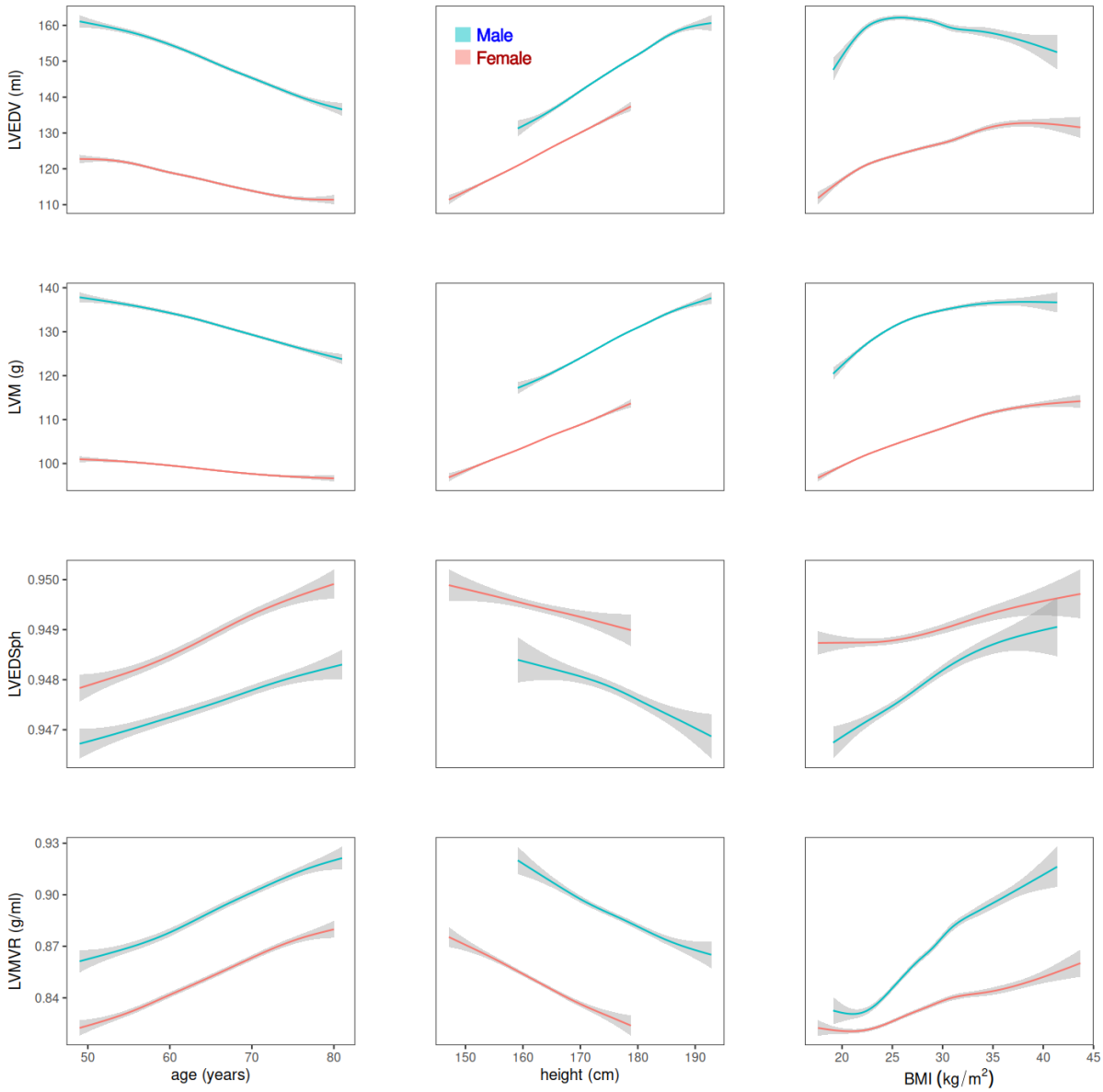

Figure 5.3: Smoothed curve plots depicting the influence of age, height and BMI on different LV indices: LVEDV, LVM, LVEDSph and LVMVR.

LVEDSph and LVMVR, coloured by genetic sex. Fig. 5.3 shows the dependence of the same indices with demographic and anthropometric variables via smoothed curves, also coloured by genetic sex. Note that the raw values are not shown so that the trends can be readily appreciated.

Reassuringly, we reconfirm trends observed in [12], while extending our analysis to also include LV sphericity index. We observed an effect of male genetic sex of +3.0 ml, +3.8 g and +0.0095 g/ml ($p < 10^{-62}$) on LVEDV, LVM and LVMVR after controlling for height, weight, age, body surface area, imaging centre, smoking status and alcohol intake frequency (using terms

of up to second degree including interactions)[1]. Additionally, a small but significant increase ($\hat{\beta} = +3 \times 10^{-4}$, $p = 6.0 \times 10^{-15}$) in LVEDSph was associated with female sex after controlling for the same variables. We have also detected a small but significant increase of LVEDSph with age ($\hat{\beta} = +5.0 \times 10^{-5} \mathrm{yr}^{-1}$, $p = 1.1 \times 10^{-54}$), while a decrease with height ($\hat{\beta} = -6.5 \times 10^{-5}\,\mathrm{cm}^{-1}$, $p = 2.6 \times 10^{-133}$) and an increase with BMI ($\hat{\beta} = +6.2 \times 10^{-5}\,\frac{\mathrm{m}^2}{\mathrm{kg}}$, $p = 5.3 \times 10^{-29}$) are observed (note that we are not controlling for other variables here).

5.3 Genome-wide association studies (GWAS)

In this section, we provide the details of the implementation of GWAS. Unless otherwise stated, these steps are conducted similarly in Chapter 6 and Chapter 7.

5.3.1 Model

According to the traditional GWAS scheme, we tested each genetic variant, $X_i \in \{0, 1, 2\}$, for association with each phenotype y through a univariate linear additive model of genetic effects:

$$Y = \beta_\ell X_\ell + \varepsilon_\ell \tag{5.2}$$

where ε_i is the component not explained by the genotype, assumed to be normally distributed. Note that the α parameter from Eq. 2.1 is not required since inverse rank normalisation produces a centered phenotypic score. The null hypothesis tested is that $\beta_\ell = 0$.

5.3.2 Exclusion of samples

To avoid issues related to population stratification, only unrelated individuals with self-reported British ancestry were included in the study. This produced a sample size of 54,121 individuals. Summary statistics of demographic data from these subsample can be found in Table 5.1. For the results presented in the main text, no individuals were filtered out based on previous diagnoses or image-derived cardiac function parameters (such as ejection fraction).

The above sample was split into two groups, one for discovery and one for replication, which consisted of 48,651 and 5,470 individuals, respectively.

[1] The effect of male genetic sex without adjusting for additional variables is 32.1 ml.

5.3.3 Adjustment for covariates

Before GWAS, the phenotypes were adjusted for a set of covariates, with the rationale given in subsection 2.2.3. In the following, we describe these variables, classifying them as "demographic" and "genetic principal components".

Demographic variables. The demographic variables used as covariates were the following (when appropriate, UKB field codes are provided):

- height (field code 50, `"standing height"`),
- genetic sex (field code 22001, `"genetic sex"`),
- age (field code 21003, `"age when attended assessment centre"`, instance 2 corresponding to the first imaging visit),
- BMI (field code 21001, `"body mass index"`),
- body surface area (BSA), estimated as $0.20247 \times (\text{weight}^{0.425}) \times (\text{height}^{0.725})$, where weight is expressed in kg and height is expressed in meters,
- systolic blood pressure (SBP, field code 4080, with name `"systolic blood pressure"`),
- diastolic blood pressure (DBP, field code 4079, with name `"diastolic blood pressure"`),
- smoking status, categorised into `"current"`, `"previous"` or `"never"` according to field 20116, and
- drinking status: regular alcohol use was defined as a binary variable based on whether the participant reported consumption of alcohol at least three times per week (field 1558).

SBP and DBP were adjusted for those participants taking blood-pressure-lowering effect from the verbal interview (field 20003); SBP was adjusted by adding 15 mmHg, whereas DBP was adjusted by adding 10 mmHg (following the procedure in [9]).

Note that we did not compute BSA-indexed versions of the phenotypes as [78] does, i.e. phenotypes divided by BSA, since we used BSA as a covariate instead.

Genetic principal components. To compute the genomic PC loadings, a similar set of criteria for selecting SNPs as those detailed in the UKB genotyping QC report guide were followed. They are reproduced here for convenience:

- Minor allele frequency $\geq 2.5\%$ and missingness $\leq 1.5\%$. (Checking that HWE holds in a subset of samples with European descent was part of the SNP QC procedures.)
- Pairwise Pearson $r^2 \leq 0.1$, to exclude SNPs in high linkage disequilibrium. (The r^2 coefficient was computed using `plink` and its `indep-pairwise` function with a moving window of size 1000 bp).
- Removed C/G and A/T SNPs to avoid unresolvable strand mismatches.
- Excluded SNPs in several regions with long-range LD [79]. (The list includes the MHC and 22 other regions.)

We computed the genetic PC loadings specifically on the individuals self-reported as British, to capture population structure only on this subset. For this, the software `flashpca` was utilised, which is capable of performing fast PCA on SNP data [2].

Implementation

GWAS was performed using the `bgenie` tool (version 1.3). The genotype data for each chromosome are provided by the UK Biobank as a separate BGEN file. We split these files into 1703 files, one for each of the nearly linkage-disequilibrium-independent regions of about 2Mb defined in [15]. For GWAS, these regions, of similar size, are processed in parallel for all the phenotypes on the pool of nodes available on the ARC3 and ARC4 HPCs at the University of Leeds. Subsequently, results for each region (for all the phenotypes) are split by phenotype and then merged into a single file per phenotype.

Removal of related individuals. The command line tool `GreedyRelated` was employed to remove related individuals. This program aims to minimise the number of individuals that need to be removed due to relatedness.

5.4 Genetic findings

In this section, we focus on genetic findings obtained by testing in a GWAS the different handcrafted phenotypes discussed so far. For LV, these indices were LVEDV, LVESV, LVSV, LVEF, LV sphericity index at end-diastole (LVEDSph), LV myocardial mass (LVM) and LV mass-to-volume ratio (LVMVR = LVM / LVEDV). For RV, the phenotypes RVEDV, RVESV, RVSV

and RVEF were studied (remember that since only endocardial is present in the ground truth, the RV myocardial mass can't be determined).

5.4.1 Left ventricle: volumes and volume-derived quantities

The corresponding Manhattan plots are displayed in Figures 5.4, 5.6, 5.5 and 5.8. In these plots, the 22 autosomes (non-sexual chromosomes) are juxtaposed along the horizontal axis and each dot represents a SNP. The $-\log_{10}(p-\text{value})$ is shown on the vertical axis. In the following, we discuss the associations found for each of these phenotypes.

LVEDV and LVESV. For this phenotype, we discover 6 independent associations. The strongest association, at rs11153730 ($p = 2.0 \times 10^{-29}$), is linked to *PLN* with high confidence, based on prior knowledge. This gene plays a crucial role in cardiomyocyte calcium handling by acting as a primary regulator of the SERCA protein (sarco/endoplasmic reticulum Ca^{2+}-ATPase), which transports calcium from the cytosol into the SR1 [62]. Mutations in *PLN* have a well-established relationship with dilated cardiomyopathy (DCM) [34]. In [78], *PLN* was found to be associated with LVEDV and LVESV. However, [9] does not report this locus for the same phenotypes.

The locus on chromosome 2 (with lead SNP rs2042995) is widely known to be associated with *TTN*. This gene encodes the protein titin, which is responsible for assembling myocyte sarcomere, and determines the stretching, contraction and passive stiffness of the myocardium [45]. This gene has been reported by [9, 78, 67].

rs375034445 lies within the body of *BAG3*: this is a well-known cardiac gene coding for a cellular protein that is predominantly expressed in skeletal and cardiac muscle, which plays a role in myocyte homeostasis and in the development of heart failure [52]; also, it shows a stronger association with LVESV and LVEF, as found in previous studies [9, 78].

The locus near the *ATXN2* gene has previously been reported for LVEDV and LVSV. [78]. A candidate casual gene for this association is gene *MYL2*. The lead SNP (rs35350651) lies 558808 bp away from this gene's transcription start site (TSS). [91]

The gene *TMEM43* has been found in [78] in association with LVESV and LVEF.

Finally, gene *MYH6* harbors SNP rs365990. This gene provides instructions for making a protein known as the cardiac α-myosin heavy chain, which is expressed throughout the myocardium

during early cardiac development [3]. Mutations in this gene, as well as the neighboring *MYH7* responsible for the β-myosin heavy chain, have been linked to several pathologies: cardiomyopathies, arrhythmias and congenital heart disease (CHD).

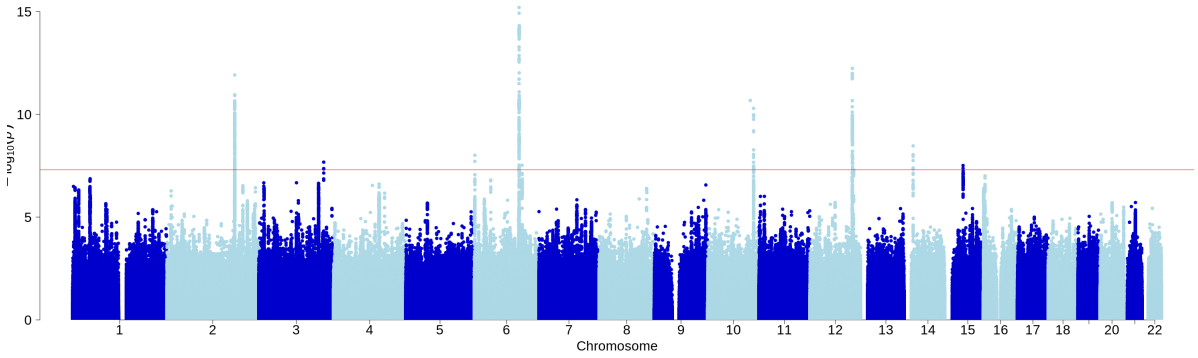

Figure 5.4: Manhattan plot of GWAS of left-ventricular end-diastolic volume (LVEDV)

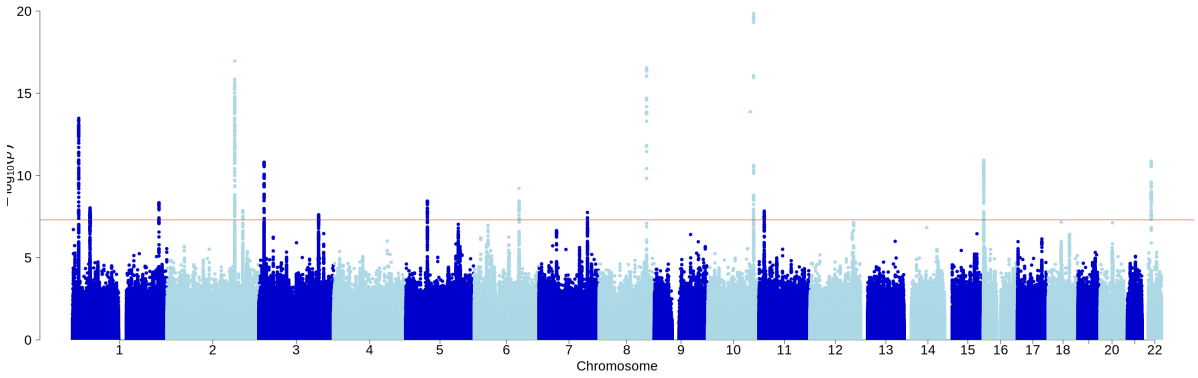

Figure 5.5: Manhattan plot of GWAS of left-ventricular end-systolic volume (LVESV).

LVEDSph. Note that LVEDSph has not been investigated in the context of GWAS before our work in [18] and the more recent work by Vukadinovic *et al.*[105], albeit a related phenotype, named "LV internal dimensions" was studied in an early GWAS of echocardiography-derived LV traits, [99].

For this phenotype, we find 9 additional independent associations, apart from the *PLN* locus. rs35564079 is located 8250 bp upstream of the TSS of *NKX2-5*, in chromosome 5. This gene plays a crucial role in heart development; in particular, in the formation of the heart tube, which is a structure that will eventually give rise to the heart and great vessels. *NKX2-5* helps determine the heart's position in the chest and also develops the heart valves and septa. Mutations in the *NKX2-5* gene have been associated with several types of congenital heart defects, including atrial septal defects and atrioventricular block [111]. It has not been reported

in [9] or [78], but shows borderline significance with the fractal dimension of the LV trabeculae [67]. rs72007904 is located 300kb upstream of the TSS of the gene *ABRA*. *ABRA* codes for a cardiac and skeletal muscle-specific actin-binding protein located in the Z disc and M-line and binds with actin. Consistent with this, it is differentially expressed in cardiac tissues and skeletal muscle in the GTEx data. *ABRA* has been associated with DCM in mice [58]. rs35001652 is close to *KDM1A*, a gene that codes for a histone demethylase involved in cardiac development, according to studies in mice [4]. rs463106 lies in the body of gene *PRDM6*. The mouse homologue of this gene, *Prdm6*, has been found to be important in early cardiac development [46]. An interesting association, with SNP rs162746, is close to gene *EN1*, however we were not able to find a strong candidate gene in this region; the causal gene for this association thus remains elusive. Finally, rs573709385 lies in a gene desert in chromosome 2, the closest protein-coding genes being *ACVR2A* and *ZEB2* (both at around 1.6Mb).

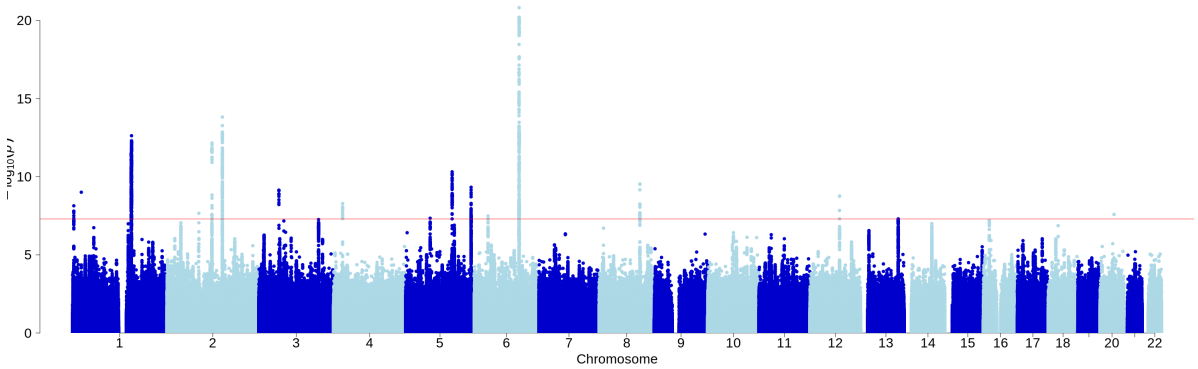

Figure 5.6: Manhattan plot of GWAS of left-ventricular end-diastolic sphericity (LVEDSph).

LVM and LVMVR. For LVM, 4 associatons are found: rs4767239 is probably related to developmental gene *TBX5* (T-box transcription factor 5), which has a known role in developing the heart and the limbs [94]. Through familial studies, mutations in this gene have been associated with Holt-Oram syndrome, a developmental disorder affecting the heart and upper limbs. In particular, there have been no recent reports on GWAS on LV phenotypes.

The locus near the *CENPW* gene has a cardiac gene, *HEY2*, possibly causal for this association. *HEY2* has been shown to suppress cardiac hypertrophy through an inhibitory interaction with *GATA4*, a transcription factor that plays a key role in cardiac development and hypertrophy. [109]. HEY proteins are direct targets of Notch signaling and have been shown to regulate multiple key steps in cardiovascular development. Studies have found that the loss of *HEY2*

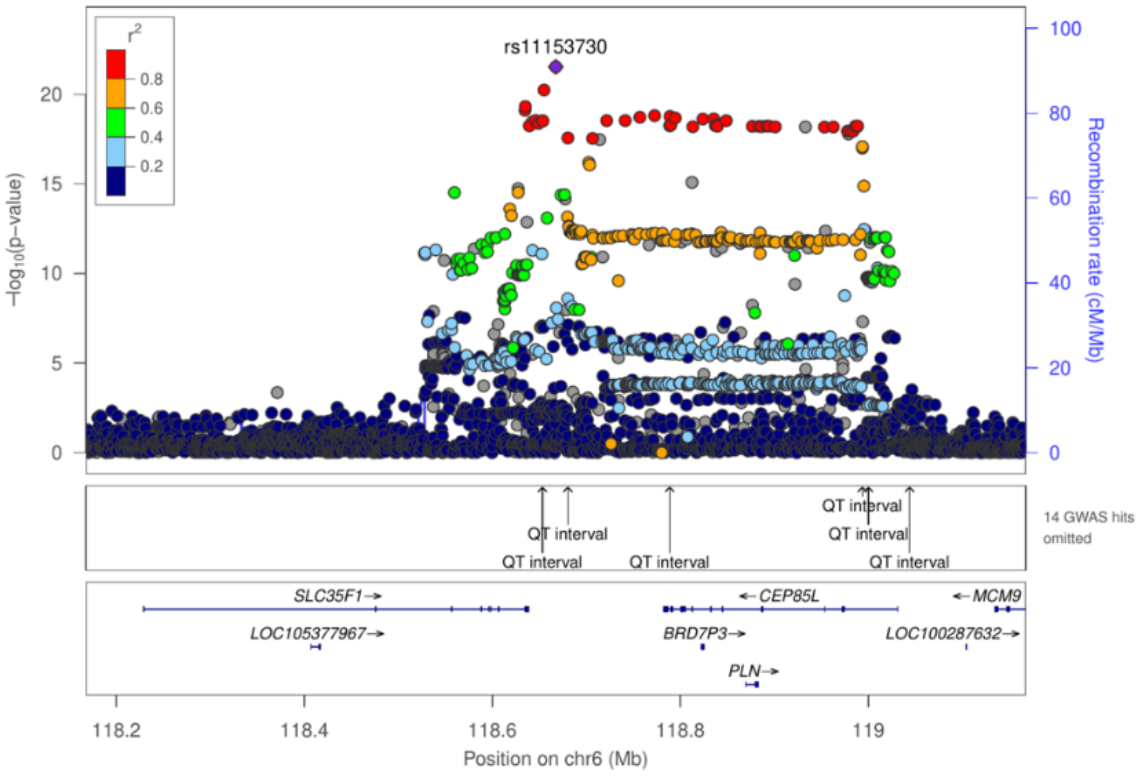

Figure 5.7: LocusZoom of *PLN* locus for LVEDSph.

in mice leads to cardiac defects with high postnatal lethality [36]. This locus has also been reported as associated to right-ventricular phenotypes [77].

rs3740293 overlaps gene *SYNPO2L*, which is highly expressed in cardiac tissues (LV and atrial appendage) and skeletal muscle, making it a strong candidate gene. This SNP is also close to gene *MYOZ1*, which is also supported by our GWAS study (see Section on TWAS). Both genes have been previously proposed as candidates for cardiac phenotypes, in particular atrial fibrillation [64, 70]. However, *MYOZ1* shows very high expression only in the latter. Loss-of-function variants in this *SYNPO2L* have also been found causative of atrial fibrillation [25], supporting this gene as a more likely candidate.

Finally, for LVMVR, 2 new loci were found, apart from the *PLN* locus: rs17460016 in the *FNDC3B* gene (in chromosome 3) and rs12542527 (in chromosome 8), an eQTL for the *MTSS1* gene also linked to LV fractal dimension [67].

LVSV and LVEF. We now discuss stroke volume and ejection fraction, quantities that, as discussed, involve volumes at two different phases, ED and ES. Whereas LVSV mostly re-

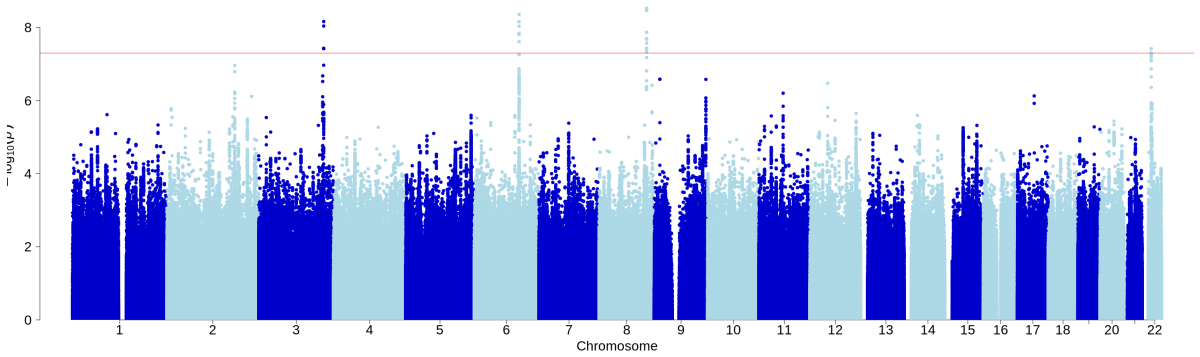

Figure 5.8: Manhattan plot of GWAS of left-ventricular mass-to-volume ratio (LVMVR).

discovers associations for LVEDV (*TTN*, *PLN*, *MYL2*, *MYH6*), LVEF shows some additional genome-wide significant associations, with loci near genes: *HSPB7*, *BAG3*, *LMF1*, *SMARCB1*, *MTSS1* and *CDKN1A*.

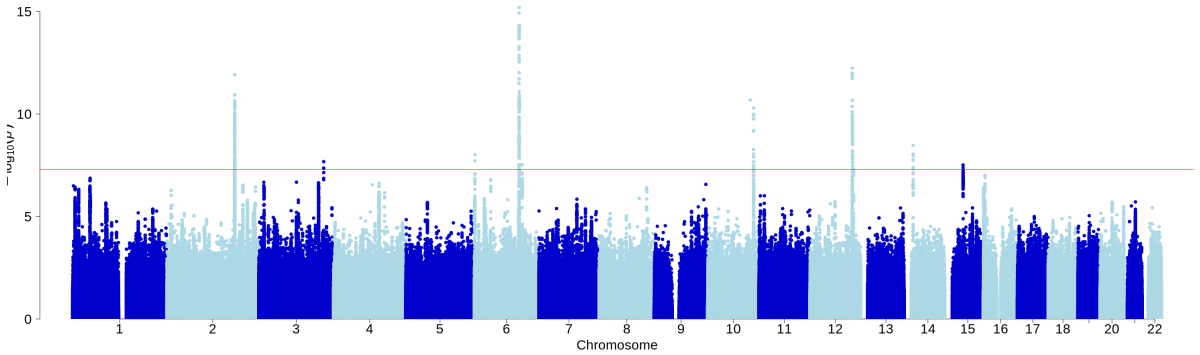

Figure 5.9: Manhattan plot of GWAS of stroke computed from left-ventricular volumes (LVSV).

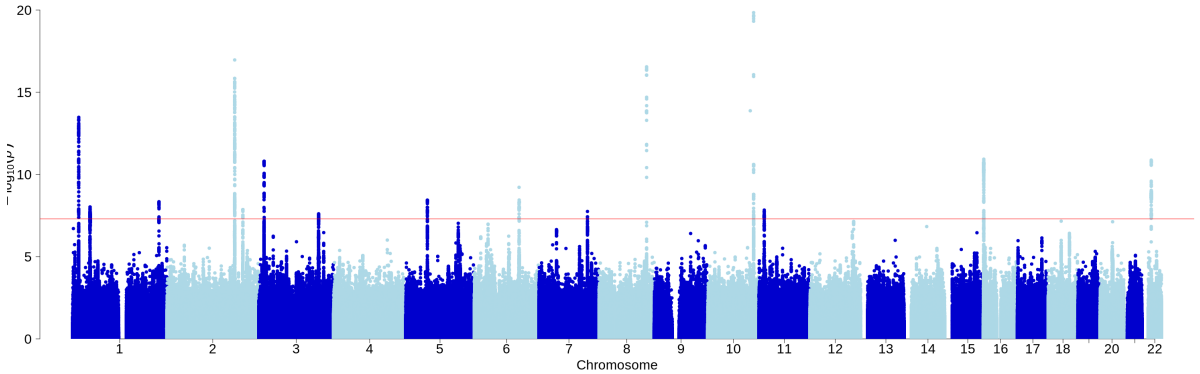

Figure 5.10: Manhattan plot of GWAS of left-ventricular ejection fraction (LVEF).

5.4.2 Right ventricle: volumes and volume-derived quantities

RVEDV and RVESV. Figures 5.11 and 5.12 show Manhattan plots for RVEDV and RVESV, respectively. For RVEDV, associations are found for loci near genes *TTN*, *SHOX2*, *RBM20*, *PLN*, *HEY2*, *FBN1* and *RUNX2*. The RVESV GWAS hit on chromosome 18 is near gene *FHOD3*, which has been identified as a causative gene for hypertrophic cardiomyopathy [72], making it a strong candidate gene for this association. In addition, GWAs hit in chromosome 9 is close to the *RGS3*, a potentially causal gene [114].

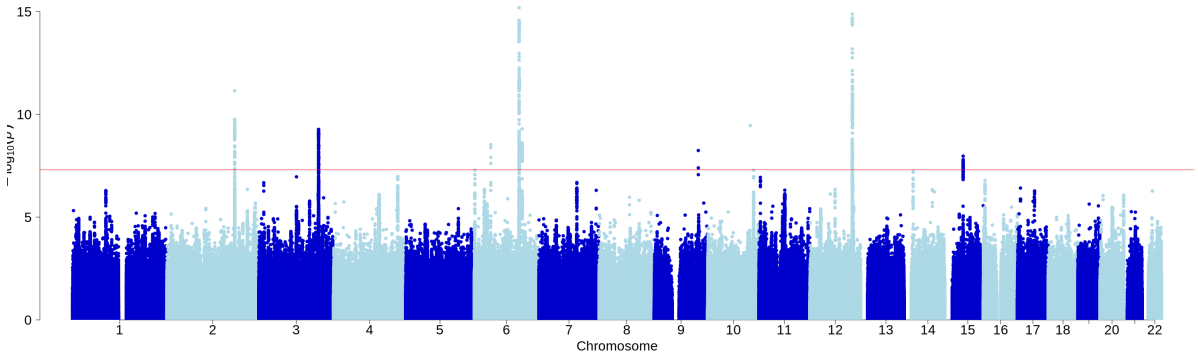

Figure 5.11: Manhattan plot of GWAS of right-ventricular end-diastolic volume ratio (RVEDV).

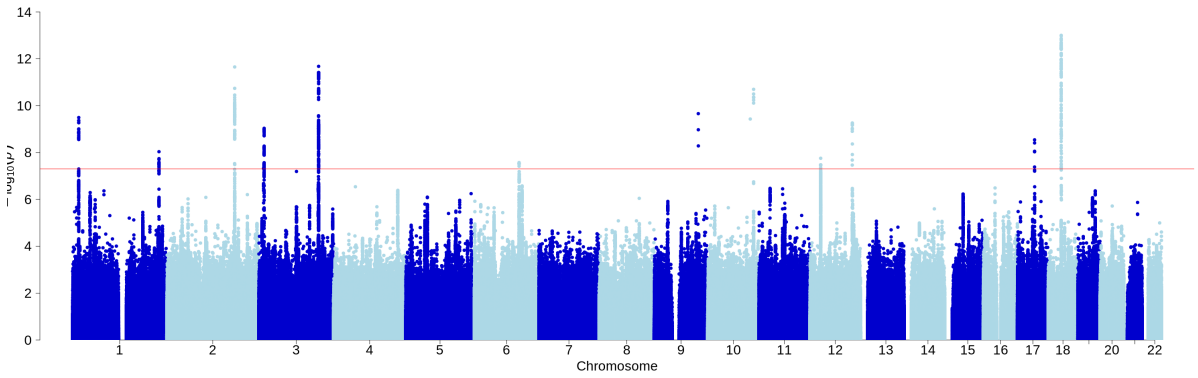

Figure 5.12: Manhattan plot of GWAS of right-ventricular end-systolic volume ratio (RVESV).

RVSV and RVEF. 5.13 and 5.14 depict the Manhattan plots for GWAS associations with RVSV and RVEF, respectively. Reassuringly, RVSV (expected to equal LVSV) identifies three of the associations from its counterpart calculated from LV volumes, albeit missing associations with *TTN* and *BAG3*.

GWAS on RVEF pinpoints a strong association with the *FHOD3* locus at chromosome, which only has borderline significance for LVEF. In addition, it identifies *HSPB7*.

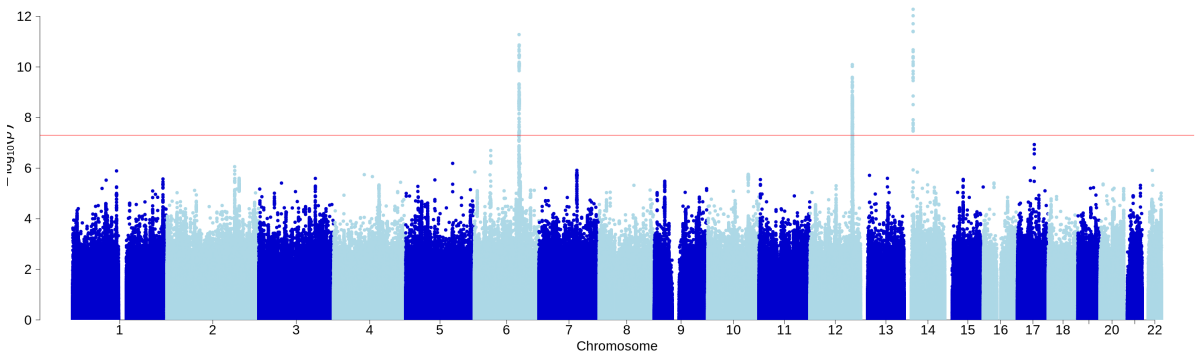

Figure 5.13: Manhattan plot of GWAS of stroke volume computed with right-ventricular volumes (RVSV).

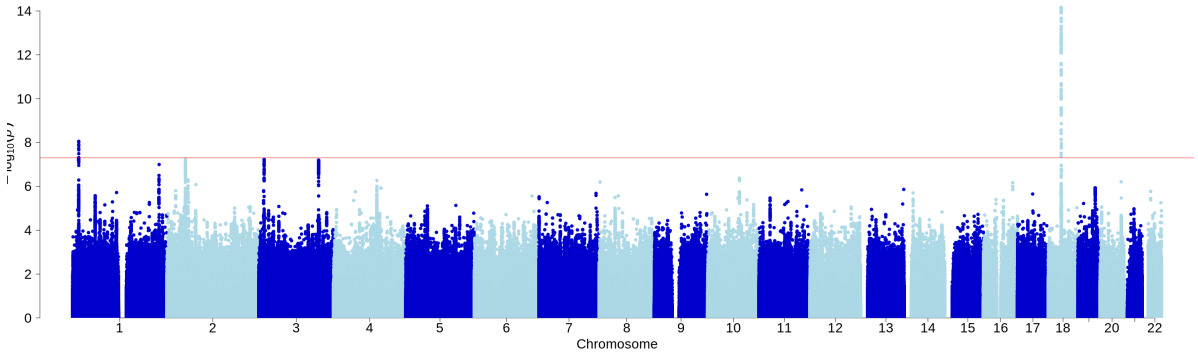

Figure 5.14: Manhattan plot of GWAS of right-ventricular ejection fraction (RVEF).

5.4.3 Left ventricle: local phenotypes

In this subsection, we investigate local LV phenotypes at each of the 17 AHA segments. In particular, we examine myocardial thickness (at ED and ES) and thickening. Absolute and relative local wall thicknesses are analyzed at the ED and ES phases, totalling 68 ($= 17 \times 2 \times 2$) traits. On the other hand, absolute and relative local thickening traits amount to 34 phenotypes. In total, they constitute 102 quantitative phenotypes to be tested. Given the high number of phenotypes, we choose not to display the corresponding Manhattan plots; instead, bull's eye plots are utilised to depict summary statistics for a selection of significant loci, across the LV location.

We declare associations study-wide significant when they surpass the Bonferroni threshold of 4.9×10^{-10} ($= 5 \times 10^{-8}/102$) in at least one phenotype. Additionally, we consider associations (suggestively) significant if they surpass the genome-wide significance threshold of 5×10^{-8} in at least 5 different phenotypes. In the following discussion, we make no distinction between

Phenotype	loci with $p < 5 \times 10^{-8}$
LVEDV	*TTN, BAG3, MYL2/ATXN2, MYH6, TMEM43, RRAS2*, ATG4D*, FBN1*
LVEDSph	*PLN, NKX2-5, EN1, KDMA1, PRDM6, ABRA, FDPS, ACVR2A/ZEB2, PPARGC1A, ADAMTS6*
LVM	*TTN, MYOZ2, TBX5, HEY2, FDPS, SYNOP2L*
LVMVR	*FNDC3B, PLN, MTSS1, SMARCB1*
LVESV	*HSPB7, TTN, RBM20, SMARCB1, BAG3, SHOX2, TMEM43, PLN, FLNC, MTSS1, SPATS2L, SPON1, LMF1, OBSCN*
LVSV	*TTN, PLN, MYL2, MYH6*
LVEF	*HSPB7, BAG3, MTSS1, CDKN1A, SMARCB1, LMF1*
RVEDV	*RBM20, MYL2, TTN, SHOX2, PLN, HEY2, FBN1, VEGFA*
RVESV	*HSPB7, OBSCN, RBM20, BAG3, MYL2, GOSR2, FHOD3, TTN , TMEM43, SHOX2, PLN, RGS3, CAND2*
RVSV	*MYL2, MYH6, PLN*
RVEF	*HSPB7, FHOD3*

Table 5.2: Summary of loci associated to handcrafted global LV and RV cardiac indices.

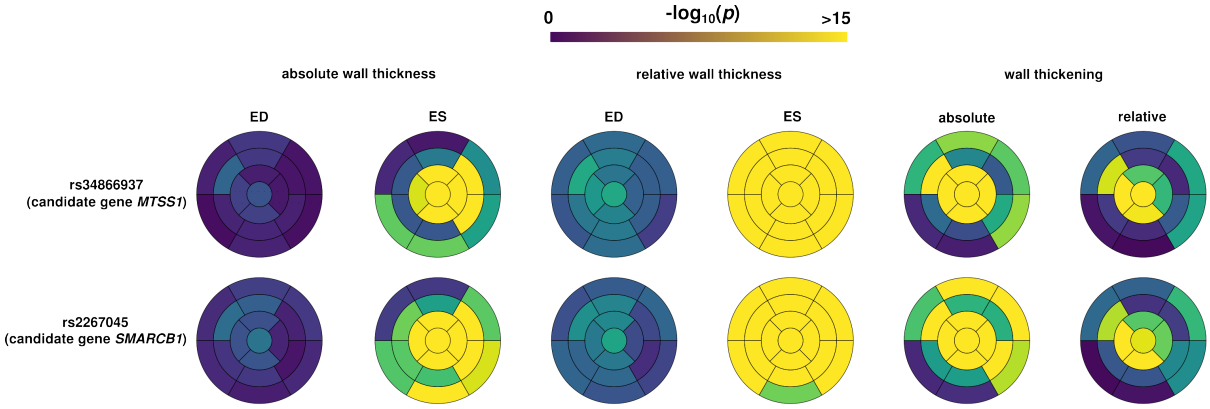

Figure 5.15: Bull's eye plots depicting $-\log_{10}(p)$ values for genetic associations with two representative loci: *MTSS1* and *SMARCB1*. Figure 5.1 can be consulted as a reference for the 17 AHA segments depicted here.

both types of associations to thus alleviate wordiness.

Absolute and relative wall thickness

As expectable, these associations reconfirmed those for LVMVR, since the latter is also a measure of relative LV wall thickness. We found the following loci (which are termed according to the most likely candidate genes): *MYO18B*, *CDKN1A*, *PRDM16*, *PLN*, *MTSS1*, *STRN*, *BAG3*, *FHOD3*, *LMF1*, *HSPB7*, *TMEM43*, *SMARCB1*, *GOSR2*, *RGS6* and *ADAMTSL3*.

5.5 Discussion

Taken together, the GWAS on the LV volume-derived phenotypes yielded 15 independent genome-wide associations. On the other hand, RV volume-derived produced 13 independent genome-wide associations, where 5 are distinct from LV ones. Interestingly, LVED sphericity index discovered a remarkable number of new loci.

Finally, examination of LV myocardial thickness and thickening could identify additional as-

sociations. Reassuringly, these associations mostly overlap with what has been reported in [7] where maximum wall thickness was studied. However, we highlight that use of a dimensionless wall thickness, which had not been tested in previous studies, was able to detect stronger associations than its absolute counterpart, especially when assessed at the end-systolic phase. In particular, associations with loci *MTSS1*, *SMARCB1* and *BAG3* were remarkably stronger for the dimensionless phenotype (see Figure 5.15.)

We observe that some loci are associated to chamber phenotypes distinctly at either distinctly their contracted or relaxed states.

5.6 Conclusions

In this chapter, we have leveraged our CMR-derived cardiac 3D meshes to derive several handcrafted indices and perform GWAS on them to discover possibly new genetic factors. We investigate global parameters, like volumes, volume-derived quantities and LV sphericity index, as well as wall thickness and thickening values computed locally at different LV locations.

These findings highlight the power of traditional phenotyping approaches, while also motivating the need for more flexible and expressive methods. In the next chapter, we explore how unsupervised, data-driven, methodologies can uncover latent phenotypes that carry complementary genetic signal.

5.7 Code availability

The repository `CardioMesh`, previously cited, contains the code to carry out the different mesh-based quantifications necessary to reproduce the results from this chapter. This repository is publicly available at `www.github.com/rbonazzola/CardioMesh`.

Chapter 6

Unsupervised static CMR-derived phenotypes

In this chapter, we investigate the genetic factors underlying shape features derived from left-ventricular meshes at end-diastole (LVED), this time in an unsupervised manner. Our hypothesis is that the recovery of more subtle shape features will enable the discovery of additional genetic variation driving LV morphology.

This chapter is composed of the following sections:

- **Section 6.1** provides a motivation for the need of unsupervised phenotyping approaches in imaging genetics in general, and in cardiac imaging genetics in particular.
- In **Section 6.2** we explore a traditional classical analysis method, shape PCA, applying it to meshes of LV at end-diastole (LVED) and describing the genetic loci that arise when testing shape PCs in GWAS.
- **Section 6.3** explores the usage of a state-of-the-art deep-learning technique, convolutional mesh autoencoders (CoMA), for shape representation learning.
- **Section 6.4** proposes an ensemble methodology that builds a highly expressive representation by pooling together representations from CoMA networks with different hyperparameters. We demonstrate that this ensemble-based approach boosts the discoverability of genetic loci, both with respect to shape-PCA and to a single-run approach. Discovered loci are discussed at length in view of recent literature.

- **Section 6.5** provides evidence supporting the identification of the different discovered loci with functional elements, i.e. genes.
- **Section 6.6** Explores the possibility of improving a learned representation for association with genetic variants via fine-tuning (this piece of work corresponds to the conference communication [18] which precedes chronologically the rest of the work in this chapter).
- **Section 6.7** summarises the contributions and findings of this chapter, and highlights some limitations.
- **Section 6.8** provides an overview of the codebase used for this chapter, which is publicly available on GitHub.

The work from this chapter has been published on and adapted from a paper in the journal *Nature Machine Intelligence* [17].

6.1 Motivation

The overarching theme of this thesis is finding patterns of variation in biomedical images, in particular CMR, which have not been characterised by means of handcrafted phenotypes, with the purpose of improving genetic discovery with respect to those more traditional measures (which we have already studied in the preceding chapter).

6.2 Methods

In this chapter, we leverage the dimensionality reduction methods exposed in Chapter 4 for producing data-driven phenotypes from static 3D LV meshes.

6.2.1 Overview

We first provide an overview of our methodology. The proposed method is outlined in Figure 6.3. It starts with extracting 3D meshes representations of LVED from CMR images using an automatic segmentation method, as explained in detail in Chapter 3. We then train several models with different metaparameters (network architecture, random seeds controlling weight initialisation and dataset partitioning, and relative weight of the variational loss) to learn low-dimensional representations of the 3D meshes which capture anatomical variations using an encoder-decoder model. All meshes are then projected to this latent space to derive a few shape

descriptors (or latent variables) for each mesh. To take advantage of the variability induced in the representation obtained by the meta-parameters, we pooled the latent vectors to obtain a richer representation. The features that make up this pooled representation are finally used in GWAS to discover genetic variants associated with shape patterns.

6.3 Shape PCA on cardiac meshes

Firstly, we explore the use of a common statistical shape analysis method, shape PCA, in the context of genetic association studies. Strikingly, to our knowledge this methodology has not been explored for heart phenotyping in prior literature, in the context of genetic association studies.

A shape PCA model (see Methods section) was fit to our set of LVED meshes and novel and interesting results are obtained, as described in the remainder of this section.

6.3.1 Morphological interpretation

The effect on LVED shape for the first 8 modes is shown in Figure 6.1. We discuss these effects, for the first four PCs, PC1 through PC4:

- PC1 shows mainly an association with LVEDV
- PC2 seems to be linked to LV conicity without changes in size.
- PC3 is highly correlated with LVEDSph.
- PC4 models a change in mitral orientation, without any change in size or sphericity.

The rest of the PCs are not described as their effect is not clear by inspection of these shapes.

6.3.2 GWAS results on LV shape PCs

GWAS was performed for the first 16 modes and 18 independent loci were found with study-wide significance ($p < 3.1 \times 10^{-9}$). It is worth mentioning that the GWAS is performed on all the individuals, including those on which the shape PCA model was trained. This is reasonable because the algorithm does not optimise for association with genetic variants, and therefore a uniform distribution of p-values under the null distribution can be safely assumed, even when including these subjects in the sample. In the following, we discuss the genetic associations

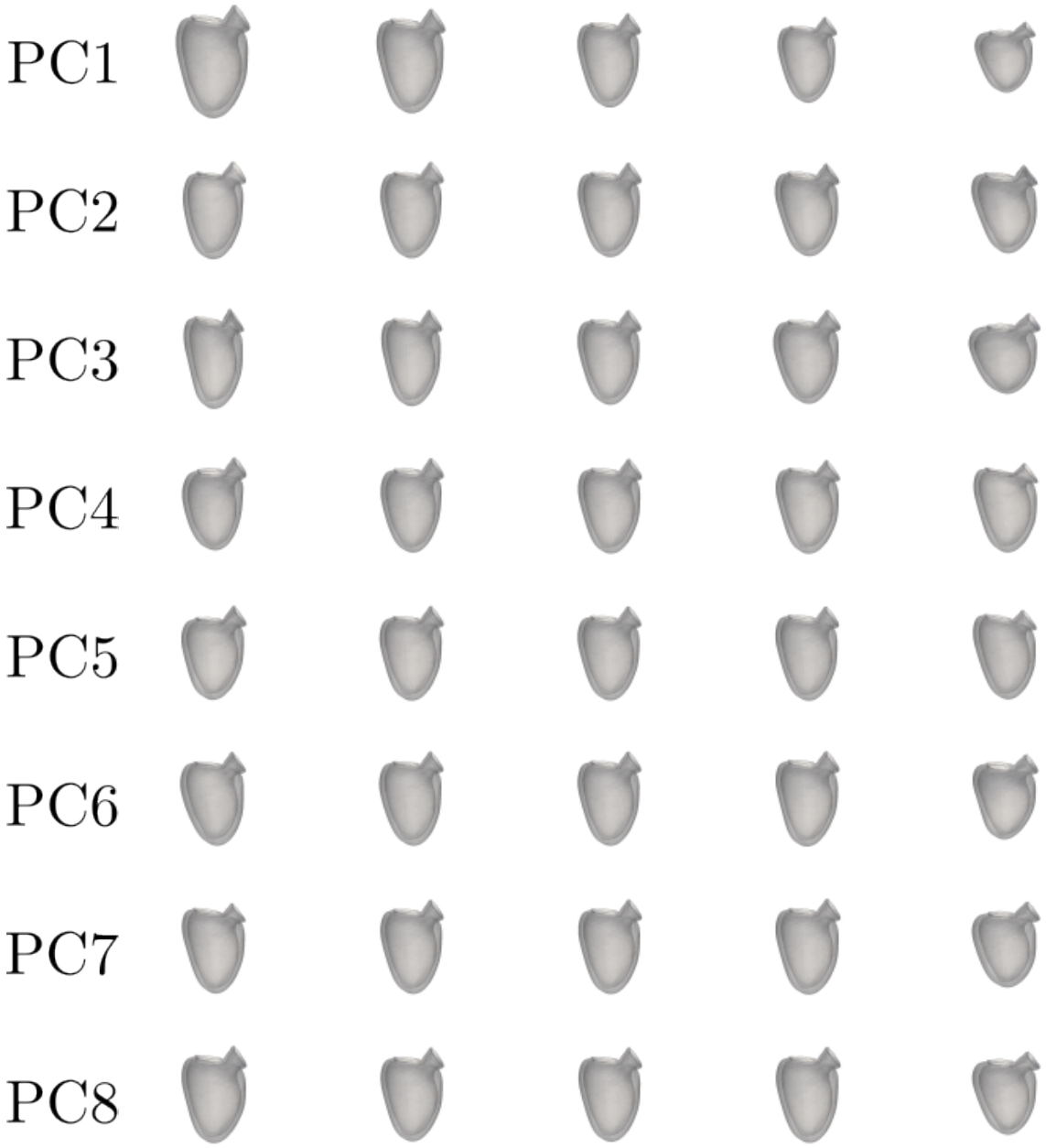

Figure 6.1: Effect on left-ventricular morphology of the first 8 shape PCs. Each shape corresponds to the average of the following quantiles: [0, 0.01], [0.095, 0.105], [0.495, 0.505], [0.895, 0.905] and [0.99, 1]. Since the population encompasses more than 60,000, each shape is the average of around 600 meshes. As explained in the text, the shapes are unscaled before averaging, and then scaled back after averaging.

found for each shape mode:

- **PC1** reconfirms the LVEDV associations with *TTN*, *MYL2* and *MYH6* discussed in Chapter 5. A new association, in chromosome 4, is an indel (chr4:120304290_GC_G) located 200kb downstream of *MYOZ2*. This gene codes for protein that functions by tethering calcineurin to alpha-actinin at Z-discs in muscle cells and inhibits the pathological cardiac hypertrophic response [86]. Another candidate gene in this locus is *PDE5A*. Indeed, some of the strongest associations overlap the body of this gene (although not the lead variant which is the indel mentioned above). It has been shown that *PDE5A* is expressed in cardiac myocytes and may have pro-hypertrophic effects [113].

- **PC2** is strongly linked with a new locus in chromosome 17, *GOSR2*. Interestingly, [67] reports the *GOSR2* locus as significantly associated with trabecular fractal dimension in slices 3 and 4, however previous GWAS in global LV indices have not reported this locus. More broadly in the literature on genetics of cardiovascular phenotypes, it has been reported as associated to ascending aorta distensibility [76], mitral valve geometry [112] and congenital heart disease [56].

- **PC3** re-discovers the *PLN* and *NKX2-5* loci, already present for LVEDSph. It also adds an association in chromosome 1, the SNP rs12142143, which lies within the *ACTN2* gene. This gene codes for the Z disk protein α-actinin-2. This locus has been reported for LVSV in [78].

- **PC6** has hits in the *TBX5* and *NKX2-5* loci, with a new association near the *NAV3* gene, that has been found to play a role in heart development in zebrafish [61].

- **PC7** is associated to a SNP near the transcription start site (TSS) of *PITX2* gene. It encodes for a transcription factor required for mammalian development, and disruption in its expression in humans causes congenital heart diseases and is associated with atrial fibrillation.

- **PC10** is linked to the *PRDM6* locus (discussed before in connection with LVEDSph).

- **PC11** is associated to SNPs rs59894072 (close to *TBX3*, a known cardiac gene [13]) and rs56229089. The second, in turn, is close (1Mb) to two possible candidate genes: *KCNJ2*, a potassium channel gene which is active in skeletal muscles and cardiac muscles [84]; and *SOX9*, a gene implicated in cardiac development [31].

- Finally, **PC16** has shown two associations, with the previously discussed *MTSS1* and *MYH6* loci.

6.4 Convolutional mesh autoencoders

In this Section, we turn to non-linear representations of the LV shapes, by means of convolutional mesh autoencoders (CoMA), a type of autoencoder that has been designed specifically for 3D mesh processing and which has been described in **Section 4.4**.

6.4.1 Why a non-linear representation for a PCA-generated population?

An anticipated concern is whether a non-linear representation is actually needed, considering the fact that the MCSI-Net segmentation algorithm (described in subsection 3.4) produces meshes by means of a PCA-based PDM which is inherently linear. Here, we argue that a non-linear representation can still be useful.

Firstly, recall that the PDM has been trained with unscaled meshes, and the scale information is re-introduced as an extra transformation; instead, our mesh-derived latent embeddings do account for scale information. This could potentionally introduce non-linear interactions between the scale and the rest of the shape features, thus enriching the representation. On the other hand, the PDM contains 70 principal components, whereas our reduced representations are much smaller ($n_z = 8$ or $n_z = 16$), which could cause the network to learn non-linear combinations of the input coordinates in order to account for as much shape variation as possible, leading to a more compact representation than the linear one.

6.4.2 Phenotype ensembling procedure

Given that the evaluation metric which guides the training, i.e. reconstruction error with a variational loss, is not necessarily aligned with the final purpose of discovering genes that influence LV shape, there is no reason to adopt the single run with the best value for such metric. This was the approach followed in our conference paper [18], discussed in Section 6.6 and which precedes chronologically the results in this section.

For this reason, here we choose to adopt an ensemble-based approach, in which we pool the different phenotypes together. Based on the observation that different network hyperparameters, dataset partitioning and weight initialisations yielded latent encodings with different genome-

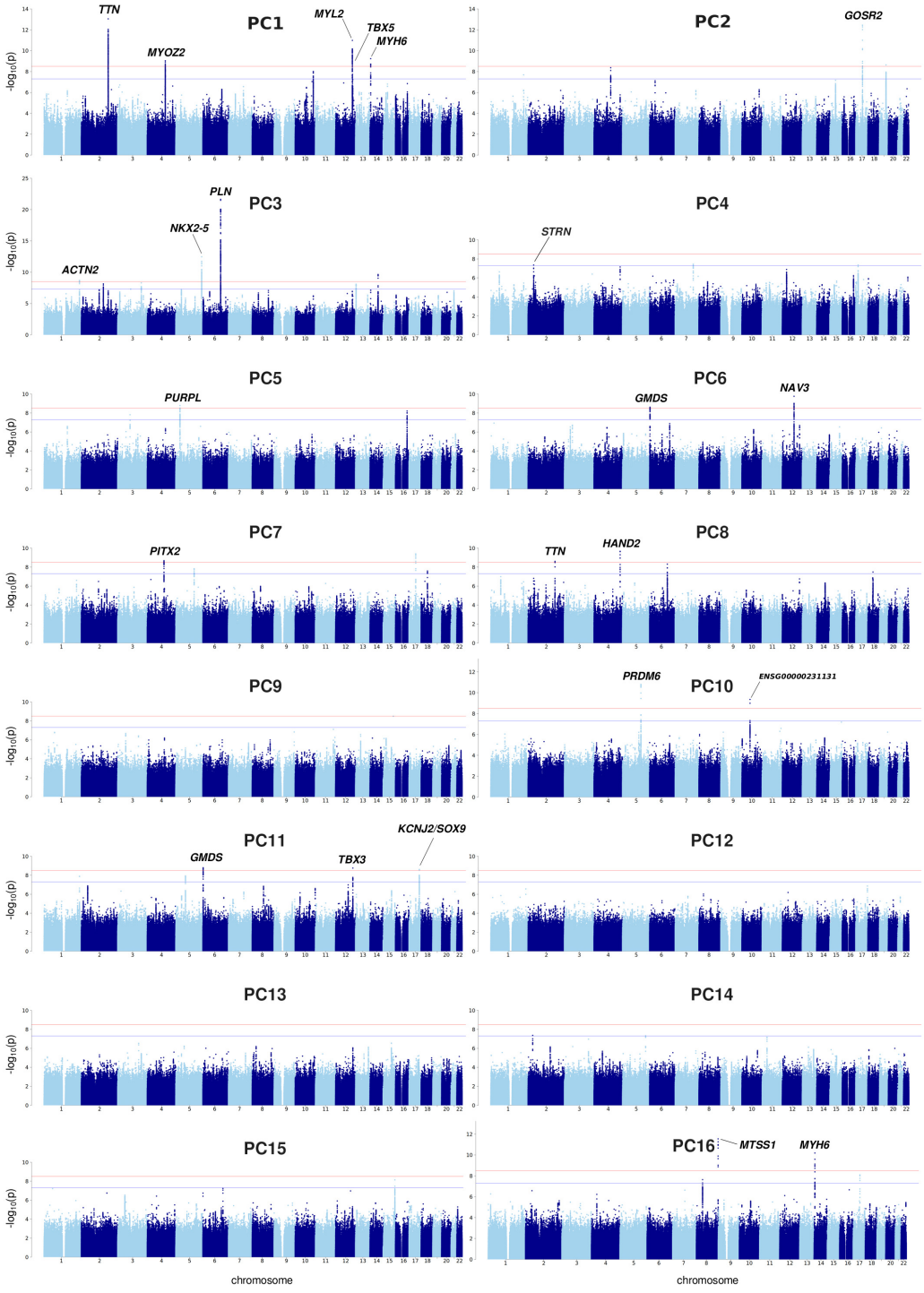

Figure 6.2: GWAS on first 16 shape PCs on left-ventricular meshes.

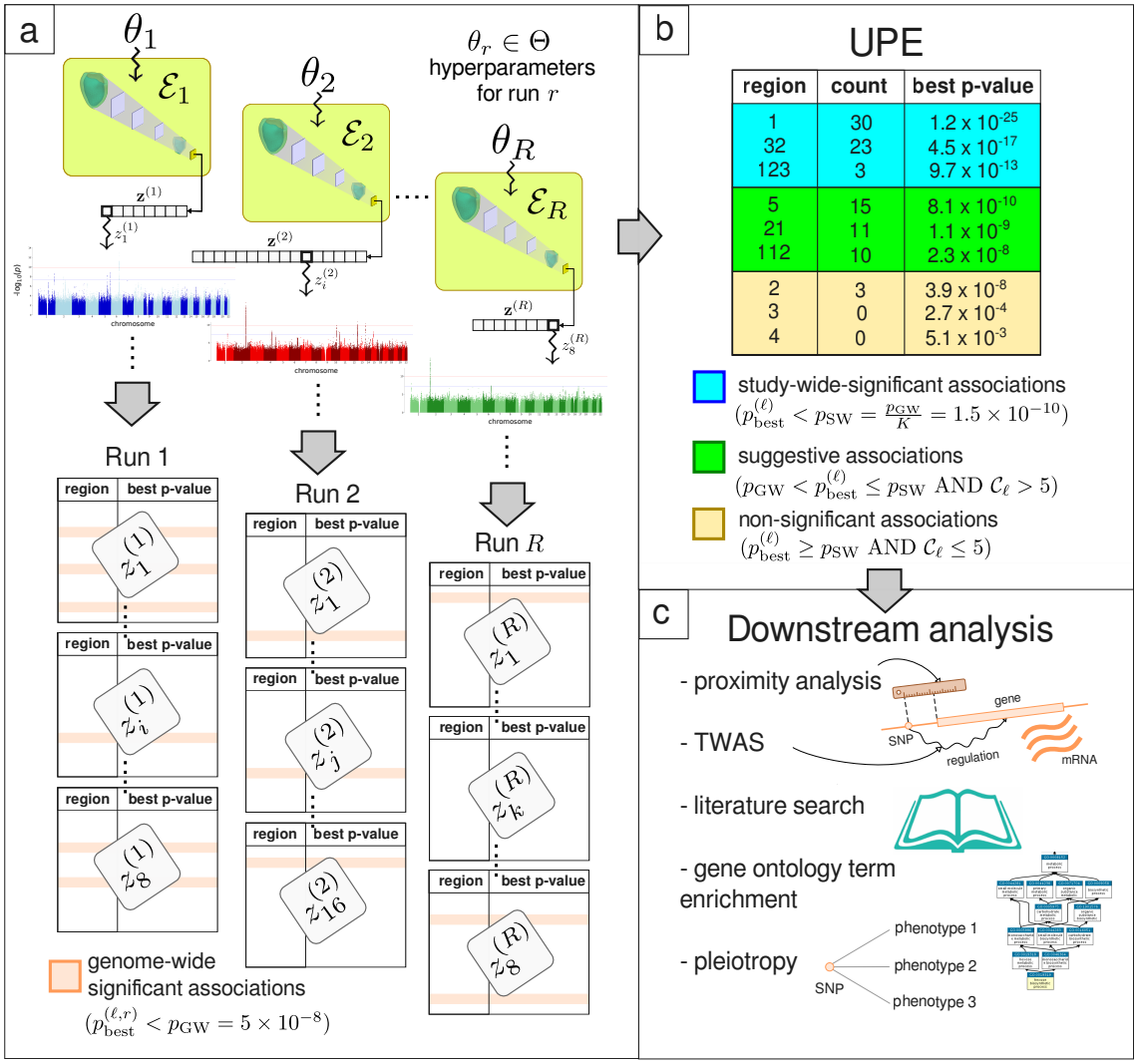

Figure 6.3: Flowchart of the proposed Unsupervised Phenotype Ensembles (UPE) framework. a) The different latent vectors obtained for each CoMA run, $\{\mathbf{z}_r\}_{r=1}^{R}$ are tested in a GWAS component by component, for association with genetic variants. The best association, with p-value $p_{\text{best}}^{(\ell,r)}$, is found for each run r and region ℓ (each region is of around 2Mb in length). c) Associations are then aggregated across the ensemble and classified as significant, suggestive and non-significant according to $p_{\text{best}}^{(\ell)}$ and the count $\mathcal{C}_\ell$ (as indicated by the rules). c) Finally, for each significant association, a downstream analysis is conducted to identify potentially causative genes.

wide-significant loci, we built an ensemble of phenotypes by concatenating the latent vectors for each individual run. This composite representation provides a redundant yet more expressive representation of LV shape at end-diastole. We coin this methodology *Unsupervised Phenotype Ensembles* (UPE).

These pooled runs covered a wide range of w_{KL}, and variations in network architectures; most importantly, in the latent space dimensionality, n_z. Also, for a given combination of hyperpa-

	Input	**Output**
ChebConv	5220×3	$5220 \times C_1$
DS	$5220 \times C_1$	$2610 \times C_1$
ChebConv	$2610 \times C_1$	$2610 \times C_2$
DS	$2610 \times C_2$	$1305 \times C_2$
ChebConv	$1305 \times C_2$	$1305 \times C_3$
DS	$1305 \times C_3$	$652 \times C_3$
ChebConv	$652 \times C_3$	$652 \times C_4$
DS	$652 \times C_4$	$326 \times C_4$
ChebConv	$326 \times C_4$	$326 \times C_5$
FC	$326 \times C_5$	$n_z \times 1$

Table 6.1: Architecture of the encoder part used for each of the cardiac chambers. The decoder has the same architecture but reading from the bottom upwards and inverting input and output. (ChebConv: Chebyshev convolution, DS: downsampling, FC: fully connected layer.) The architectural hyperparameters were $(C_1, C_2, C_3, C_4, C_5) \in \{(16, 32, 64, 128), (128, 128, 128, 128), (1024, 512, 256, 128)\}$ and $n_z \in \{8, 16\}$.

rameters (including architecture), an optimal learning rate was selected empirically based on validation set reconstruction error, and then five different random seeds were utilised to initialize the network's weights and to partition the full dataset into training, validation and test partitions (each seed constitutes a different run). Details on the architectural parameters are given in Table 6.1.

Run selection. From the full set of runs, we selected 36 training runs that achieved a good reconstruction performance: a root mean squared deviation (RMSD) of less than 1 mm (averaged over the subjects from the test set) was chosen as the selection criterion. The deviation is taken to be the vertex-wise Euclidean distance, and the mean is taken over the $M = 5220$ vertices of the LV mesh. In other words, the RMSD for subject i in run r is:

$$\text{RMSD}_i^{(r)} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} ||\mathbf{x}_{i,j} - \hat{\mathbf{x}}_{i,j}^{(r)}||_2^2}, \tag{6.1}$$

where $\mathbf{x}_{i,j}$ denotes the triad of spatial coordinates for vertex j in the mesh of subject i, and $\hat{\mathbf{x}}_{i,j}^{(r)}$ is the same for the reconstructed mesh in run r of the autoencoder. $|| \cdot ||_2^2$ denotes the squared Euclidean norm. Note that the in-plane resolution in this CMR dataset is 1.8 mm, i.e. almost twice the selected threshold.

Importantly, the runs were selected based *only* on mesh reconstruction error and not on the presence or absence of GWAS hits. This allows to assume a uniform distribution of p-values

over the $[0, 1]$ interval under the null hypothesis of no effect.

These 36 autoencoder runs produced a total of 384 phenotypes (where the latent dimension was 8 for some runs and 16 for others). With the aim of controlling for the false discovery rate, this procedure requires correcting the usual genome-wide Bonferroni p-value threshold, $p_{\text{GW}} = 5 \times 10^{-8}$, since the number of statistical tests that are performed grows with the size of the (pooled) representation. With the aim of not overcorrecting this threshold, whenever a pair of latent variables (within the same run or not) had a Spearman correlation coefficient greater than 0.95 in absolute value, one of them was dropped at random. This procedure resulted in 324 phenotypes to be tested in GWAS. The new study-wide threshold utilised was, therefore, $p_{\text{SW}} = \frac{p_{\text{GW}}}{324} = 1.5 \times 10^{-10}$.

Locus counting

To perform locus counting, a set of nearly LD-independent regions of around 2Mb was utilised to partition the genome [15].

Given that, for each genomic locus, the lead variant might vary across different phenotypes by virtue of high linkage disequilibrium with close genetic variants, we adopt the following approach for locus counting: the genome is partitioned into 1703 approximately LD-independent regions, where each is region is nearly 2 megabases (Mb) in length. We compute the number of autoencoder runs in which each region ℓ was genome-wide significant, denoting this quantity $\mathcal{C}_\ell$: for each run r and region ℓ, we retrieve the minimum p-value, across the different latent variables $z_k^{(r)}$ (remember that $1 \leq k \leq 8$ or $1 \leq k \leq 16$, depending on the run r) which we call $p_{\ell,r}$. We then count the number of runs for which $p_{\ell,r} < p_{\text{GW}}$: $\mathcal{C}_\ell = \sum_{r=1}^{R} \mathbf{1}_{p_{\ell,r} < p_{\text{GW}}}$, where $\mathbf{1}$ denotes the indicator function and $R = 36$. This $\mathcal{C}_\ell$ is the quantity labelled 'count' in Table 6.2. Also, note that $\mathbf{z}_r \equiv \mathbf{z}_{\theta_r}$ from Figure 6.3, where subindex θ_r was utilised to stress the fact that different hyperparameters were used in each run.

6.4.3 Implementation

Deep learning pipeline. The autoencoder architecture is detailed in Table 6.1. Different dimensions of the latent space n_z and weights w_{KL} were studied. With the aim of achieving a compromise between reconstruction error and interpretability of the components, we only study runs with $n_z \in \{8, 16\}$.

After each convolutional layer, a ReLU activation function was applied. The number of samples used for training was 5,000, whereas the validation set contained 1,000 individuals. The sample partition, as well as the weight initialisation (and the sampling process in the case of VAEs), were controlled by a random seed that was stored along with the trained model for reproducibility; 3 to 5 different random seeds were utilised for each parameter configuration. The Adam optimiser was used to find optimal network parameters, by minimising the KL-regularised MSE reconstruction loss [49]. The learning rates that achieved good performance utilised were in the range $[10^{-4}, 3 \times 10^{-4}]$.

The network training was performed on a Nvidia DGX A100 workstation located at the University of Leeds. This machine is endowed with Nvidia A100 GPUs.

Genetic pipeline. The genetic pipeline was executed on the University of Leeds' high performance computing cluster, ARC. The workload has been parallelised across computing nodes using the Son of the Grid Engine (SGE) queue management system. `qctool` was used to filter the original genotype files. `bgenie` version 1.4.1 was used to execute GWAS. A `bash` script was used to format the files to include the following columns: chromosome, variant ID, position, allele frequency, reference and alternative alleles, imputation INFO score, estimated effect size, standard error for the effect size, t-score, and p-value. Also, the rows were ordered by chromosome and position. Finally, the files were `tabix`-indexed so that position-based queries can be performed efficiently.

Each of the 324 GWAS summary statistics (one per phenotype), requiring around 1GB of disk space, was partitioned into the 1703 LD-independent blocks, and the smallest p-value for each block was identified for each phenotype and saved in corresponding files, to enable for an efficient and fast subsequent locus count and analysis.

The `R` was used Package was used to `qqman` was used to produce the Manhattan plots and the Q-Q plots for the GWAS. To establish candidate genes for the different loci, we first conducted a proximity analysis. For this, the `biomaRt` R package was utilised. Firstly, the locations of all the human genes were queried, using the Ensembl version 27 (which in turn uses the Human Genome Reference GRCh37). This includes not only protein-coding genes, but also pseudo-genes, long non-coding RNA and short non-coding RNA.

We filtered these genes according to whether their transcription start sites lie within a 500kb win-

dow centered at the lead SNP of each genomic region. To determine the transcription start site (TSS), the `Biomart` database provides three attributes: `"start_position"`, `"end_position"` and the strand (1 or -1). TSS corresponds to the attribute `"start_position"` when the strand is positive, and to `"end_position"` when it is negative.

We use the tool g:Profiler via its R API to find pathways for which our sets of genes were enriched. To define the gene sets, we selected a region of 200 kb around each lead variant and chose the genes whose transcription start site (TSS) was located in that region. Gene ontology terms belong to one of three different categories: molecular functions (MF), cellular components (CC), and biological processes (BP).

6.5 Results for unsupervised phenotype ensembles (UPE)

6.5.1 Reconstruction performance

Figure 6.4 shows a comparison of the reconstruction error obtained through principal component analysis (PCA) and convolutional mesh autoencoders (CoMA), as a function of the number of the components n_z of the latent space (values of 8 and 16 were used). CoMA and PCA yielded comparable reconstruction errors, with the best CoMA runs outperforming PCA slightly for $n_z = 8$, and PCA outperforming CoMA for $n_z = 16$. No significant difference was found between non-variational and variational CoMA in terms of the reconstruction error, in spite of the variational loss term which in principle could negatively affect the competing reconstruction term [1].

6.5.2 GWAS

GWAS was performed on all latent variables, for all training runs achieving a good reconstruction performance. The number of such runs was $R = 36$. The results obtained with $n_z = 8$ and $n_z = 16$ (eight and sixteen latent variables, respectively) are reported, with a total number of 384 latent variables in the pooled representation.

Loci counting across the ensemble. First, we examine the prevalence of significant GWAS loci found in all runs of our ensemble. To count the loci, we split the genome into approximately

[1] This fact also suggests the presence of multiple local optima in the loss function, further justifying the usefulness of an ensemble.

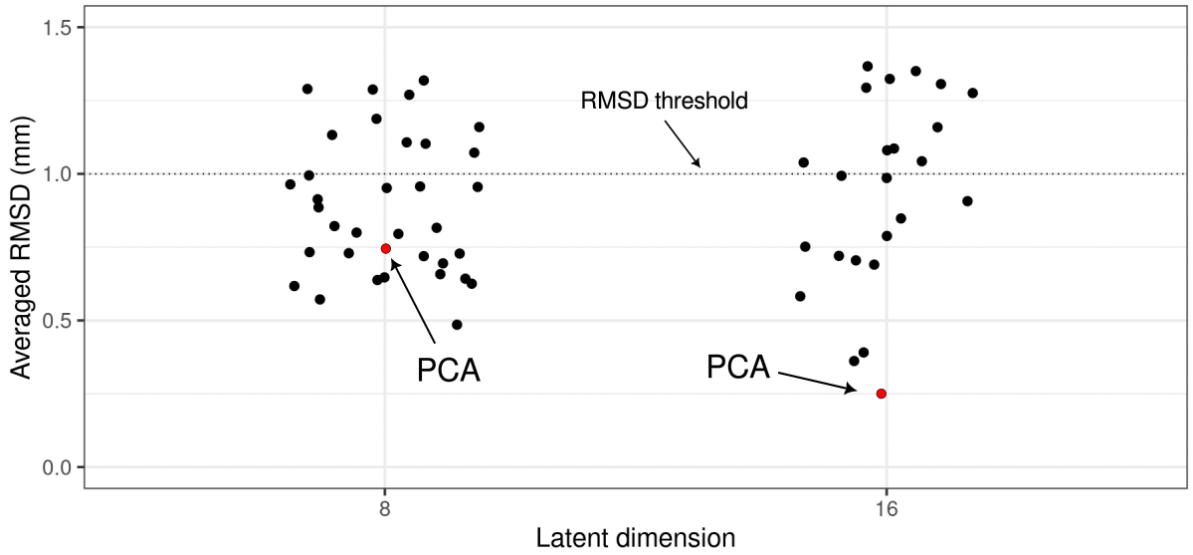

Figure 6.4: Averaged reconstruction errors (measured as the RMSD averaged across the test set of 1000 subjects) for different CoMA models (black dots) and for PCA (red dots). A threshold of 1 mm in this metric is used to select the runs.

LD-independent genomic regions [15] and computed the number of loci below the usual genome-wide significance threshold of 5×10^{-8} (see details in the Methods section); Table 1 shows the results.

We found 49 independent associations with study-wide significance, i.e. $p < p_{\mathrm{SW}}$. Importantly, all of the previously discussed findings are recovered by UPE with study-wide significance, except the following: *MTSS1*, *TBX3*, *PPARGC1A* and *FNDC3B* (the last two appear with suggestive significance in UPE). The summary statistics of the GWAS for the best latent variable of each of these 49 loci are displayed in Table 6.2. When a gene name is displayed in bold letters in this table, it means that this locus was found *only* via the ensemble approach (neither via traditional cardiac indices or shape PCA). Most loci have prior evidence supporting their plausible role in cardiac pathways. In addition, many of them are totally novel and represent interesting avenues for further research. In what follows, we perform an in-depth analysis of our novel genetic findings in the light of recent literature.

Loci with prior evidence. We now describe loci not previously associated with left-ventricular phenotypes, but supported by other types of evidence linking them to cardiac pathways.

rs11706187 is likely linked to developmental gene *SHOX2*. The mouse homologue of *SHOX2*, *Shox2*, is essential to differentiate cardiac pacemaker cells by repressing *Nkx2-5* [35]. Whereas both *TBX5* and *NKX2-5* are highly expressed in adult cardiac tissues according to GTEx

chr.	region	candidate gene	count	min. p-value	lead variant	NEA / EA	EAF (%)	$\lvert\hat{\beta}\rvert \pm \mathrm{se}(\hat{\beta})(\times 10^{-2})$
10	120591353-122407323	BAG3	35	4.1×10^{-18}	rs375034445	A / AT	21.2	5.29 ± 0.79
2	178553183-181312739	TTN	35	1.4×10^{-17}	rs2042995	T / C	23.2	5.70 ± 0.77
6	117672972-118963115	PLN	35	2.0×10^{-29}	rs11153730	T / C	48.6	4.29 ± 0.64
14	23018665-24905123	MYH6	34	2.7×10^{-14}	rs365990	A / G	36.9	4.04 ± 0.66
12	113986709-115036602	TBX5	34	2.3×10^{-11}	rs4767239	G / C	82.5	4.12 ± 0.85
12	110336719-113263518	MYL2	34	2.8×10^{-15}	rs35350651	A / AC	51.4	4.36 ± 0.64
4	119933512-120392684	MYOZ2	33	2.4×10^{-13}	4:120304290_GC_G	GC / G	29.0	3.96 ± 0.71
10	73508512-75422550	SYNPO2L	31	2.5×10^{-15}	rs3740293	A / C	14.3	5.07 ± 0.92
3	157312028-159477890	**SHOX2**	30	7.0×10^{-15}	rs11706187	A / G	50.1	3.23 ± 0.64
1	154770403-156336133	**FDPS**	29	9.7×10^{-13}	rs41314549	T / C	2.81	13.7 ± 1.9
17	43056905-45876022	GOSR2	26	8.3×10^{-22}	rs17608766	T / C	14.3	5.73 ± 0.90
3	99373762-100592217	**FILIP1L***	25	8.4×10^{-14}	rs9811920	G / A	40.8	3.23 ± 0.65
16	4001196-5118345	**SRL**	24	9.4×10^{-12}	rs889807	T / C	50.8	3.24 ± 0.64
1	21736588-23086883	KDM1A	22	2.0×10^{-12}	rs35001652	G / A	37.3	2.59 ± 0.67
7	45952922-46986720	IGFBP3	22	2.0×10^{-11}	rs143741275	A / AGTGTGT	42.4	2.59 ± 0.66
5	171074292-172678327	NKX2-5	21	9.0×10^{-14}	rs35564079	C / CT	28.5	2.76 ± 0.72
3	13070799-14816900	TMEM43	21	2.7×10^{-11}	rs900173	T / C	34.0	3.48 ± 0.68
1	144977494-148361253	**GJA5**	20	1.1×10^{-10}	rs12046416	A / G	33.5	2.81 ± 0.68
1	235819436-237555628	ACTN2	20	2.0×10^{-12}	rs12142143	T / C	53.1	4.11 ± 0.65
13	20686720-22242174	FGF9	20	8.8×10^{-13}	rs10628955	G / GAA	47.7	3.95 ± 0.67
4	111256567-113870102	PITX2	19	4.4×10^{-14}	rs2723294	C / T	69.5	4.65 ± 0.70
1	14891511-16897730	HSPB7	17	2.3×10^{-11}	rs1763605	T / G	67.5	2.80 ± 0.68
2	146445570-147277162	ACVR2A/ZEB2	16	2.9×10^{-11}	rs573709385	A / AT	44.9	1.93 ± 0.65
6	125424383-127540461	HEY2	16	2.9×10^{-11}	rs11423823	C / CT	50.2	3.23 ± 0.66
2	36122006-38132712	**STRN**	15	9.9×10^{-16}	rs2110944	T / C	52.6	3.71 ± 0.64
16	79134815-80297374	**MAF***	15	5.2×10^{-11}	rs558328129	A / AT	45.4	2.97 ± 0.70
8	107410754-108648177	ABRA	15	8.1×10^{-11}	rs72007904	A / AACTATTC	50.0	3.87 ± 0.64
21	27271019-29125226	**ADAMTS1**	15	1.4×10^{-10}	rs2830977	G / A	22.0	4.29 ± 0.78
1	49894177-51713726	**RNF11***	13	3.1×10^{-11}	rs7555411	C / T	1.42	13.8 ± 2.7
13	98938919-100574095	**DOCK9***	13	1.2×10^{-10}	rs34138434	C / A	29.7	4.62 ± 0.72
2	118367466-121303783	EN1*	13	6.7×10^{-14}	rs162746	A / G	67.5	4.01 ± 0.68
12	27799773-29651255	**CCDC91***	12	2.0×10^{-14}	rs5797270	G / GT	20.3	3.35 ± 0.81
13	25784362-27284362	WASF3*	11	9.3×10^{-11}	rs61944841	G / A	41.3	3.07 ± 0.67
18	19485844-20649472	**GATA6**	11	6.6×10^{-11}	rs62094198	T / A	39.6	2.21 ± 0.66
14	19002084-21589402	**NDRG2**	10	2.5×10^{-11}	rs12889267	A / G	16.7	3.43 ± 0.85
6	35455756-37572596	CDKN1A	10	3.2×10^{-13}	rs3176326	G / A	19.8	3.47 ± 0.81
7	118351581-121045273	**WNT16**	10	2.4×10^{-11}	rs3801387	A / G	28.1	2.79 ± 0.72
16	75977954-77523678	**ADAMTS18**	10	5.2×10^{-13}	rs62046468	C / T	37.6	4.80 ± 0.66
17	41772087-43056905	**SOST**	10	5.5×10^{-11}	rs17881550	G / GC	43.4	3.65 ± 0.65
11	27020461-28481593	**CCDC34***	8	5.7×10^{-12}	rs10835164	C / T	25.6	3.84 ± 0.74
5	120452166-122556905	PRDM6	8	6.2×10^{-11}	rs463106	T / C	47.2	3.68 ± 0.65
17	67858770-69387817	KCNJ2/SOX9	7	3.6×10^{-12}	rs56229089	G / C	55.7	2.89 ± 0.65
5	63968304-65911286	ADAMTS6	7	1.6×10^{-11}	rs753963943	ATT / A	42.5	2.88 ± 0.66
18	8498931-11075913	**NDUFV2**	6	6.1×10^{-12}	rs206524	T / C	70.7	3.01 ± 0.71
11	65898631-68005825	**KDM2A**	4	2.1×10^{-11}	rs12785906	G / C	5.84	7.40 ± 1.38
1	1892607-3582736	PRDM16	3	1.7×10^{-11}	rs781212641	G / GC	9.33	5.68 ± 1.12
4	7539692-8152235	**AFAP1**	3	1.4×10^{-11}	rs28542374	G / A	63.5	3.10 ± 0.67
12	76511314-78570570	NAV3	3	2.7×10^{-12}	rs7965680	T / C	55.7	2.43 ± 0.65
17	15019097-16412342	**CENPV***	3	3.4×10^{-11}	rs7477	A / C	49.8	3.18 ± 0.64

Table 6.2: Counts of GWAS hits across runs in the UPE framework, $\mathcal{C}_\ell$ for each locus ℓ, which represents the number of runs for which the corresponding locus shows at least one association with $p < p_{\mathrm{GW}} = 5 \times 10^{-8}$ (see details in the Methods section). Note that the total number of runs was 36. Genes with an asterisk were annotated based purely on proximity to the lead variant in that region. Gene names with no asterisk have additional prior evidence of a link to cardiac physiology. When a gene name is written in bold letters, it means that the corresponding locus is only discovered with UPE, i.e. the remaining loci were already found by testing either the handcrafted phenotypes or the shape PCs.

data, *SHOX2* is not highly expressed in these tissues. A possible hypothesis is that rs11706187 regulates the expression of *SHOX2* in developmental or pre-adult stages.

Genes *HSPB7* and *CDKN1A* have been previously found in [78] in association with LVESV and LVEF.

A particularly interesting association, with the SNP rs2245109, is located within the body of

the *STRN* gene on chromosome 2 and is probably causally related to it: this gene encodes the protein striatin, which is expressed in cardiomyocytes and has been shown to interact with other proteins involved in the mechanism of myocardial function [68]. Mutations in this gene have been shown to lead to DCM in dogs [66]. In humans, there has been a recent GWAS on heart failure (HF) that reported this locus [57], but ours is the first study to link it with cardiac morphology. Moreover, our estimated effect size is significantly higher; suggesting that this latent variable is an endophenotype closer to the underlying biology. This could provide insight to unravel the aetiology of a heterogeneous condition such as HF. Interestingly, the lead SNP has a high minor allele frequency (MAF), of 47.4%. This locus also contains eQTLs for this gene, as evidenced by TWAS (see subsection Transcriptome-wide association analysis).

A comparable pattern is observed at the RNF11 locus, although this does not reach genome-wide significance for HF ($p = 3.2 \times 10^{-6}$). The lead variant for this locus is an indel with low frequency (MAF=1.4%) and large estimated effect size (standardised $\hat{\beta}$ =0.138). This locus has also been linked to the QRS interval, although the causative gene is not clear [93], some candidates being *RNF11* itself, *CDKN2C*, *C1orf185* and *FAF1*.

The *SRL* gene, which encodes the sarcalumenin protein, harbours the SNP rs889807. Sarcalumenin is a protein that binds Ca^{2+} located in the longitudinal sarcoplasmic reticulum (SR) of the heart. Its main function is to regulate Ca^{2+} reuptake in the SR by interacting with the cardiac sarco (endo)plasmic reticulum Ca^{2+}-ATPase 2a (SERCA2a). According to GTEx data, this gene is highly expressed in adult cardiac tissue (both in the left ventricle and atrial appendage tissues) and skeletal muscle.

Interestingly, several associations lie near genes of the *ADATMS* (a disintegrin and metalloproteinase with thrombospondin motifs) family [89]: *ADAMTS1* and *ADAMTS5* (near rs2830977 on chromosome 21, with $p = 1.4 \times 10^{-10}$), *ADAMTS6* (rs753963943 on chromosome 5, $p = 5.6 \times 10^{-11}$) and *ADAMTS18* (chromosome 16, $p = 5.2 \times 10^{-13}$).

An association lies 260 kb upstream of *GATA6*, a transcription factor that plays a critical role in the development of the heart. It has been found to regulate the hypertrophic response [98]. Sequence variants in this gene have been discovered to predispose for CHD phenotypes [63, 107].

rs12889267 lies 3700 kb upstream of the TSS of *NDRG2*. This gene has been demonstrated to play a role in protection against ischemia/reperfusion injury, in a study in rats [96].

One SNP overlaps *KDM2A*. As *KDM1A*, it is a histone demethylase gene. Although its link to the heart is less clear, there exists evidence from knockout studies in mice that supports its importance in embrionic development, including heart development [48].

rs206524 is located within a gene for long non-coding RNA, *LINC01254*. A possible candidate protein-coding gene is *NDUVF2*, located 1.3 Mb upstream of the SNP. According to the GTEx dataset, *NDUFV2* is highly expressed in cardiac and skeletal muscle tissue.

Novel loci. In addition to the loci with prior evidence discussed above, we report a number of novel genetic loci with $p < p_{\mathrm{SW}}$, which have not been previously reported in connection with cardiac phenotypes or pathways. These loci were annotated based on the closest gene: *CCDC91*, *FILIP1L*, *EN1*, *GJA5*, *ACVR2A*, *AFAP1*, *IGFBP3*, *CCDC34*, *WASF3*, *DOCK9*, and *MAF*. Of particular interest are those loci with a small number of counts, e.g. $\mathcal{C}_\ell \leq 15$. These are the loci for which the ensemble approach seems most relevant, since they are unlikely to be pinpointed by one particular run. Furthermore, they are typically not found by testing the shape PCs.

Loci with suggestive significance. In addition to genetic loci with $p < p_{\mathrm{SW}}$, several SNPs show $p_{\mathrm{SW}} < p < p_{\mathrm{GW}}$ in 5 or more independent runs. We consider these associations suggestive and briefly discuss some of them here. The summary statistics for these associations are shown in Table 6.3.

Some of these loci have been found in previous investigations: GWAS, familial studies, or studies with model organisms. For example, variants in gene *RBM20* are associated to DCM [53]. We observe that the lead SNP in this region has a low MAF (1.4%), and the effect size estimate is high (standardised $\hat{\beta} = 0.20$).

A cluster of associations in chromosome 1 is located in a region that includes the *S100* family of genes. In particular, the lead SNP in this region, rs985242, is located within the body of genes *S100A1* and *S100A13*. The S100 is a family of low-weight Ca^{2+}-binding EF-hand proteins, with 25 human genes identified.

The SNP rs28681517 lies within gene *ADAMTSL3*, whose associated protein has been shown to play a crucial role in maintaining cardiac structure and function in mice [88].

SNP rs569550 lies 578846 base pairs away from *KCNQ1*, which belongs to a large family of genes that provide instructions for making potassium channels. *KCNQ1* encodes the alpha

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF (%)	$\lvert\hat{\beta}\rvert \pm \mathrm{se}(\hat{\beta})(\times 10^{-2})$
10	110317705-112561493	RBM20	21	2.8×10^{-10}	rs189569984	C	T	0.9	19.3 ± 3.4
15	48136048-50008043	FBN1	21	1.7×10^{-10}	15:48690566_TC_T	TC	T	85.6	3.23 ± 0.96
12	53039004-54778823	ATG4D/S1PR5*	15	6.4×10^{-09}	rs12810579	G	C	1.25	19.2 ± 4.1
15	84260468-86652905	ADAMTSL3	13	3.9×10^{-10}	rs2585058	G	A	49.5	4.02 ± 0.64
20	34960446-36909530	KIAA1755	13	2.3×10^{-10}	rs41282820	G	A	1.73	10.7 ± 2.48
16	60054-1207206	LMF1*	11	3.8×10^{-10}	rs79523980	T	C	35.7	3.80 ± 0.69
17	45876022-47517400	SKAP1*	11	1.2×10^{-09}	rs17697950	A	G	24.7	2.37 ± 0.74
1	59890409-61922365	NFIA*	11	1.6×10^{-10}	rs2474370	C	T	67.9	2.89 ± 0.69
3	170964909-172295731	FNDC3B	10	7.4×10^{-10}	rs17460016	G	C	40.1	3.34 ± 0.66
17	56-1172399	VPS53*	9	4.0×10^{-09}	rs16954854	G	A	10.0	6.29 ± 1.07
6	45406563-47311898	RUNX2	9	1.9×10^{-09}	6:45452929_CT_C	CT	C	24.9	3.62 ± 0.79
1	153180829-154770403	S100A1	9	9.4×10^{-10}	rs985242	C	G	53.2	3.33 ± 0.65
11	1213590-3665481	KCNQ1	9	4.6×10^{-10}	rs569550	T	G	38.6	3.1 ± 0.66
3	137371083-139954597	NME9*	8	2.9×10^{-10}	rs13059110	G	T	13.0	5.69 ± 0.96
8	17387876-17836399	PDGFRL*	7	5.7×10^{-09}	rs2299575	C	T	83.1	3.88 ± 0.86
1	200137649-201589975	ZNF281	7	1.2×10^{-09}	rs10753873	A	T	58.3	2.58 ± 0.65
5	127344604-129519025	SLC27A6	7	1.1×10^{-09}	rs1898547	G	A	32.1	3.79 ± 0.69
4	22319347-24135529	PPARGC1A	7	3.1×10^{-10}	rs73243622	C	T	25.4	4.17 ± 0.74
4	174264132-176570716	HAND2	7	1.2×10^{-09}	rs12502027	A	G	36.1	2.49 ± 0.69
7	116780178-118351581	WNT2	6	1.6×10^{-09}	rs5004797	T	C	18.2	4.31 ± 0.84
6	1452362-2458936	GMDS	6	1.6×10^{-09}	rs6934958	T	C	54.9	3.07 ± 0.65
3	168580960-170964909	SAMD7*	6	2.6×10^{-09}	rs201527389	A	AT	3.1	7.83 ± 1.96
3	133252173-135456906	EPHB1	6	5.3×10^{-09}	rs79656429	A	G	6.99	5.34 ± 1.28
3	7083387-8648561	LMCD1	6	9.4×10^{-10}	rs9814240	G	A	42.0	2.83 ± 0.65
2	54685226-56203345	EFEMP1	6	2.0×10^{-10}	rs1430197	G	A	37.8	2.39 ± 0.67
5	36433954-38802410	NIPBL*	5	5.6×10^{-09}	rs292178	C	G	43.9	3.30 ± 0.65
11	90966490-92077144	FAT3	5	1.2×10^{-08}	rs72972727	T	G	10.2	4.98 ± 1.06
1	219590571-221858231	HLX*	5	4.1×10^{-09}	rs115466747	C	T	6.04	6.16 ± 1.35
17	63148128-64800430	PRKCA	5	1.8×10^{-09}	rs569258685	A	AT	53.4	3.08 ± 0.66
12	65559695-67181144	HMGA2	5	8.9×10^{-09}	rs8756	C	A	51.9	2.96 ± 0.64

Table 6.3: Summary statistics for the 24 suggestive associations obtained via the UPE framework. The count column presents the number of runs $\mathcal{C}_\ell$ for which locus ℓ shows at least one association with $p < p_{\mathrm{GW}} = 5 \times 10^{-8}$ (see details in the Methods section). p-values are one-sided and derived from a linear association t-statistic (no adjustments were made for multiple comparisons). Note that, as before, the total number of runs was 36. Genes with an asterisk were annotated based purely on proximity to the lead variant in that region. Gene names with no asterisk have additional prior evidence of a link to cardiac physiology.

subunit of the potassium channel KvLQT1. Mutations in *KCNQ1* are responsible for the long QT syndrome [19].

Deletion 15:48690566_TC_T is a relatively common variant (MAF = 14.4%), and is located 10 kb downstream of the TES of *FBN1*. Mutations in this gene are associated with Marfan syndrome, a genetic disorder that affects connective tissues in the body. It can have various manifestations, including cardiovascular complications.

rs9814240 is a coding variant in the *LMCD1* gene. Interestingly, mutations in this gene are causative of HCM in mice [39]; however, no association had been found between variants in this gene and human cardiac phenotypes. Moreover, this gene has been found to interact with (the homologous of) *GATA6* in mice [82] (*GATA6* is located near one of the loci discovered with study-wide significance.)

6.5.3 Effect of latent variables on LV morphology

It is of key importance, not only to discover significant associations, but to understand what is their effect on shape. An interpretation of the impact on LV morphology of the latent variables linked to different loci was achieved by examining the average shape of subjects located at different quantiles throughout the distribution. Prior to averaging, the sets of meshes were unscaled, and then scaled back after averaging. The quantile ranges used were: [0, 0.01], [0.095, 0.105], [0.495, 0.505], [0.895, 0.905] and [0.99, 1]. Note that, since there are over 50,000 subjects in our database, each quantile range (of 1% in width) encompasses more than 500 subjects. Reassuringly, in all cases we observe a smooth transition in shape from lower to higher values. This is shown in Fig. 6.5, along with the associated Manhattan plots, for the loci *PLN*, *TTN* and *STRN*.

The effect of these loci on the LV morphology was evaluated by selecting the single phenotype with the strongest p-value for the associated locus.

Correlation of latent factors with traditional indices. To further help characterise the effect of these latent variables, the Spearman correlation coefficient between the latter and the handcrafted LV indices were calculated and are shown in Table 6.4. These indices were LVEDV, LV sphericity index at end-diastole (LVEDSph), LV myocardial mass (LVM) and LV mass-to-volume ratio (LVMVR = LVM / LVEDV).

Interestingly, we observe a very distinct effect of each of these SNPs on the morphology. While the *PLN* variant influences a latent variable that is mainly linked to LV sphericity (Spearman $r = 0.625$) with a relatively small effect on LVEDV ($r = 0.434$), the *TTN* gene shows a greater correlation with the latter ($r = 0.889$). Consistent with this, the GWAS on LVEDSph shows no signal for *TTN*, but a strong one for *PLN* ($p = 10^{-15}$, see Figure 5.6), which is also in line with a previous finding of ours [18].

The SNP in the *STRN* gene is associated with a subtle phenotype that controls mitral orientation without a concomitant change in LV size (see Figure 6.5). This is consistent with the fact that it was not discovered in previous studies of structural LV phenotypes.

On the other hand, the *TBX5* gene is linked to a latent variable that, as for *TTN*, is mainly correlated with LVEDV with no correlation with LVEDSph. For the developmental gene *NKX2-5*, we note that the associated latent variable has an effect both on LVEDV ($r = 0.865$) and on

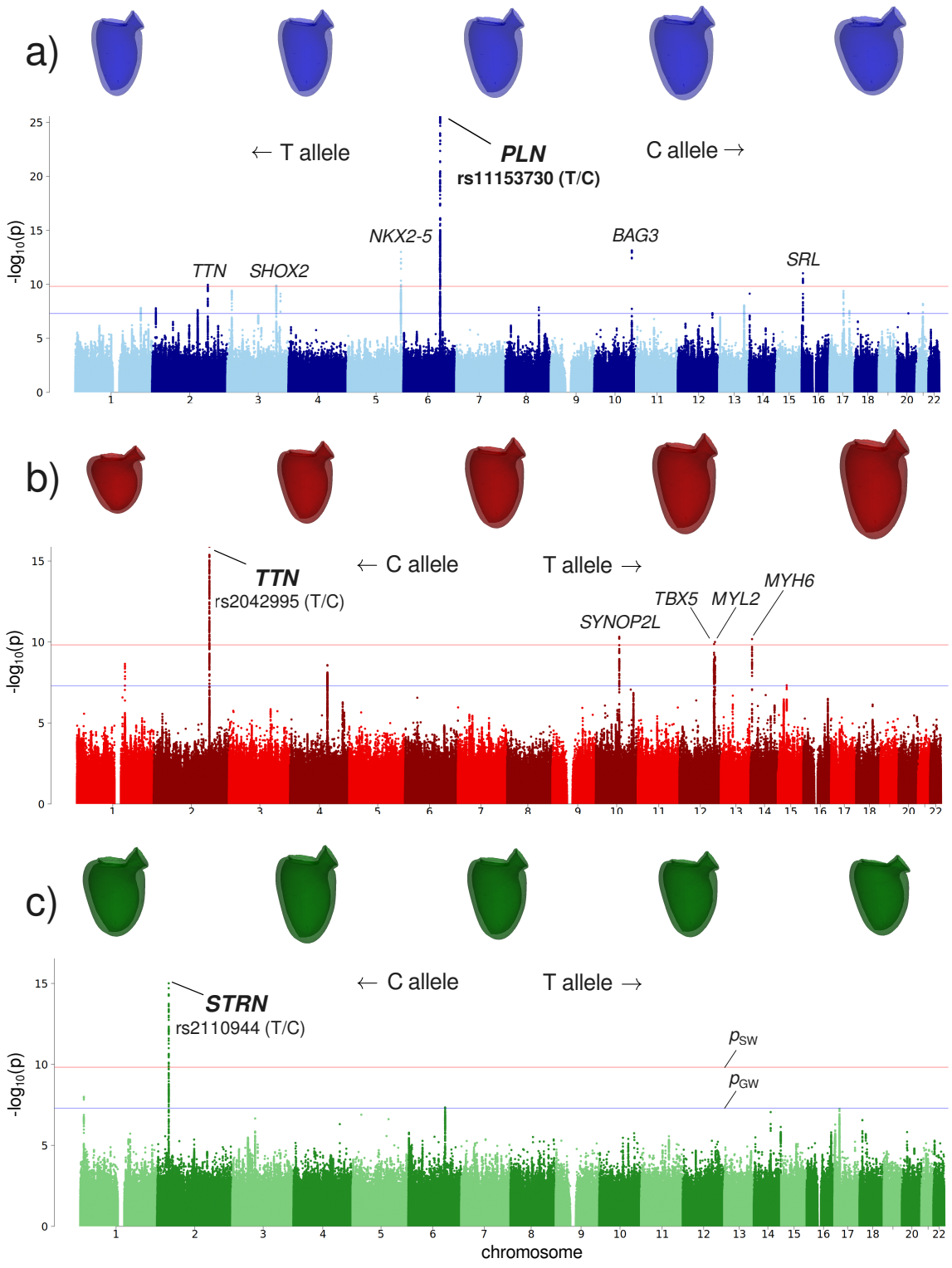

Figure 6.5: Manhattan plots for LV latent variables with best association for SNPs at the a) *PLN*, b) *TTN* and c) *STRN* loci. On top are shown the average meshes corresponding to the following range of quantiles, for each latent variable: [0, 0.01], [0.095, 0.105], [0.495, 0.505], [0.895, 0.905] and [0.99, 1].

LVEDSph ($r = 0.372$).

6.5.4 Replication study

We set apart a subset of 5,470 UKB subjects of British ancestry for which the whole pipeline was run identically to the individuals from the discovery set. While merging the replication and discovery sets would have increased statistical power, we decided to leave a subset out of the discovery phase to provide reassurance as to the lack of data dredging during that phase.

We report the detailed results in Table 6.5, including the estimated statistical power for each SNP, based on the effect size estimate $\hat{\beta}$ from the discovery phase. Among the 49 study-wide significant loci, we report 28 that replicate with $p < 0.05$ (whereas 7 replicate with the more stringent Bonferroni threshold of $p < 0.05/49$), as well as 47 loci for which the estimated direction of effect is consistent with that found in the discovery phase. For the suggestive associations, 11 loci replicated (out of 24) with the threshold of $p < 0.05$, whereas 22 have a concordant direction of effect between the discovery and replication phases. These replication rates are consistent with the estimates of the power (present in last column of Table 6.5), given the size of the replication sample.

6.5.5 Gene enrichment analysis

We use the tool g:Profiler to find pathways for which our sets of genes were enriched. To define the gene sets, we selected a region of 100 kb around each lead variant and chose the genes whose transcription start site (TSS) was located within that window. Gene ontology terms belong to one of three different categories: molecular functions (MF), cellular components (CC), and biological processes (BP).

Within the CC category, we have found a relevant enriched term, "Sarcomere", comprising the following 9 genes from our query: *ACTN2*, *MYOZ1*, *SYNPO2L*, *BAG3*, *TNNT3*, *TNNI2*, *MYH6*, *MYH7*, *KY* ($p_{\text{adj}} = 9.2 \times 10^{-3}$).

Within the BP category, the terms "Myofibril assembly", "striated muscle cell development" and "sarcomere organization" result enriched ($p_{\text{adj}} = 1.2 \times 10^{-3}$, $p_{\text{adj}} = 1.4 \times 10^{-3}$ and $p_{\text{adj}} = 1.5 \times 10^{-3}$, respectively).

Within the MF category, the term "calcium-dependent protein binding" is enriched ($p_{\text{adj}} = 2.9 \times 10^{-8}$), although it is comprised of 9 members of the *S100A* family (which encompass a

Loci	LVEDV	LVEDSph	LVM	LVMVR
NKX2-5	-0.546	-0.470	-0.287	0.409
EN1*	0.165	0.678	0.161	0.020
LMF1*	0.839	0.273	0.697	-0.194
KCNQ1, MYL2	0.885	0.218	0.809	-0.086
FILIP1L*	-0.815	-0.121	-0.749	0.077
HEY2	0.907	0.092	0.894	0.011
FDPS	0.197	-0.526	0.251	0.090
BAG3, TMEM43	-0.809	-0.445	-0.685	0.167
STRN	-0.331	0.452	-0.341	-0.045
EPHB1	-0.728	0.311	-0.696	0.047
DOCK9*	0.239	0.675	0.221	-0.023
ACTN2	0.335	-0.521	0.486	0.251
SAMD7*	0.093	-0.147	0.114	0.041
CDKN1A	0.316	-0.486	0.416	0.161
NDUFV2, SKAP1*	0.436	0.351	0.480	0.112
WASF3*	-0.263	0.527	-0.286	-0.016
GATA6	0.240	-0.281	0.079	-0.278
KDM2A, CCDC34*, CCDC91*	0.101	0.431	0.019	-0.118
PITX2	-0.091	-0.237	-0.204	-0.207
MAF*	-0.711	0.256	-0.658	0.063
ADAMTS1	0.094	0.617	0.004	-0.119
KIAA1755	-0.223	-0.575	-0.136	0.100
MYH6	0.904	-0.183	0.837	-0.080
PIGL*	-0.499	-0.122	-0.605	-0.198
TTN	0.910	-0.187	0.855	-0.067
WNT16	0.055	0.036	0.279	0.366
PRDM16	-0.683	0.432	-0.600	0.123
AFAP1	0.673	0.367	0.724	0.119
WNT2	0.034	0.456	0.048	0.049
HAND2	0.250	-0.306	0.246	-0.009
ZNF281	0.755	-0.286	0.819	0.133
NME9*, VPS53*	0.210	-0.663	0.170	-0.070
ACVR2A	0.191	-0.682	0.260	0.093
FBN1	-0.381	-0.264	-0.282	0.139
ADAMTS18	-0.231	0.446	-0.428	-0.334
EFEMP1	0.518	0.334	0.398	-0.146
GOSR2	-0.748	0.064	-0.771	-0.070
SRL, PLN, FNDC3B	0.722	0.532	0.567	-0.212
KDM1A	-0.257	0.702	-0.317	-0.141
RUNX2	-0.384	-0.436	-0.473	-0.193
PDGFRL*	-0.369	-0.339	-0.422	-0.132
RBM20	-0.937	-0.111	-0.847	0.108
ABRA	0.366	0.540	0.346	-0.001
PRDM6	-0.496	0.377	-0.648	-0.257
GJA5	0.549	-0.514	0.534	-0.024
ADAMTS6	-0.249	-0.549	-0.311	-0.134
HSPB7	-0.761	-0.148	-0.578	0.281
FGF9, RNF11	-0.100	0.540	-0.166	-0.089
LMCD1	0.726	0.350	0.566	-0.233
SOST	-0.365	-0.426	-0.110	0.397
KCNJ2, S100A1	0.143	0.429	0.094	-0.079
MYOZ2	0.912	-0.060	0.845	-0.079
SHOX2	0.734	0.232	0.596	-0.193
IGFBP3	0.886	-0.035	0.880	0.034
SLC27A6	0.293	0.510	0.038	-0.395
NFIA*, PPARGC1A	-0.273	0.566	-0.363	-0.130
NDRG2	-0.657	-0.317	-0.700	-0.101
TBX5	-0.728	0.403	-0.714	0.018
ADAMTSL3	0.158	-0.015	0.154	-0.013
ATG4D/S1PR5*	-0.926	0.002	-0.758	0.255
NAV3	-0.341	0.670	-0.380	-0.062
SYNPO2L	-0.734	0.353	-0.821	-0.178

Table 6.4: Spearman correlation between the best latent variable per locus from UPE, and the four LV handcrafted indices. The signs of the correlations were switched so that the effect sizes from Table 1 are always positive, with the choice of effect alleles described in that table.

gene	direction concordance	replication p-value	power ($\alpha = 0.05$)
ABRA	✓	3.0×10^{-2}	0.645
ACTN2	✓	6.8×10^{-2}	0.627
ACVR2A	✓	5.4×10^{-3}	0.622
ADAMTS1	✓	4.4×10^{-2}	0.569
ADAMTS18	✓	1.2×10^{-2}	0.665
ADAMTS6	✓	1.4×10^{-2}	0.614
AFAP1	✓	2.4×10^{-2}	0.648
BAG3	✓	3.6×10^{-2}	0.822
CCDC34*	✓	1.4×10^{-1}	0.652
CCDC91*	✓	5.5×10^{-2}	0.719
CDKN1A	✓	5.4×10^{-3}	0.665
DOCK9*	✓	1.6×10^{-1}	0.594
EN1*	✓	5.9×10^{-2}	0.691
FDPS	✗	2.9×10^{-1}	0.665
FGF9	✓	1.0×10^{-1}	0.699
FILIP1L*	✓	2.6×10^{-3}	0.709
GATA6	✓	2.5×10^{-2}	0.574
GJA5	✓	4.8×10^{-2}	0.585
GOSR2	✓	8.2×10^{-5}	0.892
HEY2	✓	3.8×10^{-5}	0.617
HSPB7	✓	5.5×10^{-3}	0.592
IGFBP3	✓	2.2×10^{-2}	0.623
KCNJ2	✓	1.3×10^{-1}	0.635
KDM1A	✓	8.1×10^{-6}	0.638
KDM2A	✓	1.7×10^{-1}	0.612
MAF*	✗	3.9×10^{-1}	0.652
MYH6	✓	4.1×10^{-2}	0.700
MYL2	✓	7.7×10^{-7}	0.735
MYOZ2	✓	1.8×10^{-1}	0.697
NAV3	✓	9.2×10^{-2}	0.638
NDRG2	✓	3.2×10^{-2}	0.594
NDUFV2	✓	4.8×10^{-4}	0.634
NKX2-5	✓	1.1×10^{-1}	0.691
PIGL*	✓	1.4×10^{-3}	0.568
PITX2	✓	5.5×10^{-2}	0.707
PLN	✓	2.5×10^{-7}	0.960
PRDM16	✓	4.0×10^{-2}	0.620
PRDM6	✓	6.0×10^{-2}	0.582
RNF11	✓	5.5×10^{-2}	0.622
SHOX2	✓	6.2×10^{-2}	0.740
SOST	✓	2.9×10^{-2}	0.589
SRL	✓	2.7×10^{-2}	0.603
STRN	✓	1.2×10^{-2}	0.766
SYNPO2L	✓	6.0×10^{-1}	0.761
TBX5	✓	2.2×10^{-1}	0.634
TMEM43	✓	9.1×10^{-2}	0.601
TTN	✓	2.1×10^{-4}	0.790
WASF3*	✓	4.0×10^{-3}	0.614
WNT16	✓	2.8×10^{-1}	0.621

Table 6.5: Replication results for the study-wide significant loci from the discovery phase. The estimated power for each genetic variant is for a level $\alpha = 0.05$.

single locus), apart from *SYT8* and *TNNT3*.

These enriched pathways further support the biological relevance of the phenotypes extracted via our unsupervised approach.

6.5.6 Transcriptome-wide association analysis

We performed TWAS using the S-PrediXcan tool [14], to test the possibility of a mediating effect of gene expression and intron excision events on structural phenotypes. This tool is fed with models that impute gene expression and intron excision data based on the genotype, and is trained using data from the Genotype-Tissue Expression (GTEx) project, version 8 [28].

Our focus was on cardiovascular tissues, specifically the left ventricle, atrial appendage, and coronary, aortic, and tibial arteries. To maintain statistical rigor, we applied a significance threshold of $p_{\mathrm{GEx}} = 2.2 \times 10^{-9}$, which adjusts for multiple comparisons (324 phenotypes and 68,919 tissue-gene pairs). Similarly, for alternative splicing, the threshold was set at $p_{\mathrm{AS}} = 8.2 \times 10^{-10}$, considering the same multiple testing correction (the number of intron-tissue pairs tested being 187,535).

In the cardiac tissues (left ventricle and atrial appendage), we identified genes located within loci of previously reported genes. In the left ventricle, these included *NKX2-5*, *STRN*, *SYNPO2L* (*FUT11*, *SEC24C*, and *SYNPO2L* itself), *PLN*, *HEY2* (*CENPW* gene), *TTN* (*FKBP7* gene), *CENPV*, *GOSR2* (*MAPT* and *GOSR2* itself), and *FDPS* (*SCAMP3*, *ARHGEF2*, *RIT1*, *GOSR2*, *MAPT*, *HCN3*, *GBA*, *MSTO1*, *RUSC1*, *FUT11*, *SYT11*, *ADAM15*, and *FDPS* itself). For the atrial appendage, the genes included *PLN*, *STRN*, *NKX2-5*, *SYNPO2L*, and *MYOZ1* within the *SYNPO2L* locus, as well as *FKBP7* and *SCAMP3*. Many of these genes had been previously implicated based on independent knowledge, bolstering the evidence for their potential causal roles.

Notably, our analysis also revealed the direction of the effect on gene expression: higher *PLN* expression was associated with a more spherical left-ventricular morphology, while lower *NKX2-5* expression was linked to the same phenotype (refer to Figure 6.5b). Furthermore, an elevated *STRN* expression (in both cardiac tissues) was associated with a more horizontal mitral orientation (see Figure 6.5c).

In the case of arterial tissues, we found significant associations within various loci, such as *SYNPO2L* (*AGAP5*, *FUT11*, *SEC24C*, and *ARHGAP27*), *FDPS* (*ARHGEF2*, *CLK2*, *FAM189B*, *GBA*, *GON4L*, *HCN3*, *NPR1*, and *SYT11*), *CENPW*, *TTN* (*PRKRA* and *FKBP7* genes), *PLN* (*CEP85L* and *PLN*), *GOSR2* (*WNT3*, *CRHR1*, *LRRC37A*, and *MAPT*), *KDM2A*, *LINC01562*, *MYH6* (*MYH6* and *MYH7*), *RP11-383I23.2*, *RP11-574K11.29*, *SCAMP3*, *MYL2* (*SH2B3*

gene), *SOST*, and *TCF21*.

6.5.7 Link of our GWAS hits to other phenotypes and diseases.

IEU OpenGWAS. To detect pleiotropic effects, we performed a phenome-wide association study (PheWAS) of the lead SNPs from Table 6.2. For this, we queried the IEU OpenGWAS Project's database. The results are included in the Supplementary Data File for the associated publication [17]. We discuss briefly here some associations with cardiovascular phenotypes.

A number of loci were associated to cardiac electrical phenotypes: *CDKN1A*, *NDRG2*, *PLN*, *TBX5*, *MYH6*. The following loci were associated to pulse rate: *SYNPO2L*, *NDRG2*, *MYH6*, *SRL*, *GOSR2*, *GATA6*, *ACTN2*, *KIAA1755*, *TMEM43*, *SLC27A6* and *FNDC3B*. The lead SNP at the *PRDM6* locus was associated to heart rate recovery post-exercise.

The following loci were associated to blood pressure phenotypes (diastolic, systolic or hypertension): *SYNPO2L*, *KCNQ1*, *MYL2*, *NDRG2*, *MYH6*, *SRL*, *GOSR2*, *GATA6*, *HSPB7*, *RNF11*, *EFEMP1*, *FNDC3B*, *NME9*, *PRDM6* and *PLN*.

Finally, *SYNPO2L*, *TBX5*, *MYH6*, *GOSR2*, *PITX2* and *CDKN1A* were associated to cardiac arrhytmias.

Comparison with GWAS on traditional LV indices

For comparison, we collected the GWAS summary statistics from previous studies on LV phenotypes, derived also from UKB CMR images, namely: [9], [78] and [67]. We also include LVESV, LVSV and LVEF. Note, however, that the unsupervised features studied in this chapter are static and were extracted using only the end-diastolic (ED) phase.

The comparison can be seen in figure 6.6. For each locus in 6.2 (which all passed the more stringent Bonferroni threshold, p_{SW}), this figure displays the association p-value found in previous GWAS. Shades of red represent non-genome-wide significant associations, whereas shades of blue represent genome-wide significant ones, and white corresponds to the p_{GW} threshold. The second row represents the best p-value across all the traditional phenotypes for the loci given in the columns.

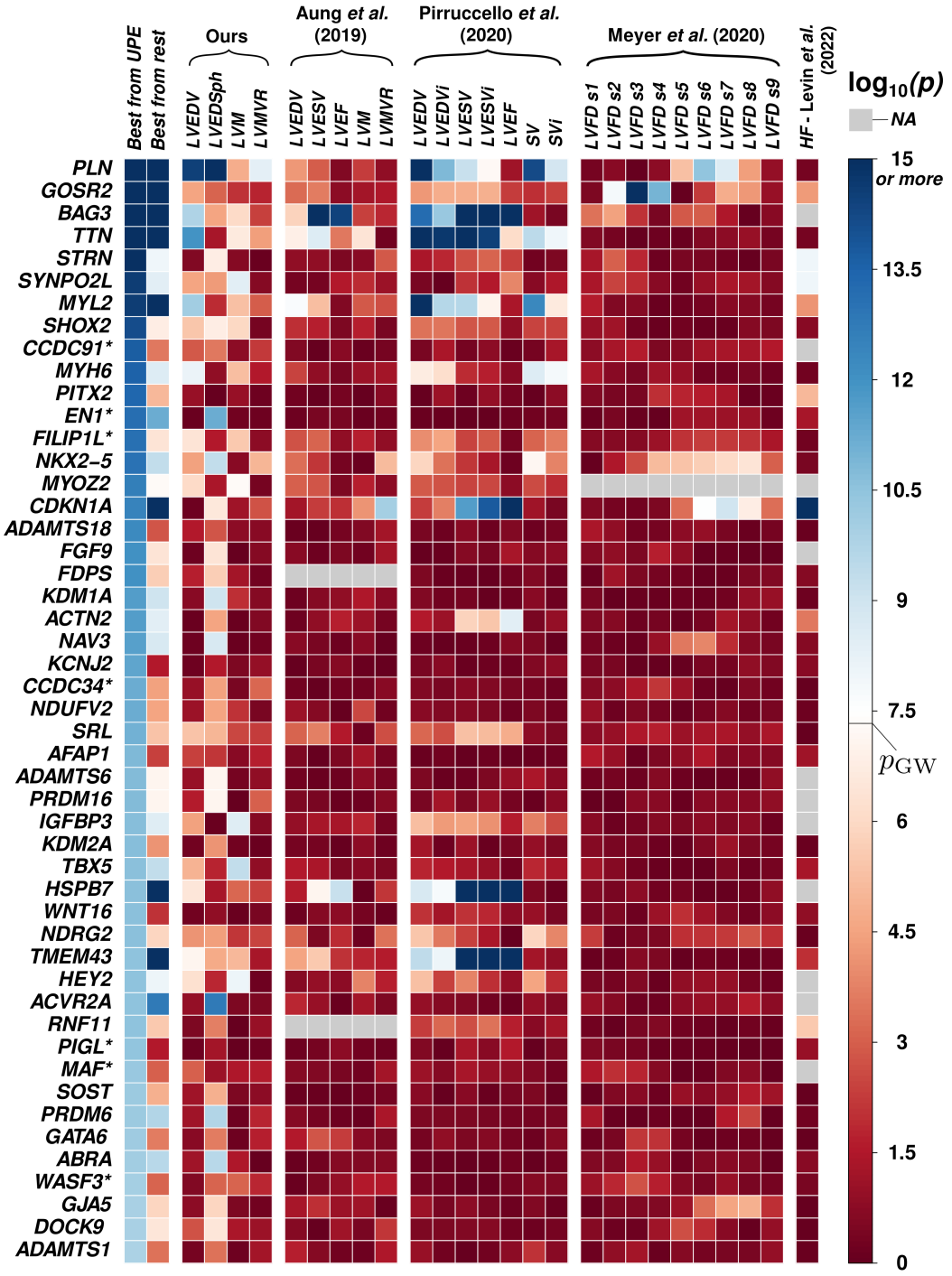

Figure 6.6: Comparison of the $-\log_{10}(p)$ values for the 49 study-wide significant genetic loci found in this work, with GWAS on handcrafted cardiac indices. The top row corresponds to the best association found for that locus across the ensemble of phenotypes, whereas the second row corresponds to the best p-value for that locus across the previous GWAS. White colour corresponds to the genome-wide significance threshold of 5×10^{-8}, whereas shades of red and blue correspond to weaker and stronger associations, respectively. SV denotes stroke volume. LVEDVi, LVESVi and SVi denote the indexed versions of the phenotypes, i.e. the phenotype divided by the subject's body surface area.

6.6 Can the learned features be refined for SNP associations?

In this section, we ask whether it is possible to refine the CoMA-derived shape phenotypes for genetic association. These results are chronologically previous to the ones exposed so far, and were presented in the MICCAI 2021 conference. For this piece of work, we used the output of the SPASM segmentation model on over 29,000 subjects of British ancestry [6].

We propose training a mesh autoencoder in two stages, described by loss functions $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. The first one is genotype-agnostic and includes only the reconstruction and KL regularisation terms, whereas the second also incorporates genotype information via a SNP-embedding correlation term that induces an alignment between the phenotypic scores (the embeddings) and SNP dosages (for a set of selected SNPs). This is described in Algorithm 1.

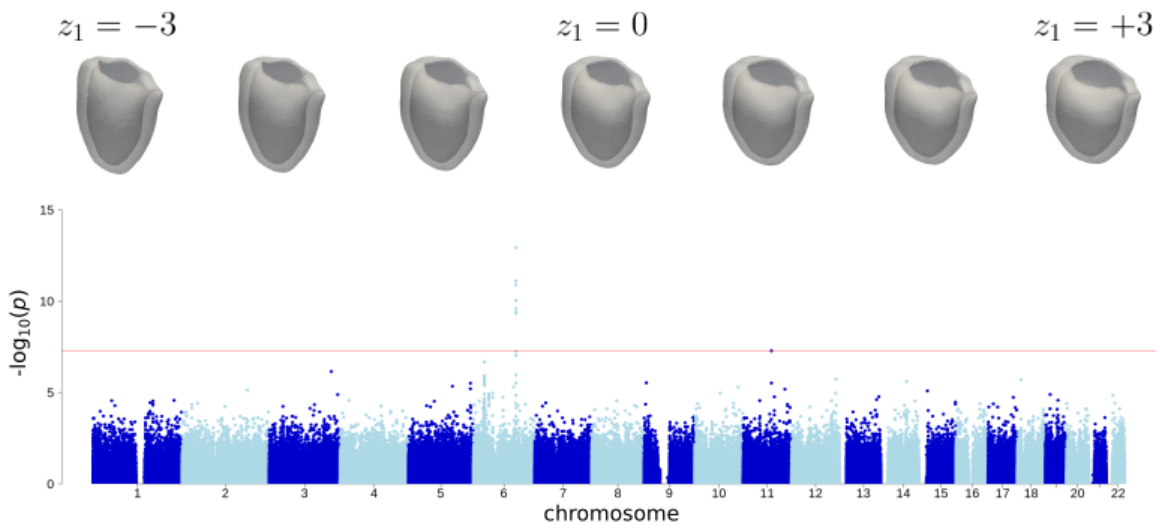

Figure 6.7: Effect of latent variable z_1 (before fine-tuning) and corresponding Manhattan plot.

$$\mathcal{L}_2 = \mathcal{L}_1 + w_{\text{SNP}}\mathcal{L}_{\text{SNP}} \tag{6.2}$$

$$\mathcal{L}_{\text{SNP}} = -\sum_k \sum_{l_k \in S_k} \text{Corr}(X_{l_k}, z_k) \tag{6.3}$$

Since this study constituted a proof of concept, only 4 latent variables were used in the bottleneck of the mesh autoencoder. 5,000 subjects were used for both the first training stage and the second fine-tuning stage. One latent variable, z_1 was found to be linked to the *PLN* locus (with

lead SNP rs11153730), as seen in Fig. 6.7. This variable was used for further fine-tuning on the second stage, with a single term on the summation from Eq. 6.3. By conducting this procedure repeated times with different values of w_{SNP}, we can assess whether it is effective in producing a better representation by computing the association p-value on the remaining 24,000 subjects. Fig. 6.8 shows that, while the effect is modest, this is the case. On the other hand, Fig. 6.9 shows the change in the effect of z_1 after the fine-tuning stage, for a particular run.

Data: 3D meshes $\mathbf{S}_i$ and linked genotype dosages X_{il}
Result: Network weights, GWAS summary statistics.
Hyperparameters: network architecture, $w_{\text{KL}} > 0$, $w_{\text{SNP}} > 0$.
`ProcrustesAlignment(`$\boldsymbol{S}_i$`);`
`PartitionDataset(`$\boldsymbol{S}_i$`,` X_{il}`);`
`InitialiseWeights();`
while *Stop criterion is not met* **do**
  Perform optimisation step with loss $\mathcal{L}_1$;
end
Select the best epoch within the validation set;
for *each latent variable* k **do**
  Perform GWAS within the training set;
  Extract set of genetic markers S_k based on a significance criterion;
  for *each SNP* l_k *in* S_k **do**
    if $Corr(X_{l_k}, z_k) < 0$ **then**
      $X_{l_k} \mapsto 2 - X_{l_k}$
    end
  end
end
while *Stop criterion is not met* **do**
  Perform optimisation step with loss $\mathcal{L}_2$;
end
Select the best epoch within the validation set;
Perform GWAS within the held-out set;

Algorithm 1: Workflow of the proposed method genotype-inform phenotyping method.

Conclusions. To the best of our knowledge, this is the first study that has shown systematically that it is possible to refine a phenotype by incorporating a term in the loss function inducing an alignment between SNP dosages and the learned phenotype ($\mathcal{L}_{\text{SNP}}$).

Importantly, note that it uses a somewhat heterodox data partitioning scheme: a relatively small subset of the data is used to *refine* the phenotype, whereas the rest is used to perform the actual GWAS, on the refined phenotype. The following paragraph suggests another potential use of this approach.

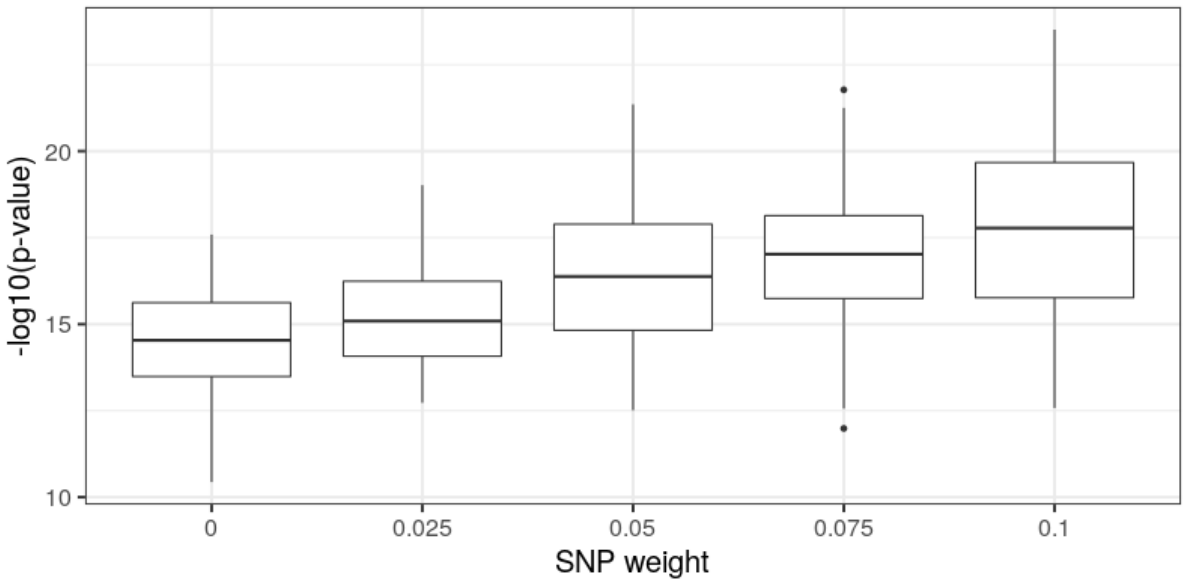

Figure 6.8: Box plot depicting the distribution of p-values for the z_1-rs11153730 associations after fine-tuning, for experiments with different values of w_{SNP} with $w_{\text{KL}} = 0.1$. Each box contains 60 ± 10 runs, where the 5,000 training subjects are changed each time, and the p-value is computed on the remaining 24,000 subjects.

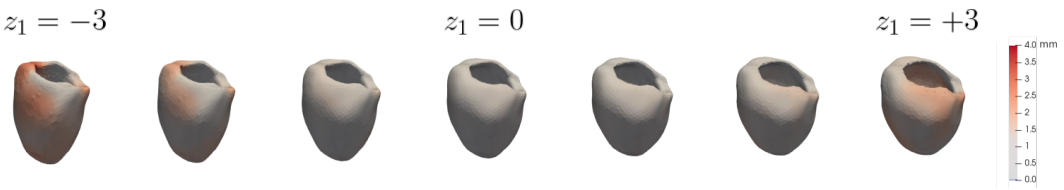

Figure 6.9: Morphologic effect of latent variable z_1, after fine-tuning. Colour represents the deviations in shape with respect to the meshes from Fig. 6.7.

An in-sample validation framework for genetic loci. We further argue that Figure 6.8 suggests this method could prove itself useful for *validating* genetic loci, *within* the discovery sample. In fact, note that the observed behaviour (that the association strength increases with increasing w_{SNP}) would only be present if the association is a true one. Therefore, a statistical test could be performed for the hypothesis of whether the procedure carried out in Figure 6.8 is effective in improving the strength of the genetic association, via a regression t-test. While we have not explored this avenue due to time constraints (note that it requires fine-tuning a great number of models with different w_{SNP} for each SNP), we believe it is a promising idea to validate genetic loci associated to autoencoder-derived phenotypes or, rather, to any neural-network-derived phenotypes.

6.7 Summary

This chapter has focussed on unsupervised methods for phenotyping static left-ventricular shapes at end-diastole (LVED) and its application for genome-wide association studies. The premise was that unsupervised phenotyping could lead to quantifying aspects of phenotypic variation that had remained uncharacterised in previous studies, thereby missing the opportunity to detect a number of interesting genetic signals.

It is important to note that the role of many of the discovered genes and, in particular of genetic variations affecting those genes, is not known at this time. This is true not only of the present study but of GWAS in general. Our study provides a starting point for elucidating the role of these loci in cardiac morphology.

Limitations and challenges. One of the most conspicuous limitations of UPE as conducted in this work, is the procedure for declaring associations significant. We opted for a conservative approach, which is using a Bonferroni threshold by dividing the genome-wide threshold by K, where K is the number of phenotypes in the ensemble. This approach seems to lead to a tradeoff between the size of the ensemble and the ability to detect significant loci, however we note that UPE was still able to detect an important number of novel loci, even when choosing this conservative approach, with $K = 36$.

We leave the development of a more optimal, principled methodology for declaring significant associations as future work. However, we argue here that using UPE with a genome-wide significance threshold, followed in tandem by the procedure depicted in Fig. 6.8, provides a promising framework for validating the detected candidate genetic loci, thereby removing the necessity for a study-wide p-value threshold. As mentioned, this procedure is resource-intensive as it requires generating several fine-tuned models for each locus. For this reason, time constraints prevented me from conducting this procedure at scale among the UPE-discovered SNPs. However, it is a promising direction for future research.

6.8 Code and data availability

The code for this work is split into several repositories publicly available on GitHub. All of them are accessible through a main repository: `www.github.com/rbonazzola/CardiacUPE`.

The `www.github.com/rbonazzola/CardiacCOMA` repository, to which the previous points, is included as a Git submodule and contains the code for an implementation of the Chebyshev-based convolutional mesh autoencoders (CoMA), using PyTorch and PyTorch Lightning. The nature of the work in this chapter makes it essential to have a systematic way to track the different runs, and logging hyperparameters and metric values. This has been achieved with the library `MLflow`, version 1.25, through its Python API. This repository also contains code to perform shape PCA on the cardiac meshes, using the `scikit-learn` Python package. Finally, it contains instructions on how to reproduce the software environment necessary for training the networks and carrying out the inference (i.e. producing the latent variables that act as quantitative phenotypes in this work).

The second submodule, `www.github.com/rbonazzola/GWAS_pipeline`, contains the code to carry out pre-processing of genetic data for GWAS, GWAS execution, results visualization and downstream analysis. This repository is written in R and Python, and also contains bash scripts invoking standard GNU command-line tools. Additional tools required for this work are: `bgenie`, `qctool`, `flashpca` and `plink`.

Data for performing the GWAS in this work comes in its integrity from the UK Biobank. UK Biobank Accession code for this application was 11350. Individual-level data are protected and therefore need to be downloaded from the UK Biobank.

Publicly available datasets used for GWAS downstream analyses have been queried for this work: the Ensembl Biomart database (www.ensembl.org), the IEU Open GWAS Project (`gwas.mrcieu.ac.uk`) for GWAS summary statistics, g:Profiler (`biit.cs.ut.ee/gprofiler`) for gene ontology terms, and `predictdb.org` for GTEx-based prediction models and SNP covariance matrices needed to run S-PrediXcan. In all cases, the date of last access was August 12, 2023.

For comparison, GWAS summary statistics were downloaded from `http://ftp.ebi.ac.uk` using the following study accession codes: GCST009393 through GCST009397 for [9], GCST010125 through GCST010131 for [78], GCST90000287 through GCST90000295 for [67] and GCST90162626 for [57].

Relevant data for this study has been uploaded to Zenodo: network weights for the ensemble of 36 autoencoders and the GWAS summary statistics for the traditional indices (LVEDV, LVEDSph, LVM and LVMVR) and for the first 16 shape PCs.

A web application has been developed on which researchers can access detailed results derived from this work. Instructions on how to connect to this application can be found at `www.github.com/rbonazzola/CardiacUPE`. We encourage readers of this thesis to consult this web application.

Chapter 7

Unsupervised dynamic CMR-derived phenotypes

In this chapter, we expand on our previous study to now investigate CMR-derived *dynamic* phenotypes. Namely, we aim to derive a compact representation of cardiac dynamics from the population of meshes. We also extend our study to the four cardiac chambers, as opposed to only the left ventricle.

As we will see, we will also require that it disentangles the anatomical aspects from those related to patterns of cardiac contraction and relaxation. We will refer to these in more general terms as *content* and *style*. We hypothesize that disentanglement of content and style factors in the representation of the beating heart will be helpful to perform the genetic association studies.

The work in this chapter is unpublished at the time of this writing, however a journal article is currently in preparation.

The chapter is composed of the following sections:

- **Section 7.1** provides motivation for the study of unsupervised dynamic phenotypes derived from CMR images.
- **Section 7.2** describes the phenotyping methodology followed in this paper, which includes defining a new neural network architecture.
- **Section 7.3** describes a synthetic dataset that has been built in order to efficiently test our dynamic phenotyping methodology.

- **Section 7.4** implements the algorithm on the real cardiac dataset, and discusses the genetic findings in light of prior knowledge on image- and ECG-derived phenotypes.
- **Section 7.5** provides a discussion on the findings.
- **Section 7.6** concludes the chapter by summarising the findings and limitations.

Note on nomenclature: We will employ the terms "cardiac anatomy", "anatomical features" and "content" interchangeably. Likewise, we will use "patterns of cardiac contraction or relaxation", "dynamical features", and "style" interchangeably.

7.1 Motivation

The motion dynamics of the beating heart are a complex rhythmic pattern of non-linear trajectories regulated by molecular, electrical and biophysical processes. Diseases affecting the heart's movement can have significant implications for overall health. Conditions like arrhythmias (irregular heartbeats), cardiomyopathies (diseases of the heart muscle), heart failure, and others can impact the heart's ability to pump effectively, leading to symptoms ranging from mild to life-threatening. Therefore, it is crucial to have quantitative metrics to characterise cardiac motion.

Perhaps the simplest scalar used to quantify contractility is the ejection fraction (EF) of a chamber (LV, RV, LA or RA), defined as the ratio between the stroke volume and the maximum cavity volume of the chamber; for example, for LV,

$$\text{LVEF} = \frac{\text{LVEDV} - \text{LVESV}}{\text{LVEDV}}, \tag{7.1}$$

[38]. Normal values for LVEF lie in the range 50-65%. However, as pointed out in [47], the broad usage of this quantity in clinical practice stems simply from the fact that it is easily determined from echocardiographic scans.

It is common that a patient with heart failure presents with normal EF. This phenomenon is termed *heart failure with preserved ejection fraction* (HFpEF) or *diastolic heart failure*. This condition can arise if myocardial tissue stiffens, which may reduce the amount the blood entering the ventricle, and the amount of ejected blood can also be reduced, leading to a normal level of

ejection fraction.

More detailed information can be obtained by examining other global indices such as the longitudinal, radial and circumferential relative displacements [104]. The LV global longitudinal strain (GLS) has proven more valuable than LVEF as a biomarker prognostic of heart failure, because at an early stage GLS tends to become compensated by circumferential and radial displacements to keep the LVEF within the normal range [95]. The local counterparts of these coefficients are also defined. This can be done for each vertex of an atlas (typically consisting of tens or hundreds of thousands of vertices) or for each of the 17 AHA segments (by taking the segment-wise average value).

In this work, following the spirit of our previous chapter, we turn to data-driven dynamic phenotypes with the hypothesis that this approach will capture previously uncharacterised patterns of motion, which in turn would lead to novel genetic discoveries.

7.2 Methods: disentangled spatiotemporal representation of cardiac motion

We wish to find an efficient low-dimensional representation that condenses the information carried by the population of meshes $\{\mathbf{s}_i(t)\}_{i=1}^N$, $t \in \mathcal{T}$ (where $\mathcal{T}$ is the temporal domain), but which also disentangles the anatomical aspects from those related to patterns of cardiac contraction and relaxation. We will refer to these in more general terms as *content* (for anatomical features) and *style* (for dynamic patterns).

7.2.1 Overview

For this, we again consider the encoder-decoder paradigm. In such paradigm, functions $E \in \mathcal{E}$ and $D \in \mathcal{D}$ are sought such that a given loss function $\mathcal{L}$ is minimised on average over the training set, with $\mathcal{L}(\mathbf{s}, \hat{\mathbf{s}})$ quantifying the deviation from an output mesh $\hat{\mathbf{s}} := D(E(\mathbf{s}))$ with respect to an input mesh M. Here, $\mathcal{E} \subset \{f : \mathbb{R}^{3M \times T} \to \mathbb{R}^K\}$ and $\mathcal{D} \subset \{g : \mathbb{R}^K \times \mathcal{T} \to \mathbb{R}^{3M}\}$ are some appropriate families of functions (T is the number of time frames in the dataset).

Periodicity Under steady state conditions, cardiac movement is nearly periodic. Cine-CMR data acquisition is performed by averaging of the signal at a corresponding phase (as determined by the ECG) across a number of cycles and under multiple breath-holds. The resulting image

can be thought of as an approximation of the expected value of the myocardial trajectory across one cardiac cycle, taken over a large number of cardiac cycles (where the scanned subject remains in a resting state), i.e. the reconstructed image does not contain more than a single cardiac cycle albeit being fed with information from several cardiac cycles. For this reason, it is convenient to think of the topology of the temporal domain $\mathcal{T}$ as periodical.

Generative Process

Let us first consider the process of generating a cardiac mesh at a time frame $\tau \in [0, 1)$ of the cardiac cycle. We will assume there is a set of latent variables $\mathbf{z}_c$ that explain the variability in content, and a set of latent variables $\mathbf{z}_s$ that explain the variability in style. We can think of this generative procedure as drawing $\mathbf{z}_c$ and $\mathbf{z}_s$ from their joint distribution (note that we do not assume them to be independent):

$$\mathbf{z}_c, \mathbf{z}_s \sim p(\mathbf{z}_c, \mathbf{z}_s) \tag{7.2}$$

Then, we assume steady state so that the movement is perfectly periodic. In this case, we can use a "periodic time", $e^{i2\pi t}$ with $t \in [0, 1)$, as a conditional variable:

$$\mathbf{s}(t) \sim p(\mathbf{s}|\mathbf{z}_c, \mathbf{z}_s; e^{i2\pi t}) \tag{7.3}$$

7.2.2 Proposed neural network

As depicted in Figure 7.1, the network we proposed is comprised of three main branches: a shared encoder, a content decoder and a style decoder.

Shared encoder. To process the sequence of meshes $[\mathbf{s}(t)]$ we can either 1) train a different 3D-mesh encoder for each phase or 2) train the same 3D-mesh encoder, i.e. share weights across time points.

In this work, we explore the second approach, in which we see two advantages. Firstly, the network is expected to be more economical in terms of complexity; namely, if N_W^ε is the number of weights needed to guarantee a given error ε on a dataset with T time points per subject, using a different encoder per time point, we hypothesise that the number of weights for the

shared encoder approach will be significantly fewer than $T \times N_W^{\varepsilon}$. The intuition behind this is that the features that explain shape variability are to be largely shared across different time points. This intuition has already been exploited by [54] in the CineNet approach, but with a different purpose in mind; namely, to improve the acquisition of CMR images. By means of backpropagation, and using a loss that penalizes reconstruction error at each time point, the learned hidden features will be such that they best capture the dynamics of the shape.

As with the processing of static 3D meshes from the previous chapter, the shared encoder, $E_{\mathbf{w}}$, is composed of a stack of four Chebyshev graph-convolutional layers with interspersed quadric pooling operations. $E_{\mathbf{w}}$ is applied to each element of the sequence of meshes $\{\mathbf{s}_i(t)\}$ for subject i, one at a time, to produce a trajectory of hidden variables over the cardiac cycle, $\{\mathbf{h}_i(t)\}$:

$$\mathbf{h}_i(t) = E_{\mathbf{w}}(\mathbf{s}_i(t)), \tag{7.4}$$

Temporal aggregation. The process of temporal aggregation $T_{\text{AGG}}(\cdot)$ is utilised on the sequence of hidden variables $\mathbf{h}_i(t)$ to generate a latent vector $\mathbf{z}_i$ that characterizes the entire sequence rather than being dependent on a particular t.

$$\mathbf{z}_i = T_{\text{AGG}}\Big(\big[\, \mathbf{h}_i(t_1) \,|\, \mathbf{h}_i(t_2) \,|\, ... \big] \Big), \tag{7.5}$$

with $t_1, t_2, ... \in \mathcal{T}$

These features are node attributes of a (downsampled) mesh since they are generated by a stack of interleaved graph-convolutional and graph-pooling operations. Therefore, this aggregation operation allows us to "mix" features at different spatial and temporal locations, producing new features $\mathbf{z}_i$ that are global both in space and in time.

Our proposal of T_{AGG} for the purposes of this work is to concatenate the features for different time points into a single vector and pass it through a stack of two fully connected layers (FCL). Although effective for our application, this approach limits the number of time points in the sequence, T, to be the same for all subjects.

Content and style disentanglement

For our case, we regard the content information contained in the collection of meshes as that which is necessary to be able to reconstruct a *content representative* of the 3D meshes across the cardiac cycle. (As we will see, this will be the time-averaged shape across the cycle.) To effectively disentangle content and style, and inspired by previous work disentangling shape from texture factors [92], we propose a network architecture in Figure 7.1, where the latent representation is partitioned into a content and a style component:

$$\left[\mathbf{z}_c^{(i)} | \, \mathbf{z}_s^{(i)}\right] = \mathbf{z}_i \tag{7.6}$$

Phase information is embedded into the style latent vector by multiplying each of its components by a unit complex vector, where its phase is given by the phase of the mesh across the cardiac cycle (with $t = 0 \equiv 1$ being end-diastole):

$$\check{\mathbf{z}}_s^{(i)}(t) = \mathbf{z}_s^{(i)} e^{i2\pi t}. \tag{7.7}$$

Finally, the latent vector is passed through the decoder: the content partition is passed through the content decoder $\mathcal{D}_c$ to obtain the content representative, or static component, $\hat{\mathbf{s}}_c^{(i)}$, whereas the *whole* latent vector, i.e. both its content and phased style partitions, is passed through the style decoder $\mathcal{D}_s$ to produce the dynamic component. Together, they generate the reconstructed shape at phase t, $\hat{\mathbf{s}}_i(t)$:

$$\hat{\mathbf{s}}_i(t) = \underbrace{\mathcal{D}_c(\mathbf{z}_c^{(i)})}_{\hat{\mathbf{s}}_c^{(i)}} + \underbrace{\mathcal{D}_s(\mathbf{z}_c^{(i)}, \check{\mathbf{z}}_s^{(i)}(t))}_{\hat{\mathbf{s}}_i(t) - \hat{\mathbf{s}}_c^{(i)}} \tag{7.8}$$

We train the network by optimizing with respect to the following loss function:

$$\mathcal{L}_i = \sum_{t \in \mathcal{T}_i} \mathcal{L}_{\text{rec}}\Big(\mathbf{s}_i(t), \hat{\mathbf{s}}_i(t)\Big) + w_c \mathcal{L}_{\text{rec}}(\mathbf{s}_c^{(i)}, \hat{\mathbf{s}}_c^{(i)}) \tag{7.9}$$

Choice of the content representative. Note that the choice of the so-called static component, $\mathbf{s}_{i,c}$, has not been specified so far. It stands for some representative of the content information for subject i. Several plausible candidates include: $\mathbf{s}_i(0)$ (shape at end-diastole),

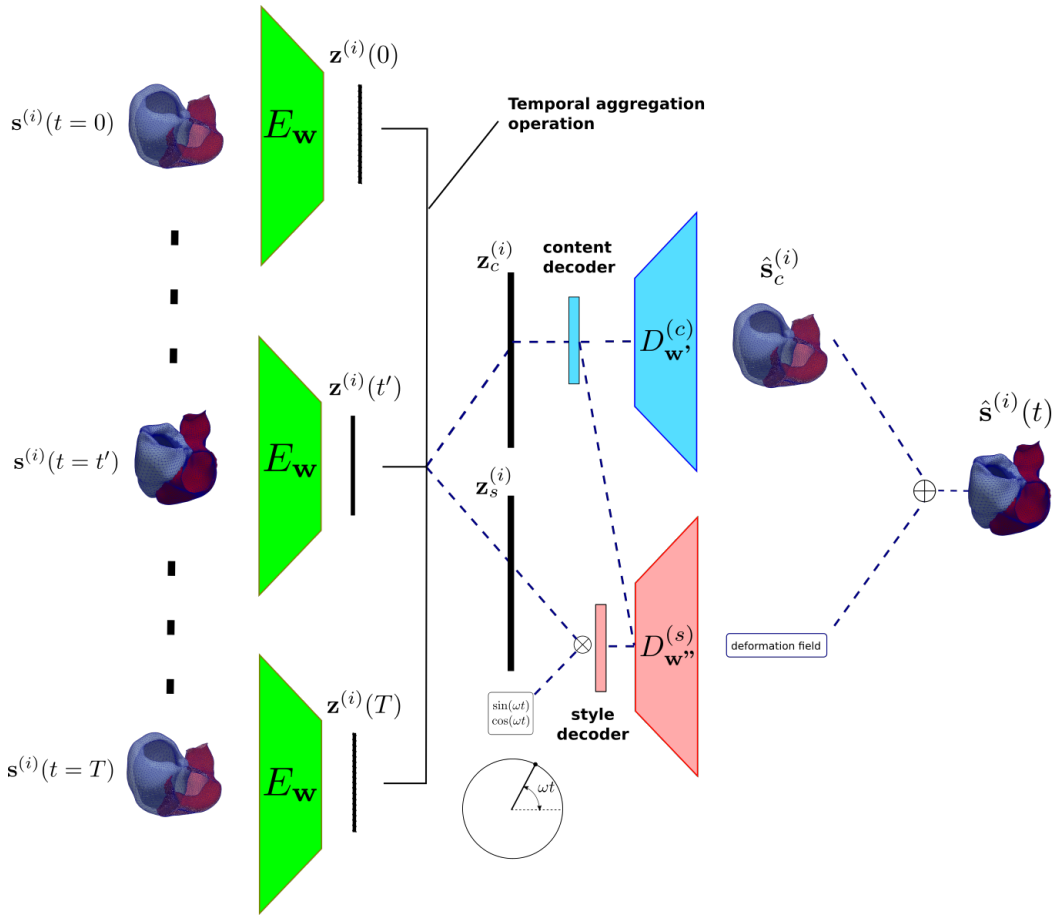

Figure 7.1: Architecture of the proposed spatio-temporal neural network: it consists of a series of weight-sharing encoders $E_{\mathbf{w}}$ that process the sequence of meshes one at a time, generating a hidden variable trajectory that is then condensed into a content and a style latent vector. The content part is passed through the content decoder $D_{\mathbf{w}'}^{(c)}$ to produce a temporally averaged mesh, and the style part is passed through the style decoder $D_{\mathbf{w}''}^{(s)}$ to generate a (pseudo-)deformation field, i.e. the vertex-wise deformation that has to be applied on the time-averaged shape $\langle \mathbf{s} \rangle$ to generate the shape at any given time point, $\mathbf{s}(t)$. We use the term "pseudo-deformation" since the content decoder does not necessarily output a real shape.

$\mathbf{s}_i\left(t_{\mathrm{ES}}^{(i)}\right)$ (shape at end-systole, where $t_{\mathrm{ES}}^{(i)} \approx \mathrm{argmin}_t \mathrm{BVV}_i(t)$), the Euclidean or Fréchet mean of the shape across the cardiac cycle. We have observed that this choice leads to detecting different loci, and therefore we decided to use it as an additional hyperparameter to thus enhance the diversity of our ensembles. However, in this chapter we will present results only for the Euclidean mean shape as content shape representative, $\langle \mathbf{s}_i \rangle$. The reason is that, due to time constraints, it was not possible to train large enough ensembles for each content representative to draw meaningful conclusions regarding the effect of using different values for this hyperparameter.

7.2.3 Network implementation

The network was implemented using `PyTorch 1.9` with `Pytorch Lightning 1.8`. As with the previous network, runs were tracked using `MLflow 1.25`.

7.3 Synthetic dataset

To test different variations of our algorithm, we built a toy dataset consisting of meshes with time-dependent periodical perturbations. This allowed us to control the size and complexity of the dataset at will, thus reducing the execution time if needed. Furthermore, the latent factors that originate each sample are known (unlike the real scenario), which facilitates testing the encoder and decoder separately in a supervised way.

As our reference shape, we chose a sphere (the shape's *population mean*, $\mathbb{E}\big(\langle \mathbf{S} \rangle\big)$), where $\mathbf{S} \in \mathbb{R}^{M\times 3}$ is the shape taken as a random variable, on which we superimposed a radial static deformation $\delta_c^{(i)}$ (of null mean), as well as a periodic dynamic deformation $\delta_s^{(i)}(t)$ (of both temporal and population null mean):

$$\mathbf{s}_i(t) = \underbrace{\mathbb{E}\big(\langle \mathbf{S} \rangle\big)}_{\text{sphere}} + \underbrace{\Big[\langle \mathbf{s}_i \rangle - \mathbb{E}\big(\langle \mathbf{S} \rangle\big)\Big]}_{\delta_c^{(i)}} + \underbrace{\Big[\mathbf{s}_i(t) - \langle \mathbf{s}_i \rangle\Big]}_{\delta_s^{(i)}(t)} \tag{7.10}$$

The total deformation $\delta_i(t)$, which defines "subject" i, is thus given by the sum of the static and the dynamic deformation at a given time point t:

$$\delta_i(t) = \delta_c^{(i)} + \delta_s^{(i)}(t) \in \mathbb{R}^{M\times 3}$$

7.3.1 Spherical harmonics

We parameterised the variations in terms of spherical harmonics $Y_{lm}(\theta, \phi)$. Here, θ and ϕ represent the spherical angles: θ is the azimuthal angle and ϕ is the polar angle (see Figure 7.2). Spherical harmonics satisfy the following equation:

$$\nabla^2 Y_{lm}(\theta, \phi) = l(l+1) Y_{lm}(\theta, \phi), \tag{7.11}$$

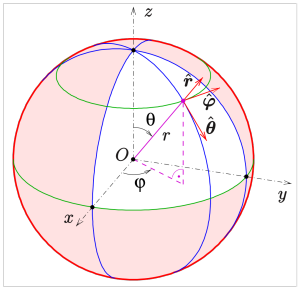

Figure 7.2: Spherical coordinates (from `https://en.wikipedia.org/wiki/Spherical_coordinate_system`)

i.e. they are eigenfunctions of the Laplacian on the sphere with eigenvalue $l(l+1)$. They are $(2l+1)$-degenerate for each distinct eigenvalue. The explicit form of these functions is given by:

$$Y_{lm}(\theta, \phi) = N_{lm} P_l^m(\cos\theta) e^{im\phi}, \tag{7.12}$$

where P_l^m are the associated Legendre polynomials and $N_{lm} = \sqrt{\frac{(2l+1)}{4\pi}\frac{(l+m)!}{(l-m)!}}$ is a normalization constant, that ensures that the polynomials constitute an orthonormal basis, that is

$$\int_0^{\pi}\int_0^{2\pi} Y_{lm}(\theta,\phi) Y^*_{l'm'}(\theta,\phi) d\theta d\phi = \delta_{ll'}\delta_{mm'},$$

where δ_{ij} is the Kronecker delta.

7.3.2 Construction of the dataset

The deformation with respect to the reference spherical shape is performed along the radial direction at each location, as a linear combination of the Y_{lm} with $l = 0, ..., l_{\max}$ and $m = -l, -l+1, ..., l$. We encoded static as well as dynamic features. Time-periodic coefficients are generated for each Y_{lm}. To make the deformation periodic in time, we build the coefficients as a linear combination of sines and cosines functions, with frequencies n that are multiples of the fundamental frequency, taken to be the unit:

$$a_{lm}(t) = b_{lm}^{(0)} + \sum_{n=1}^{n_{\max}} \left[a_{lm}^{(n)} \sin(2\pi n t) + b_{lm}^{(n)} \cos(2\pi n t) \right]$$

Finally, the shape at time t is given by the following:

$$\mathbf{d}(\theta, \phi, t) = \left(\sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} a_{lm}(t) Y_{lm}(\theta, \phi) \right) \hat{\mathbf{r}}(\theta, \phi)$$

where $\hat{\mathbf{r}}(\theta, \phi)$ is a unit vector along the radial direction (i.e. the deformations are performed along the radial direction). The generative model used was the following.

$$b_{lm}^{(0)} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_c^2)$$

$$a_{lm}^{(n)}, b_{lm}^{(n)} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_s^2), n \geq 1$$

Given that these parameters directly generate the shapes and are statistically independent, they can be thought of as the *real* latent factors.

Complexity of the mesh population. The population is completely defined by N, $l_{\max}$, σ_c^2, σ_s^2 and a random seed s. The complexity of the population of meshes is given by the number of latent factors C, which in turn is determined by $l_{\max}$ and $n_{\max}$; in particular, $C = (l_{\max}+1)^2(n_{\max}+1)$, from which $C_c = (l_{\max}+1)^2$ correspond to content and $C_s = (l_{\max}+1)^2 n_{\max}$ correspond to style. For example, with $l_{\max} = n_{\max} = 2$, we get $C = C_c + C_s = 9 + 18 = 27$.

Meshes are sampled from the population at fixed time points $t \in [0, 1)$.

7.3.3 Representation learning

In this subsection, we present some conclusions drawn from the study of the synthetic dataset.

Reconstruction performance

First, we briefly introduce the metrics utilised to assess the reconstruction performance of the network. We define a set of *ad hoc* normalized metrics that quantify deviations from the ground truth, in such way that a value below 1 means that the network's reconstruction performance is better than predicting the mean shape, whereas values close to zero represent near-perfect reconstruction.

Metric	Definition	Interpretation
ε_c	$\frac{\sum_{i=1}^{N_{\text{test}}} \left\|\left\|\delta_c^{(i)} - \hat{\delta}_c^{(i)}\right\|\right\|_2^2}{\sum_{i=1}^{N_{\text{test}}} \left\|\left\|\delta_c^{(i)}\right\|\right\|_2^2}$	Static variance failed to be captured by the model, as a fraction of the static variance in the population
$\varepsilon_s(t)$	$\frac{\sum_{i=1}^{N_{\text{test}}} \left\|\left\|\delta_s^{(i)}(t) - \hat{\delta}_s^{(i)}(t)\right\|\right\|_2^2}{\sum_{i=1}^{N_{\text{test}}} \left\|\left\|\delta_s^{(i)}(t)\right\|\right\|_2^2}$	Dynamic variance failed to be captured by the model, as a fraction of the dynamic variance in the population

Table 7.1: Normalised metrics used to evaluate reconstruction performance.

Some architectural conclusions

By testing variants of our algorithm on the toy dataset, we concluded that our approach with a fully convolutional shared encoder outperforms a shared encoder with an additional fully connected layer, presumably because this layer breaks the graph structure. For this reason, the runs with the real dataset are performed using this architecture.

Can the real latent factors be retrieved?

We then investigated whether the latent factors learned by our neural network, $\hat{\mathbf{z}}$, can be identified with the coefficients $a_{lm}^{(n)}$ and $b_{lm}^{(n)}$, which are the real latent factors. Given that these factor are well-defined only up to a linear transformation, we applied canonical correlation analysis (CCA) to obtain the best linear combination of the predicted latent factors $A\hat{\mathbf{z}}$.

We concluded that these factors can be retrieved with very high fidelity (see Figure 7.3), and that real static factors map to linear combinations of predicted static factors, and likewise for dynamic factors.

These results support the validity of our disentangling architecture and justify its application to real cardiac data, where the true generative factors are unknown.

7.4 Representation learning on dynamic cardiac meshes

In this subsection, I describe the results of our genetic association studies on the static and dynamic embeddings obtained through ensembles of models trained on sequences of cardiac meshes.

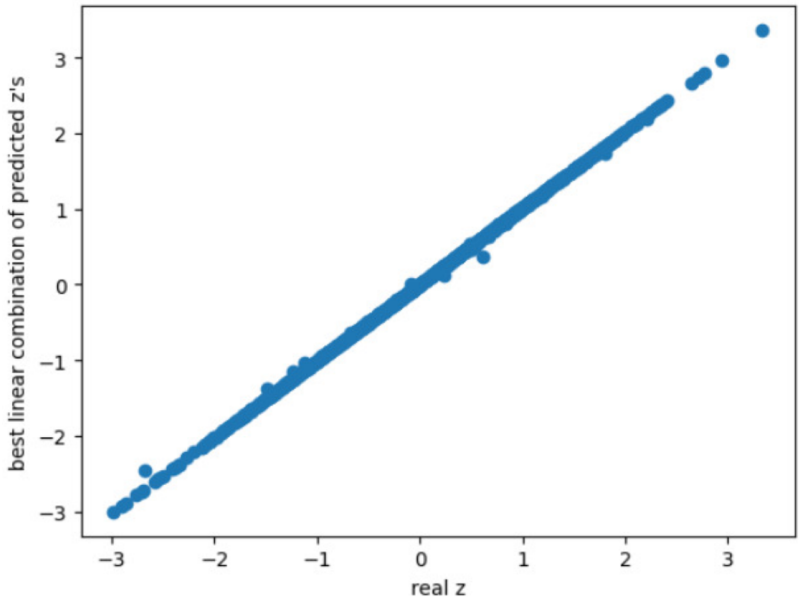

Figure 7.3: Comparison of real latent factors vs. best linear combination of predicted latent factors, from CCA. Values are predicted with almost perfect fidelity.

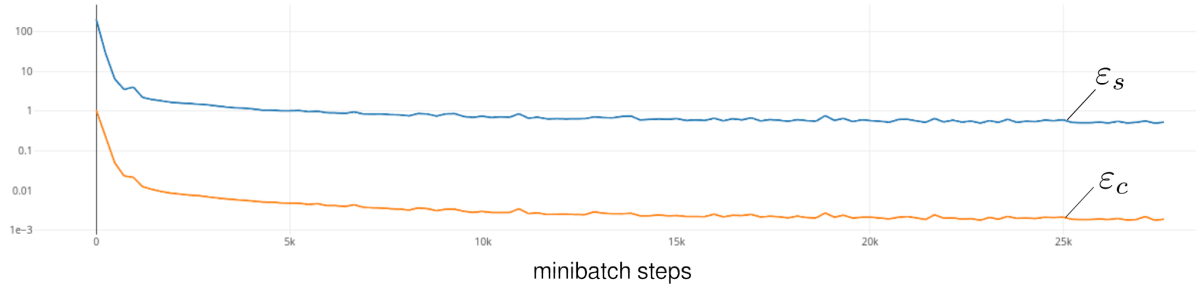

Figure 7.4: Normalised metrics curves (for ε_c and ε_s) as a function of the number of minibatches.

7.4.1 Description of the experiments

Following our previous work on unsupervised (static) LV phenotype ensembles from chapter 5, ensembles of spatio-temporal autoencoders were trained for each of the cardiac chambers: LV, RV, LA and RA. Also, LV and RV are considered jointly in a biventricular mesh (BV), constituting a fifth partition. Runs were selected as part of the ensemble based purely on whether they surpassed a threshold for reconstruction performance metrics.

We produced 5 runs per partition considered (totalling 25). The phenotypes were classified as either *static* or *dynamic* depending on the partition they belonged to (either $\mathbf{z}_c$ or $\mathbf{z}_s$, respectively). Each autoencoder produced 8 or 16 static components, and 8 or 16 dynamic components, totalling 16, 24 or 32 latent components per run.

In Figure 7.5 we display these curves for the $n_z = 8$ of a single run on LV.

7.4.2 Interpretation of the latent variables

To aid with the interpretation of the different dynamic latent variables, synthetic shape sequences are obtained by setting the latent vector $\mathbf{z} = (\mathbf{z}_c, \mathbf{z}_s) = (\mathbf{0}, \lambda \mathbf{e}_j)$ where $(\mathbf{e}_j)_i = \delta_{ij}$ and $\lambda = \{0, \pm 1, \pm 2, \pm 3\}$, and passing it through the decoder network after convolving $\mathbf{z}_s$ with $e^{i2\pi t}$ with $t \in [0, 1)$. This generates a set of synthetic cardiac sequences, and by examining the change in these sequences along the traversal from $\lambda = -3$ to $\lambda = 3$, the effect of the latent variable can be studied. As an exemplar of the full range of variability captured by the autoencoder, Figure 7.5 shows the LVV curves for the $n_z^s = 8$ traversals in one single run for LV.

In Figure 7.6, correlations with handcrafted phenotypes are displayed for the most strongly associated latent variables for each locus. In Figure 7.7, detailed per-AHA-segment correlations between latent variable and absolute and relative wall thickening are shown.

7.4.3 Genetic associations on unsupervised dynamic features

By examining the associations we found that, while some loci are linked to both static and dynamic variables, some of them are specific to either kind: some loci exist with study-wide significant associations for dynamic variables, which have no significant associations for any static variables; and the converse is also true. Based on this, we split our associations into three groups: 1) those predominantly associated to dynamic phenotypes, 2) those predominantly associated to static phenotypes and 3) those which have shown strong associations with both

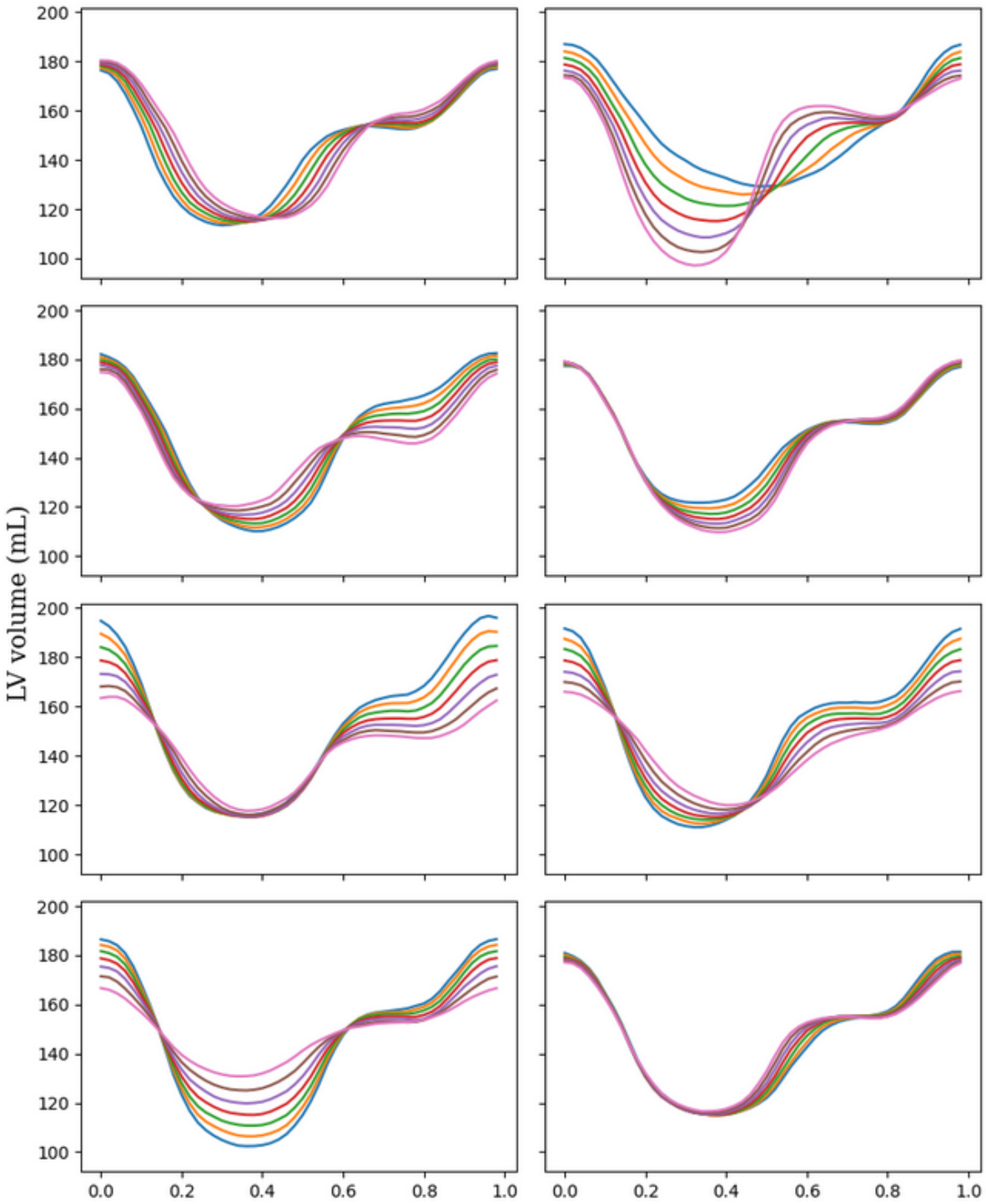

Figure 7.5: Synthetic LV volume curves obtained by varying each of the eight dynamic latent variables, one at a time, for one exemplar run with $n_z^{(s)} = 8$.

with no clear preference. These groups are comprised of 22, 12 and 40 loci, respectively. The findings for these three categories are depicted in Figure 7.8. This figure also shows how each cardiac chamber contributes to the GWAS hit counts.

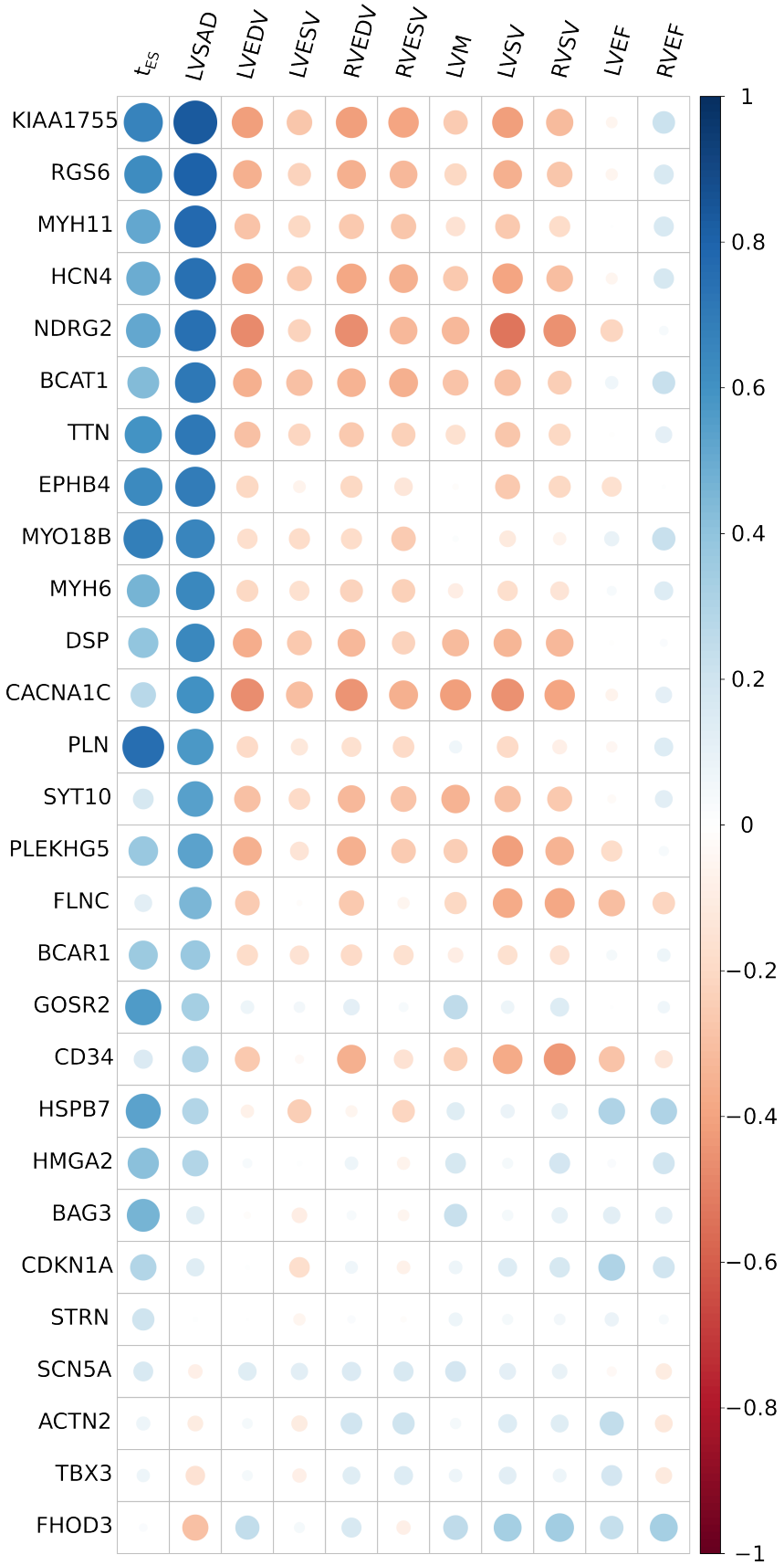

Figure 7.6: Correlation between best latent variable per locus and absolute and relative left-ventricular wall thickening (per AHA segment). The reader can refer to Fig. 5.1 to consult the location of each AHA segment.

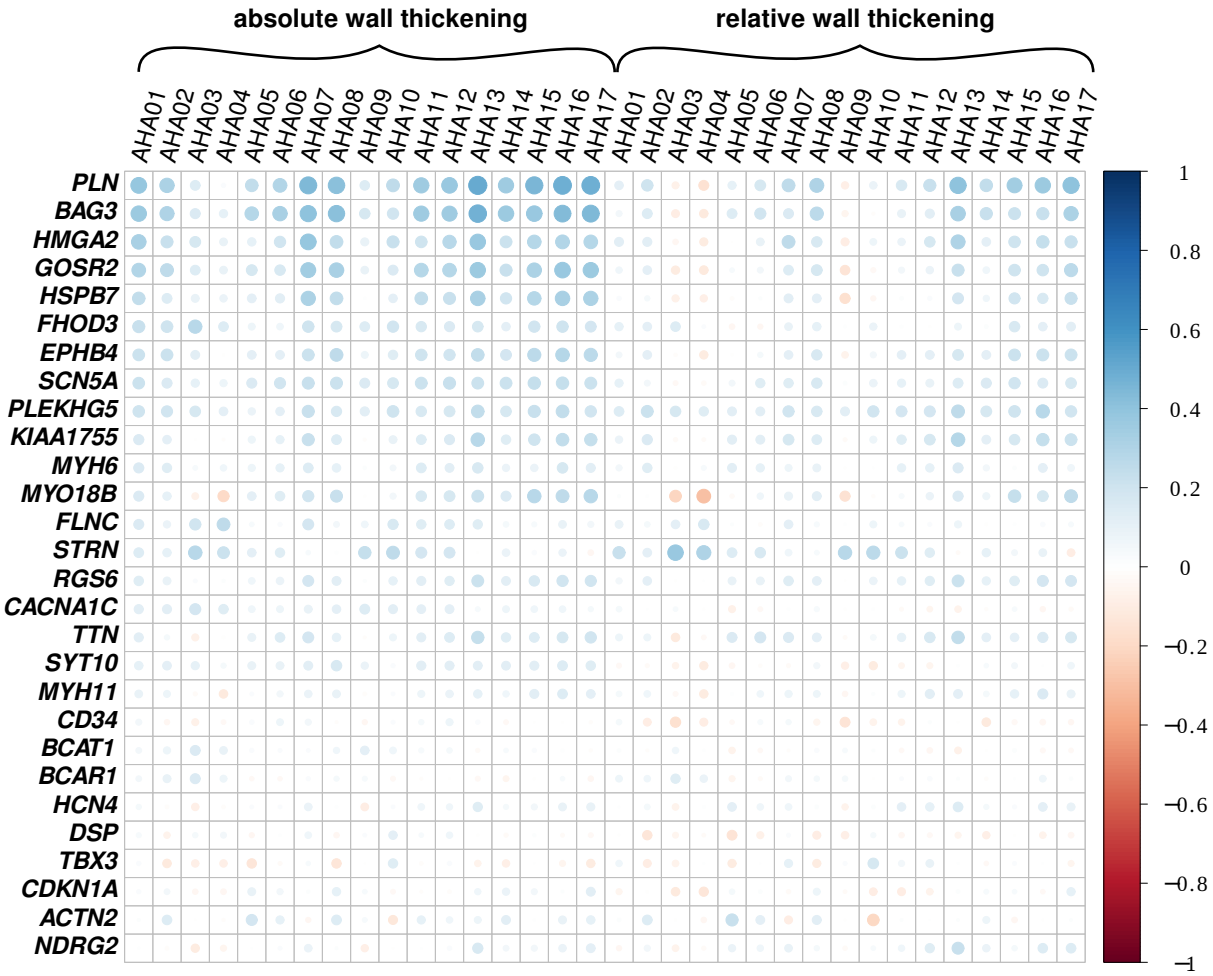

Figure 7.7: Correlation between best latent variable per locus and absolute and relative wall thickening per left-ventricular AHA segment.

Co-occurrence of GWAS associations across the ensemble We observed that certain pairs of loci were found simultaneously for a high number of latent variables. To quantify this phenomenon, we define the ensemble-wide GWAS hit matrix $H_{z,\ell} = \mathbf{1}_{p_{z,\ell}<p_{\text{GW}}}$. This is a binary matrix indicating whether a given locus, ℓ, contains a genome-wide significant SNP for the given phenotype z. The contingency matrix for pairs of loci is given by $C = H^t H$, where H_{ℓ_1,ℓ_2} is then the number of instances (phenotypes) for which loci ℓ_1 and ℓ_2 are both genome-wide significant. (Note that we are pooling all the chambers' latent variables together.) If we divide each row ℓ by the number of hits, $H_{\ell,\ell}$, we obtain a matrix of co-occurrence rate of pairs of loci, which is depicted in Figure 7.9. This matrix suggests the presence of a number of clusters of loci that are presumably linked to similar phenotypes.

Dynamic phenotypes. We first discuss genetic associations with predominantly dynamic variables. Table 7.3 shows the summary statistics for the study-wide significant loci that were

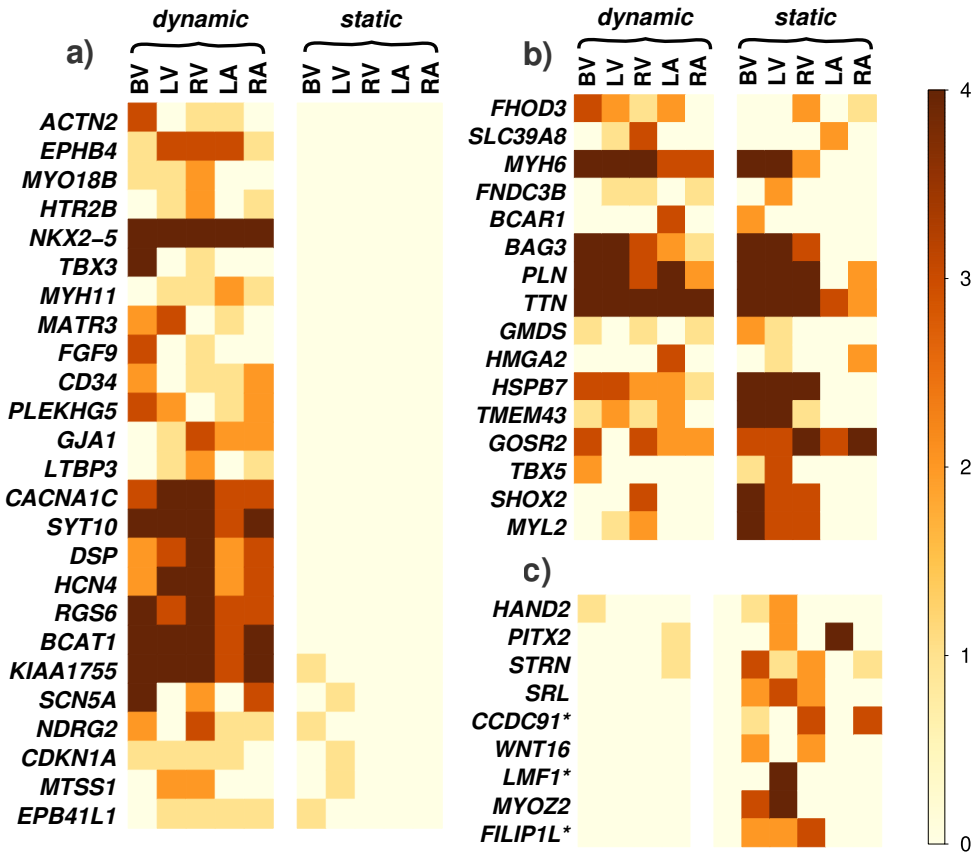

Figure 7.8: Number of runs C_ℓ ($0 \leq C_\ell \leq 5$) for which each locus was found with genome-wide significance, $p < 5 \times 10^{-8}$. Loci are classified into three groups according to whether a) they are predominantly linked to dynamic phenotypes, b) they are linked to both static and dynamic phenotypes with no clear preference or c) they are predominantly linked to static phenotypes. They are further classified by chamber to pinpoint chamber-specific loci.

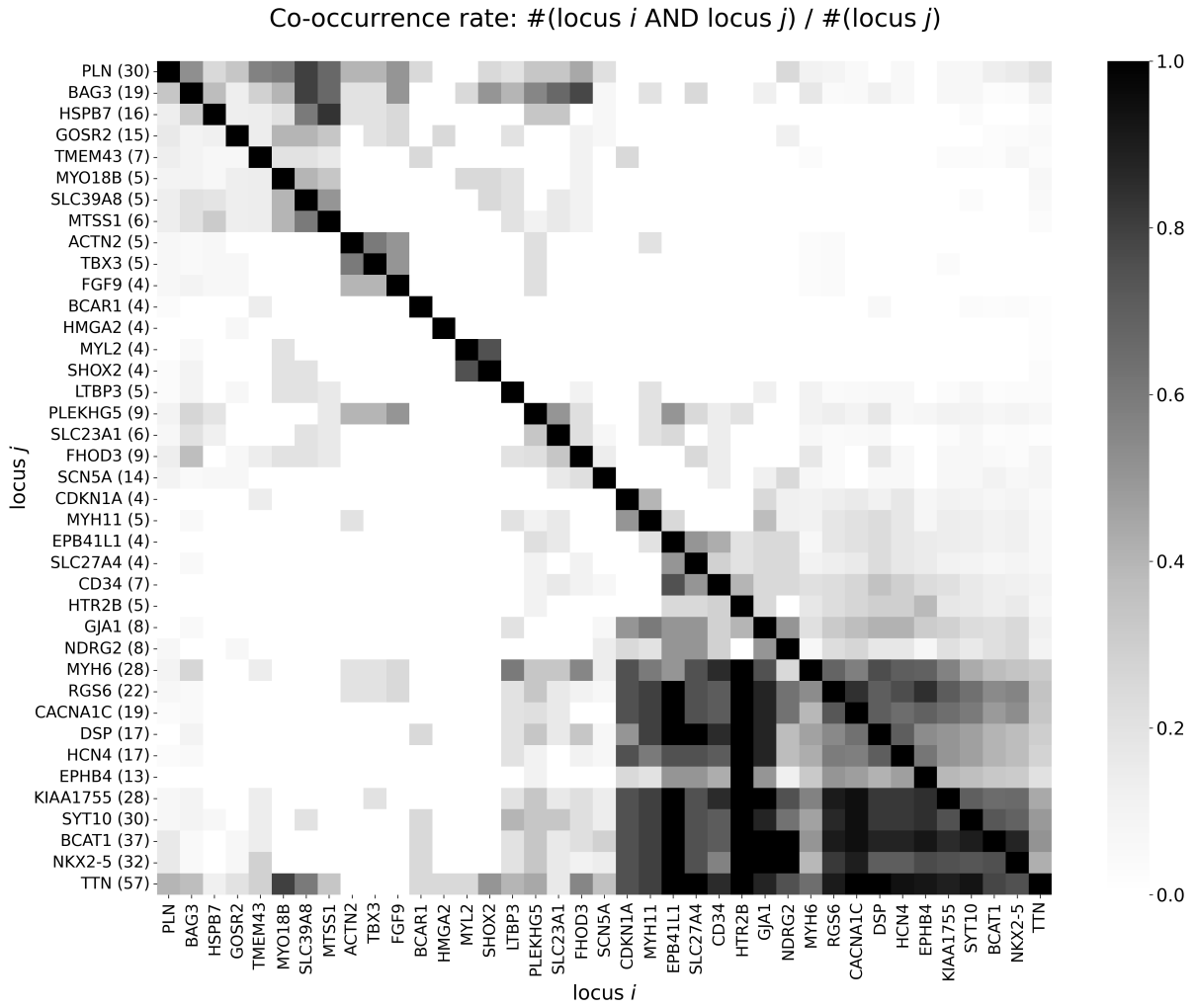

Figure 7.9: Heatmap depicting the GWAS hit *co-occurrence matrix* for dynamic variables: intensity of the gray colour indicates the fraction of times that the locus in the row is detected with $p < p_{\mathrm{GW}}$, among the dynamic variables for which the locus in the columns is also detected with the same threshold. The number between parentheses after the locus name indicates how many times this locus was discovered in total (if significant for several dynamic variables of the same run, each of these is a count). For instance, locus *PLN* is detected 30 times with genome-wide significance and, out of those, a fraction $f = 0.333$ of times, that is, 10 times, the locus *GOSR2* is also detected with the same level of significance.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
6	117672972-118963115	PLN	5	2.3×10^{-26}	rs72967533	T	C	47.7	5.29 ± 0.7
1	14891511-16897730	HSPB7	5	2.0×10^{-14}	rs1627145	C	T	65.7	5.13 ± 0.73
10	120591353-122407323	BAG3	5	6.5×10^{-14}	rs375034445	A	AT	21.2	3.97 ± 0.85
3	13070799-14816900	TMEM43	5	8.9×10^{-13}	rs900173	T	C	34.0	3.73 ± 0.74
16	60054-1207206	LMF1*	5	3.3×10^{-12}	rs12600110	T	C	38.5	-4.99 ± 0.72
2	178553183-181312739	TTN	5	2.5×10^{-19}	rs10497529	G	A	3.6	-7.78 ± 1.84
17	63148128-64800430	PRKCA	4	3.0×10^{-10}	rs7210446	G	A	58.1	3.27 ± 0.71
16	4001196-5118345	SRL	4	6.1×10^{-09}	rs62037566	C	T	23.2	4.23 ± 0.83
3	157312028-159477890	SHOX2	4	3.0×10^{-10}	rs6767008	C	G	49.8	-2.63 ± 0.7
17	43056905-45876022	GOSR2	4	4.6×10^{-10}	rs17608766	T	C	14.3	5.76 ± 0.98
12	110336719-113263518	MYL2	4	4.1×10^{-11}	rs35350651	A	AC	51.4	-2.77 ± 0.69
14	23018665-24905123	MYH6	4	4.0×10^{-11}	rs365990	A	G	36.9	4.22 ± 0.72
4	119933512-120392684	MYOZ2	4	2.4×10^{-11}	4:120323630_CT_C	CT	C	28.1	5.25 ± 0.79
3	170964909-172295731	FNDC3B	3	1.2×10^{-10}	rs4894803	A	G	40.4	3.67 ± 0.72
12	113986709-115036602	TBX5	3	1.3×10^{-09}	rs2555030	G	A	13.8	3.69 ± 1.02
3	72529329-74321817	PDZRN3	3	2.7×10^{-09}	3:73578036_AACACAC_A	AACACAC	A	39.8	-4.25 ± 0.71
3	99373762-100592217	FILIP1L*	3	6.2×10^{-09}	rs9811920	G	A	40.8	-4.04 ± 0.71
7	45952922-46986720	IGFBP3	3	1.5×10^{-08}	rs2881198	G	C	52.7	-3.97 ± 0.7
2	36122006-38132712	STRN	2	5.2×10^{-10}	rs2252032	T	C	52.7	-4.34 ± 0.7
4	111256567-113870102	PITX2	2	5.2×10^{-10}	rs976568	G	T	64.9	2.89 ± 0.74
1	226810860-229156248	OBSCN	2	2.5×10^{-09}	rs1684868	T	C	52.0	3.32 ± 0.7
12	65559695-67181144	HMGA2	2	3.4×10^{-09}	rs1979440	T	C	40.7	-4.19 ± 0.71
8	125683719-126410917	MTSS1	2	4.2×10^{-09}	rs12542527	A	G	28.1	-3.48 ± 0.78
4	174264132-176570716	HAND2	2	1.4×10^{-08}	rs4342170	T	A	97.5	-12.97 ± 2.29

Table 7.2: GWAS results for LV static variables.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
14	23018665-24905123	MYH6	5	1.4×10^{-23}	rs365990	A	G	36.9	4.22 ± 0.72
20	34960446-36909530	KIAA1755	5	7.2×10^{-15}	rs4811602	G	A	46.4	-3.35 ± 0.71
15	73628714-76398624	HCN4	5	2.2×10^{-10}	rs11630367	A	G	15.8	-4.83 ± 0.95
10	120591353-122407323	BAG3	5	8.7×10^{-19}	rs72842207	C	T	21.4	-3.26 ± 0.84
5	171074292-172678327	NKX2-5	5	4.9×10^{-14}	rs11134777	G	T	16.5	-6.56 ± 0.95
12	1080331-2544786	CACNA1C	5	1.2×10^{-11}	rs2283274	G	C	17.8	-4.32 ± 0.92
12	33076989-37856717	SYT10	5	9.8×10^{-16}	rs11052736	T	C	45.6	3.13 ± 0.7
12	23820634-25371083	BCAT1	5	1.7×10^{-16}	rs4963772	G	A	15.0	-5.9 ± 0.97
2	178553183-181312739	TTN	5	5.0×10^{-17}	rs17362588	G	A	8.7	8.61 ± 1.21
14	71131957-72889615	RGS6	4	1.9×10^{-12}	rs17180489	G	C	14.5	4.3 ± 0.93
1	14891511-16897730	HSPB7	4	3.9×10^{-12}	rs10927879	G	A	35.1	5.06 ± 0.73
6	117672972-118963115	PLN	4	1.9×10^{-11}	rs57912492	C	T	41.9	-2.9 ± 0.71
6	6785207-7808936	DSP	4	2.9×10^{-09}	rs72825054	G	C	9.7	7.15 ± 1.18
7	100196651-101199253	EPHB4	4	2.8×10^{-10}	rs144170516	A	AAAAT	26.2	-3.81 ± 0.8
5	136376050-139265072	SLC23A1	3	7.2×10^{-10}	rs11741938	A	G	66.8	4.59 ± 0.74
8	125683719-126410917	MTSS1	3	5.1×10^{-11}	8:125858538_GA_G	GA	G	31.0	3.07 ± 0.76
2	231843389-233550003	HTR2B	2	1.5×10^{-08}	rs13001643	G	A	22.4	-4.02 ± 0.84
3	13070799-14816900	TMEM43	2	1.3×10^{-08}	rs7612736	G	A	21.2	4.86 ± 0.86
18	33861964-35075250	FHOD3	2	6.6×10^{-11}	rs2644262	T	C	27.5	4.67 ± 0.78
22	23712647-24984204	SMARCB1	2	5.7×10^{-14}	rs5760032	C	T	79.5	-3.6 ± 0.87
1	5913893-7247335	PLEKHG5	2	2.4×10^{-08}	rs709208	A	G	32.2	-3.62 ± 0.76
6	35455756-37572596	CDKN1A	1	1.1×10^{-10}	rs146170154	C	CTA	19.9	-3.66 ± 0.88
12	65559695-67181144	HMGA2	1	9.5×10^{-12}	rs761210718	AAG	A	36.9	-3.14 ± 0.72

Table 7.3: GWAS results for LV dynamic variables.

found for LV. Similar tables for the rest of the chambers and for BV can be found in the Appendix.

SNP in the *KIAA1755* locus had been identified with suggestive significance (best $p = 2.3 \times 10^{-10}$) in our investigation of static LV phenotypes from the previous chapter. Interestingly, in the present study we identify it as related to dynamic phenotypes with a greater effect size. Furthermore, it has a remarkably greater presence among LV phenotypes of the ensemble. *FLNC* encodes filamin C, a protein that plays a crucial role in the structure and function of heart muscle cells. It is an actin cross-linking protein and its main role is maintenance of the structural integrity of the sarcomere by binding to several proteins in the Z-disk. The *GJA1* gene, which encodes a gap junction protein known as connexin 43 (Cx43), is associated with various aspects of heart function. *MTSS1* encodes for a cytoskeletal protein. *MYO18B* encodes for myosin XVIIIB. CN4 is a protein that is encoded by the *HCN4* gene and is prominently expressed in the pacemaker region of the mammalian heart. This protein play an important role in heart rate regulation. The *MYH11* gene encodes myosin heavy chain 11 is involved in vascular contractility, which is a definitive marker for smooth muscle cells (SMCs) during embryonic and postnatal development. Associations on chromosome 22 are likely due to two independent signals, with lead SNPs rs5760032 and rs133885, that are more than 2Mb apart. *CD34* is a protein that is expressed on the surface of hematopoietic stem cells and endothelial progenitor cells. *EPHB4*, a receptor in the Eph-ephrin signaling pathway, plays a crucial role in the cardiovascular system, particularly in the adult heart. *CACNA1C* is a gene that encodes the alpha-1 subunit of a voltage-dependent L-type calcium channel expressed in the human heart and brain. The branched-chain amino acid transaminase 1 (*BCAT1*) has been implicated in various cardiovascular diseases, including heart failure, dilated cardiomyopathy, and atherosclerosis. *HTR2B* encodes for serotonin receptor 2B [69]. *RGS6* is a regulator of G protein signaling that is expressed robustly in the heart, particularly in the sinoatrial node (SAN) and atrioventricular node (AVN), which are known to control heart rate and rhythm.

The *SCN5A* gene encodes the α subunit of the main cardiac sodium channel Nav1.5. This channel predominates inward sodium current (INa) and plays a critical role in regulation of cardiac electrophysiological function [59].

Static phenotypes. In this category we find the following genes: *SHOX2*, *PITX2*, *STRN*, *CCDC91*, *FILIP1L*, *RBM20*, *PRKCA*, *LMF1*, *MYOZ2*, *SRL* and *PDZRN3*.

The protein *PDZRN3* has been found to play a significant role in heart maturation and the development of heart failure.

Dynamic and static phenotypes. We now discuss associations with both dynamic and static variables, which represents a category with 16 discovered loci, among which are canonical cardiac genes *PLN*, *TTN*, *BAG3*, *TBX5*.

7.4.4 Gene ontology term enrichment

To elucidate possible mechanisms mediating the discovered associations, and to further assist with gene prioritization, we conducted gene ontology term enrichment using the g:Profiler R API. Results are presented in Table 7.4.

This analysis revealed significant associations with cardiac development processes, contractile structures (e.g., Z disc, myosin filament), and clinical phenotypes such as atrial fibrillation and thromboembolic stroke, supporting the relevance of the discovered loci to both structural and electrophysiological aspects of cardiac function.

7.4.5 Comparison with related traits

Handcrafted dynamical phenotypes In Figure 7.12 we display the $-\log_{10}(p)$ values for the associations between our top loci and handcrafted phenotypes. These phenotypes are: SV, LVEF/RVEF, LV absolute wall thickening and LV relative wall thickening (for the last two, the strongest association across the 17 AHA segments is shown). Also, the temporal width at half contraction was computed, as well as the position of end-systole as a fraction of the length of the full cardiac cycle. It can be seen that these simpler phenotypes fail to detect many of the loci linked to dynamic unsupervised phenotypes. Manhattan plots for the different traits are displayed in Figures 7.10 and 7.11.

Electrocardiographic traits. Given the expected overlap with ECG features, we downloaded summary statistics from a study investigating electrocardiographic traits, for comparison [100]: this study performed GWAS on the ECG signal throughout the cardiac cycle, one time point at a time.

In Figure 7.12, it can be seen that although most of our discovered loci are also found via testing of these ECG-derived phenotypes, the following loci are not found in [100] with $p < p_{GW}/500$

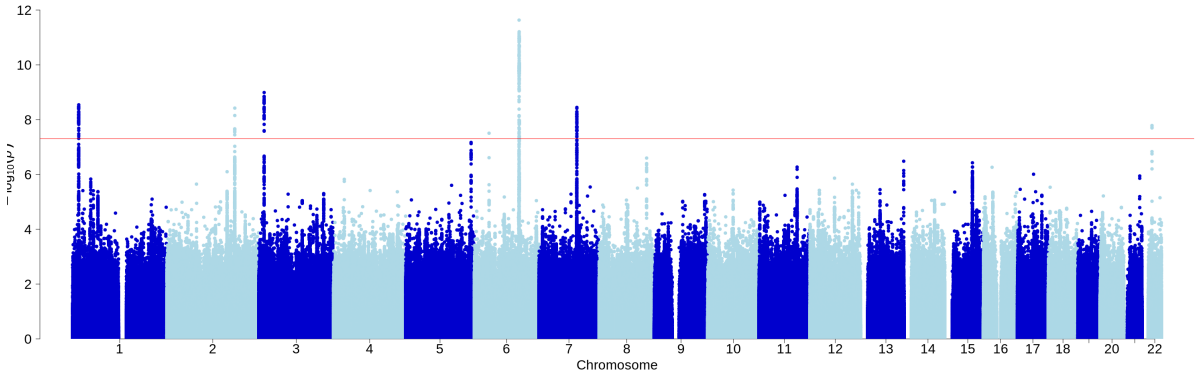

Figure 7.10: Manhattan plot for LV semi-amplitude duration: the temporal width between the time points at which the left ventricular volume reaches 50% of its stroke amplitude during contraction and relaxation, expressed as a fraction of the cardiac cycle length.

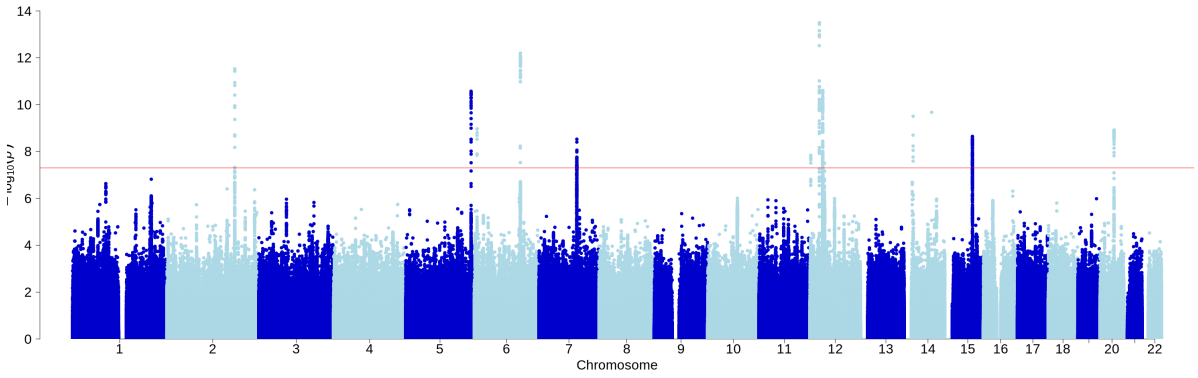

Figure 7.11: Manhattan plot for the position of the LVV minimum across the cardiac cycle, as a fraction of the length of the cycle.

(500 being the number of phenotypes, i.e. time points, tested in this study): *BCAR1*, *SYT10*, *ACTN2*, *HCN4*, *HMGA2*, *FLNC*, *GJA1*, *DSP*, *MYH11*.

7.5 Discussion

As exposed in the section 7.4, our phenotyping framework was able to retrieve loci that relate to dynamic and static cardiac phenotypes in a distinct way.

Most conspicuously, the associations termed as "predominantly dynamic" are almost entirely new with respect to our previous work; being *NKX2-5* and *NDRG2* the only two loci that had been detected in [17].

With respect to our investigation on static LV phenotypes, we remarkably extend the discoveries also by virtue of studying the four cardiac chambers across the full cardiac cycle. However, given the lower number of runs in the UPEs for each chamber, some of our previous "static" loci are not found here. This is most likely to be fixed by simply performing more runs, which will be done for the final journal version of this work (note that each training run takes approximately between 1 and 3 days). Another potential reason is that, unlike our previous results from section 6.5 , the *static* results here do not correspond to the ED phase; instead, we used the Euclidean time-averaged shape as the "content representative". This also helps explain the slight discrepancy in the found loci.

We consider potentially beneficial to employ different choices of the content representative as a hyperparameter to induce diversity in our ensemble. This idea is further supported by the observation (from previous studies, like [9, 78, 77, 8], as well as our own results on traditional phenotypes from Chapter 5) that cardiac indices computed at different cardiac phases, e.g. LVEDV and LVESV, can identify distinct loci, presumably reflecting the distinct underlying biology leading to variation in morphology in each case.

7.6 Conclusions

We have proposed an autoencoder-based network to perform unsupervised phenotyping of temporal sequences of cardiac meshes across the cardiac cycle, derived from CMR images by automatic segmentation. This network effectively incorporates the periodicity of cardiac motion by construction. In addition, we proposed to build a disentangled latent representation, where both

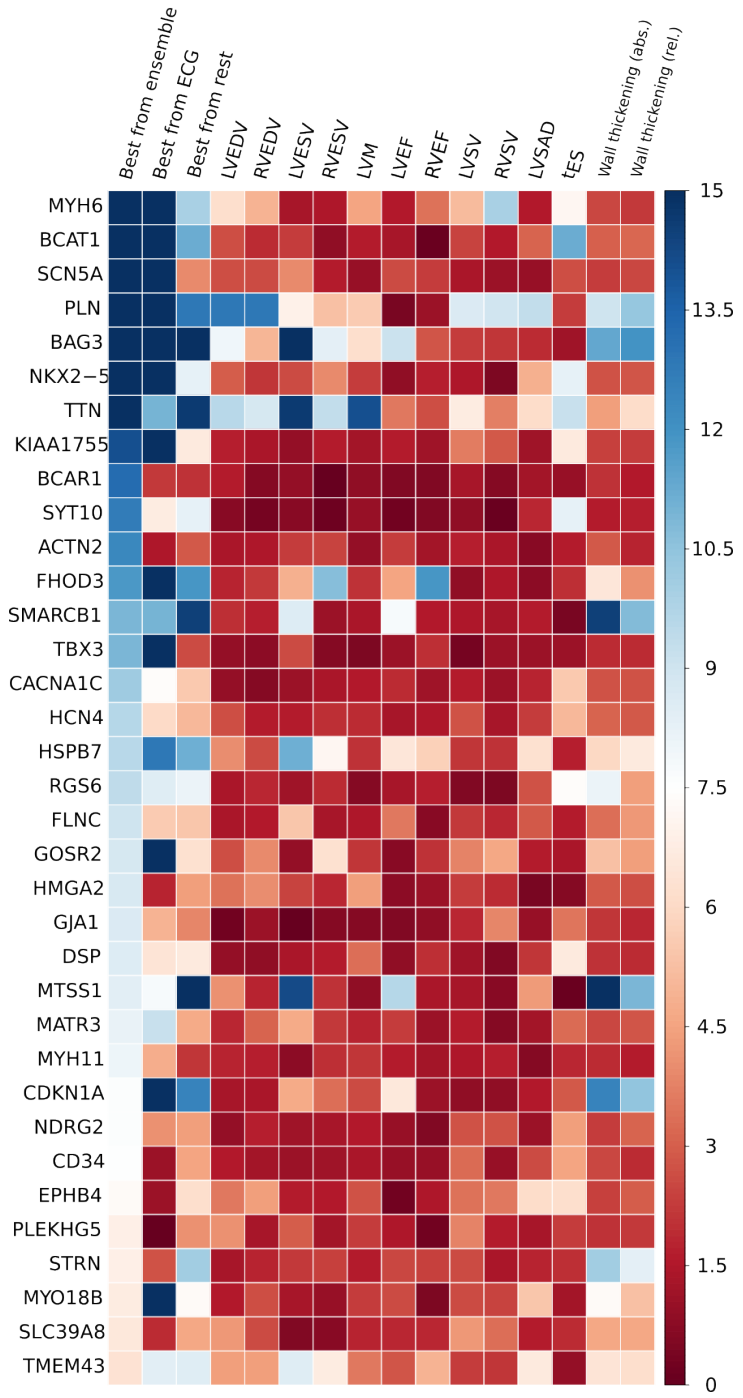

Figure 7.12: Comparison of $-\log_{10}(p)$ for the best latent variable per locus and other genetic associations for the same locus: the best from 500 electrocardiographic traits as studied by [100] and a series of handcrafted phenotypes derived by us. LVSAD: semi-amplitude duration in LVV-t curve, t_{ES}: time when end-systole occurs, normalized to the full cardiac cycle.

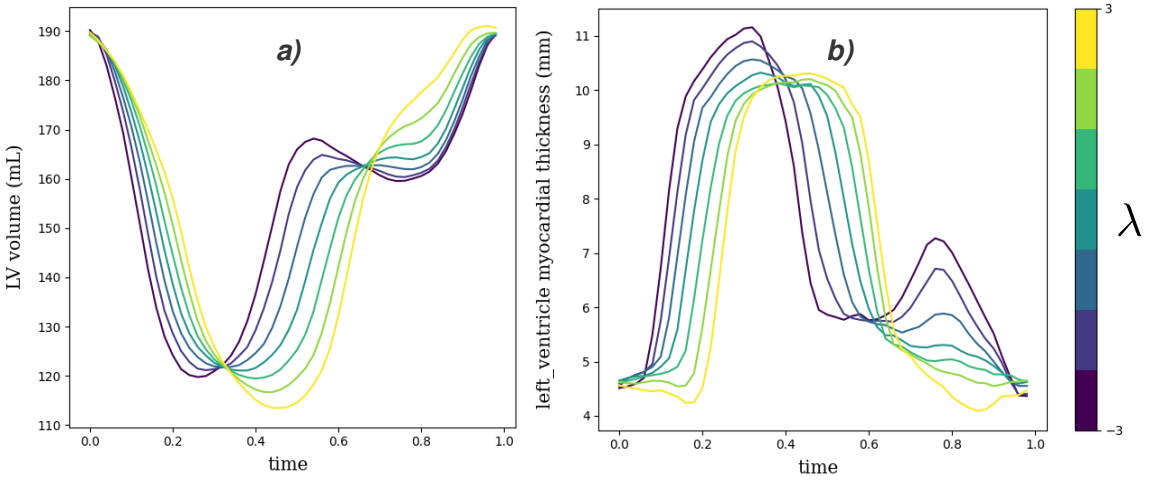

Figure 7.13: Synthetic curves for a) LVV and b) wall thickness at AHA segment 1 for the best LV latent variable of locus *SYT10*. The same phenotypic theme is observed for the best latent variable of the following loci: *GJA1*, *BCAT1*, *CACNA1C*, *HCN4*, *DSP*, *MATR3*, *MYH11* and *CD34*.

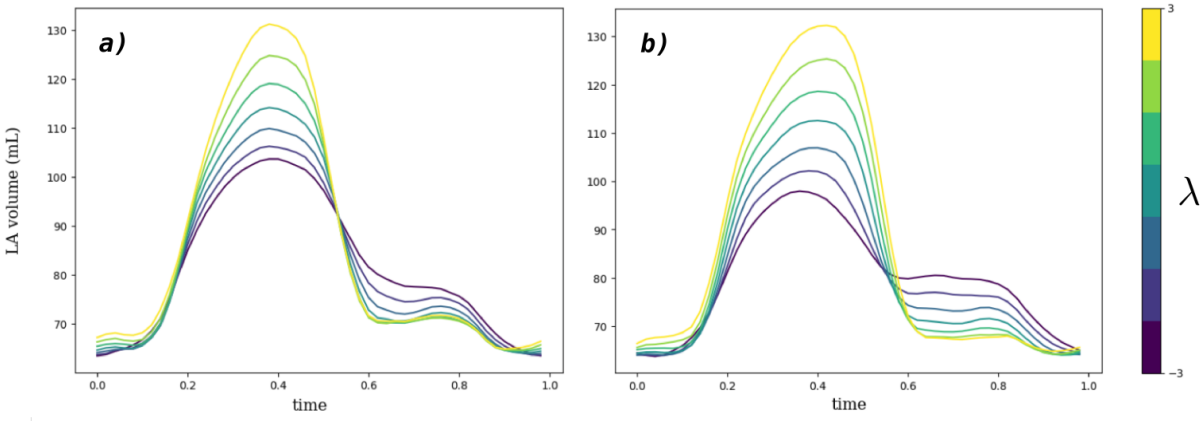

Figure 7.14: Synthetic LAV-t curves for the best LA dynamic variables for locus a) *BCAR1* and b) *HMGA2*, which are loci specific of LA phenotypes. Note that these phenotypes are linked to the maximum LA volume and very similar, however they are specific to either locus.

static and *dynamic* factors can be derived. Informed by biology, by which we expect dynamic phenotypes to be more enriched in genes with a role in the conduction system, our hypothesis is that distinct genetic factors will contribute to each type of phenotypic variation.

We have applied this phenotyping approach to the different cardiac chambers (LV, RV, LA and RA, as well as a combination of LV and RV) using the UPE framework from Chapter 6, consisting of training a set of networks for each chamber with different hyperparameters. We have succeeded in identifying distinct genetic variants for static and dynamic phenotypes, as we had hypothesised. Many of these loci had been identified in previous GWAS on ECG features, however we were able to find additional loci in spite of our lower sample size.

Limitations and challenges. While we have aimed at characterising the phenotypic variation captured with each of the dynamic latent variables by means of chamber volume-time curves and LV wall thickening, additional work needs to be carried out in order to fully understand the morphological and dynamical aspects that these latent factors model.

An additional drawback of our method is that it does not account for the differences in pulse rate in the population. In our approach, time is normalized to the length of a cardiac cycle.

Finally, it is of great interest to simultaneously study ECG- and CMR-derived features in a multimodal setting. Such study would presumably allow to further disentangle the latent representations, gaining further insights into the mechanisms underlying the genetic associations.

7.7 Code availability

The codebase for this chapter will be made publicly available upon acceptance of the corresponding publication. The repository is called `CardiacMotion`.

Code for generating synthetic shapes has already been made publicly available on GitHub, under the URL: `www.github.com/rbonazzola/SyntheticMeshGeneration`.

Source	Term name	p-value	Intersection size	Term size
GO:BP	heart development	1.1×10^{-4}	11	600
HP	Thromboembolic stroke	1.7×10^{-4}	6	74
HP	Atrial fibrillation	3.8×10^{-4}	6	84
HP	Atrial arrhythmia	6.5×10^{-4}	6	92
HP	Thromboembolism	1.1×10^{-3}	6	100
GO:BP	circulatory system development	1.3×10^{-3}	13	1134
GO:BP	cardiac muscle cell development	1.5×10^{-3}	5	79
HP	Orthopnea	1.5×10^{-3}	6	106
GO:BP	cardiocyte differentiation	1.7×10^{-3}	6	150
GO:BP	cardiac cell development	2.1×10^{-3}	5	85
GO:CC	Z disc	2.7×10^{-3}	5	134
GO:BP	striated muscle cell development	2.8×10^{-3}	6	164
GO:CC	actin cytoskeleton	3.8×10^{-3}	8	516
GO:CC	myosin filament	4.1×10^{-3}	3	23
GO:CC	I band	4.6×10^{-3}	5	149
HP	Supraventricular arrhythmia	5.7×10^{-3}	6	133
GO:CC	myosin II complex	6.0×10^{-3}	3	26
GO:BP	adult heart development	6.2×10^{-3}	3	14
GO:BP	muscle cell development	6.8×10^{-3}	6	191
HP	Tricuspid regurgitation	7.2×10^{-3}	5	77
HP	Lipoatrophy	7.6×10^{-3}	5	78
HP	Abnormal EKG	8.1×10^{-3}	6	141
GO:BP	atrioventricular node cell fate commitment	8.2×10^{-3}	2	2
HP	Abnormal tricuspid valve physiology	8.7×10^{-3}	5	80
HP	Exertional dyspnea	8.7×10^{-3}	6	143

Table 7.4: Top 25 enriched terms from g:Profiler analysis.

Chapter 8

Conclusions and future directions

In this last chapter, I bring this thesis to a closure by summarising the findings arising from this work, as well as some limitations and opportunities for future research.

8.1 Summary of key findings

Genetic basis and environmental factors influencing traditional CMR-derived cardiac indices. Firstly, we have performed genome-wide association studies (GWAS) on clinical cardiac phenotypes extracted from cine-cardiovascular magnetic resonance (cine-CMR) from UK Biobank (UKB). We have also studied how these phenotypes relate to demographic variables. In spite of its lack of novelty, this study stands out because it uses the full sample size available from UKB as of June 2023. This has enabled the discovery of a number of previously unveiled genetic loci. Furthermore, we study new handcrafted phenotypes (for instance, left-ventricular sphericity and normalized LV myocardial thickness).

A deep ensemble-based approach for genetic discovery of left-ventricular shape. Secondly, we have performed unsupervised phenotyping on static LV meshes (at end-diastole), by means of shape PCA and convolutional mesh autoencoders (CoMA). We confirmed the hypothesis that an unsupervised phenotyping approach would be beneficial in capturing more subtle patterns of shape variation, thereby enhancing genetic discovery.

A second important contribution is the idea of employing ensembles of CoMA networks, which we coin Unsupervised Phenotype Ensembles (UPE), with different hyperparameters and random

seeds, to better sample the phenotypic space. We substantiated that this approach enables to exploit more exhaustively the information contained in the mesh, which in turn leads to additional genetic discoveries.

Study of dynamic patterns Finally, we have developed a novel approach for the phenotyping of the full cine sequence. By means of this approach, we have extended the previous study to also consider dynamic phenotypes. Furthermore, we study the four cardiac chambers. After performing GWAS on the latent representations, we were able to clearly distinguish loci with an association to static phenotypes, from those with an association to dynamic phenotypes.

8.2 Implications and significance

On the biological side, we detect a remarkable number of new genetic locations linked to cardiac shape and motion phenotypes. As with other GWAS findings, ours provide a starting point for further research into the role of the candidate genes. This could lead to a better understanding of mechanisms for cardiac physiology in health and disease, which in turn may derive in the development of better therapies or assist personalised healthcare.

On the methodological side, we highlight the usefulness of full 3D meshes representing the myocardium, along with geometric deep learning techniques, for obtaining more detailed quantitative phenotypes that enhance the discovery of genetic associations. We propose a novel neural architecture for phenotyping of dynamic cardiac meshes and showcase its utility in the case of genetic association studies. Furthermore, our ensemble-based phenotyping approach provides an opportunity to increase gene discovery also for other modalities and organs, whenever deep-learning-based phenotypes is performed.

8.3 Limitations, challenges and opportunites for future research

In this subsection, we highlight some of the limitations of this work and, in some cases, ideas to overcome them in future work.

The need for more diverse ancestries Firstly, this study has focused on the major ancestry group present in UKB, which is the British ancestry. While we argue that the involvement in cardiac physiology of genes belonging to discovered loci is most likely generalisable to different

populations, the specific genetic variation that has been detected in the GBR population is possibly not shared across the board; neither are the SNP effect sizes. Likewise, differing allele frequencies and LD patterns present in other populations may help discover additional relevant genes. In view of these reasons, we argue that it is desirable to count with linked imaging and genetic data from other ethnicities apart from GBR.

Sexual chromosomes and sex-specific differences In this work, we have only investigated autosomes. This is a shortcoming not only of our work but of many published GWAS. Sex chromosomes offer some complications due to born males having one copy of X and one of Y chromosome, whereas born females have two copies of the X chromosome[1]. This implies that the range of dosages for the genetic variants in the X chromosome is $\{0, 1\}$ for males, and $\{0, 1, 2\}$ for females.

Furthermore, an interesting avenue is determining whether our genetic findings present significantly different effect sizes within the male and female subpopulations, thus providing the opportunity to pinpoint sex-specific effects.

Use of multi-organ and multi-modality data The breadth of the imaging data available in UKB spans different organs and modalities. For example, in the case of the heart, tagging CMR is a modality where a different radiofrequency pulse sequence is applied during the MRI acquisition, which allows to create a dark grid (or *tags*) in the image, enabling the tracking of physical points in the myocardium over time. This modality enables a better characterisation of the deformation of the muscle, in particular torsion patterns which are not easily detectable in cine-CMR. Also within the scope of heart modalities, electrocardiogram (ECG) data is also available in UKB and of particular interest.

8.3.1 Exome-wide studies

UK Biobank has released full exomes, with full coverage over the coding region of the genome [10]. This data has not been employed in this study, and constitutes a rich resource that could enable deeper understanding of the role of rare coding variation on image-derived phenotypes. In addition, it can help pinpoint causal coding variants by not relying on genotyped proxy SNP markers in high LD.

[1]There are also subjects with more than 2 sexual chromosomes, but they constitute a very small minority.

8.3.2 Employing texture-endowed volumetric meshes

Since we have investigated purely geometrical myocardial surface features, we are at risk of confounding geometrically similar phenotypes which arise from different biological pathways. We argue that the incorporation of texture information may help distinguish these two geometrically similar phenotypes. Take, for instance, the case of an enlarged heart: the texture of myocardial tissue can vary greatly depending on the underlying cause of the enlargement, e.g. an infection like myocarditis, athlete's heart or hypertension. A joint characterisation of tissue texture and morphology may lead to phenotypes that better map to the underlying biology, and hence are better suited for genetic association studies.

In this sense, while landmark-based surface-based segmentation produces a standardised object that effectively "distills" relevant anatomical information from an image into a surface mesh, this surface mesh representation used in this work seems inconvenient as it is unable to incorporate texture. Here, I briefly discuss an idea towards conducting a more complete retrieval of the image's information, while not sacrificing the advantages of using registered meshes.

The use of *volumetric meshes* of the myocardium, that is, meshes including vertices that live in the interior of the cardiac muscle, enable the use of voxel intensities as node features, thus preserving the texture information from the original image, while using a registered mesh representation which, as we have seen, has significant advantages. The segmentation approach in [40], a work currently published as a preprint and which I have co-authored, produces high-quality volumetric tetrahedral meshes and is therefore a good candidate for producing the mesh data for following this line of research.

8.3.3 Implication of discovered loci in disease

An important limitation of our study is that it did not focus on the pathological implications of our discovered loci. In this sense, databases that are more enriched in subjects with different cardiac pathologies of interest (as opposed to UKB where the population was recruited with the aim of being representative of the whole) will be useful in assessing the role of our novel associations. In particular, the use of Mendelian randomisation could help shed light on the causal roles of our derived structural phenotypes.

8.3.4 Electromechanical models of the heart

We anticipate that integrating an electromechanical-inspired model could offer valuable insights into the mechanisms underlying genetic associations, while also serving as an independent validation tool. In this context, the 2002 work by Denis Noble, who used biophysically-detailed simulations to link specific genetic mutations to cardiac arrhythmias, illustrates the power of mechanistic modeling in elucidating genotype–phenotype relationships [71]. Despite its potential, there is a notable scarcity of studies exploring this interdisciplinary approach. This naturally raises the question of whether such scarcity stems from inherent limitations, or rather from a lack of interdisciplinary dialogue and collaboration

8.4 Final remarks

In this thesis, we have demonstrated how state-of-the-art segmentation and representation learning techniques can provide novel insights into cardiac biology, substantiating the view that deeply multi-disciplinary work can help us gain valuable knowledge about the workings of biological systems.

We expect that the novel knowledge on the genetics of cardiac shape and function which this thesis has unveiled, as well as the methodological contributions, will constitute building blocks for understanding the complex puzzle of cardiac physiology in health and disease.

References

[1] Abdel Abdellaoui, Loic Yengo, Karin JH Verweij, and Peter M Visscher. "15 years of GWAS discovery: realizing the promise". In: *The American Journal of Human Genetics* 110.2 (2023), pp. 179–194.

[2] Gad Abraham, Yixuan Qiu, and Michael Inouye. "FlashPCA2: principal component analysis of Biobank-scale genotype datasets". In: *Bioinformatics* 33.17 (2017), pp. 2776–2778.

[3] Melissa Anfinson, Robert H Fitts, John W Lough, Jeanne M James, Pippa M Simpson, Stephanie S Handler, Michael E Mitchell, and Aoy Tomita-Mitchell. "Significance of α-Myosin Heavy Chain (MYH6) variants in hypoplastic left heart syndrome and related cardiovascular diseases". In: *Journal of Cardiovascular Development and Disease* 9.5 (2022), p. 144.

[4] Veronica Astro, Gustavo Ramirez-Calderon, Roberta Pennucci, Jonatan Caroli, Alfonso Saera-Vila, Kelly Cardona-Londoño, Chiara Forastieri, Elisabetta Fiacco, Fatima Maksoud, Maryam Alowaysi, et al. "Fine-tuned KDM1A alternative splicing regulates human cardiomyogenesis through an enzymatic-independent mechanism". In: *iScience* 25.7 (2022).

[5] Rahman Attar, Marco Pereañez, Christopher Bowles, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, and Alejandro F Frangi. "3D cardiac shape prediction with deep neural networks: simultaneous use of images and patient metadata". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 586–594.

[6] Rahman Attar, Marco Pereañez, Ali Gooya, Xènia Albà, Le Zhang, Milton Hoz de Vila, Aaron M. Lee, Nay Aung, Elena Lukaschuk, Mihir M. Sanghvi, Kenneth Fung, Jose Miguel Paiva, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, and Alejandro F. Frangi. "Quantitative CMR population imaging on 20,000 subjects of the UK Biobank imaging study: LV/RV quantification pipeline and its evaluation". en. In: *Medical Image Analysis* 56 (Aug. 2019), pp. 26–42. ISSN: 13618415. DOI: 10.1016/j.media.2019.05.006.

[7] Nay Aung, Luis R Lopes, Stefan van Duijvenboden, Andrew R Harper, Anuj Goel, Christopher Grace, Carolyn Y Ho, William S Weintraub, Christopher M Kramer, Stefan Neubauer, et al. "Genome-Wide Analysis of Left Ventricular Maximum Wall Thickness in the UK Biobank Cohort Reveals a Shared Genetic Background With Hypertrophic Cardiomyopathy". In: *Circulation: Genomic and Precision Medicine* 16.1 (2023), e003716.

[8] Nay Aung, Jose D Vargas, Chaojie Yang, Kenneth Fung, Mihir M Sanghvi, Stefan K Piechnik, Stefan Neubauer, Ani Manichaikul, Jerome I Rotter, Kent D Taylor, et al. "Genome-wide association analysis reveals insights into the genetic architecture of right ventricular structure and function". In: *Nature genetics* 54.6 (2022), pp. 783–791.

[9] Nay Aung, Jose D. Vargas, Chaojie Yang, Claudia P. Cabrera, Helen R. Warren, Kenneth Fung, Evan Tzanis, Michael R. Barnes, Jerome I. Rotter, Kent D. Taylor, Ani W. Manichaikul, Joao A.C. Lima, David A. Bluemke, Stefan K. Piechnik, Stefan Neubauer, Patricia B. Munroe, and Steffen E. Petersen. "Genome-Wide Analysis of Left Ventricular Image-Derived Phenotypes Identifies Fourteen Loci Associated With Cardiac Morphogenesis and Heart Failure Development". en. In: *Circulation* 140.16 (Oct. 2019), pp. 1318–1330. ISSN: 0009-7322, 1524-4539. DOI: 10.1161/CIRCULATIONAHA.119.041161.

[10] Joshua D Backman, Alexander H Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D Kessler, Christian Benner, Daren Liu, Adam E Locke, Suganthi Balasubramanian, et al. "Exome sequencing and analysis of 454,787 UK Biobank participants". In: *Nature* 599.7886 (2021), pp. 628–634.

[11] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. "Automated

cardiovascular magnetic resonance image analysis with fully convolutional networks". In: *Journal of Cardiovascular Magnetic Resonance* 20.1 (2018), pp. 1–12.

[12] Wenjia Bai, Hideaki Suzuki, Jian Huang, Catherine Francis, Shuo Wang, Giacomo Tarroni, Florian Guitton, Nay Aung, Kenneth Fung, Steffen E Petersen, et al. "A population-based phenome-wide association study of cardiac and aortic structure and function". In: *Nature medicine* 26.10 (2020), pp. 1654–1662.

[13] Martijn L Bakker, Bastiaan J Boukens, Mathilda TM Mommersteeg, Janynke F Brons, Vincent Wakker, Antoon FM Moorman, and Vincent M Christoffels. "Transcription factor Tbx3 is required for the specification of the atrioventricular conduction system". In: *Circulation research* 102.11 (2008), pp. 1340–1349.

[14] Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L Edwards, et al. "Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics". In: *Nature Communications* 9.1 (2018), p. 1825.

[15] Tomaz Berisa and Joseph K Pickrell. "Approximately independent linkage disequilibrium blocks in human populations". In: *Bioinformatics* 32.2 (2016), p. 283.

[16] Carlo Biffi, Antonio de Marvao, Mark I Attard, Timothy J W Dawes, Nicola Whiffin, Wenjia Bai, Wenzhe Shi, Catherine Francis, Hannah Meyer, Rachel Buchan, Stuart A Cook, Daniel Rueckert, and Declan P O'Regan. "Three-dimensional cardiovascular imaging-genetics: a mass univariate framework". en. In: *Bioinformatics* 34.1 (Jan. 2018). Ed. by Robert Murphy, pp. 97–103. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btx552.

[17] Rodrigo Bonazzola, Enzo Ferrante, Nishant Ravikumar, Yan Xia, Bernard Keavney, Sven Plein, Tanveer Syeda-Mahmood, and Alejandro F Frangi. "Unsupervised ensemble-based phenotyping enhances discoverability of genes related to left-ventricular morphology". In: *Nature Machine Intelligence* 6.3 (2024).

[18] Rodrigo Bonazzola, Nishant Ravikumar, Rahman Attar, Enzo Ferrante, Tanveer Syeda-Mahmood, and Alejandro F Frangi. "Image-Derived Phenotype Extraction for Genetic Discovery via Unsupervised Deep Learning in CMR Images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 699–708.

[19] Inge R Boulet, Adam L Raes, Natacha Ottschytsch, and Dirk J Snyders. "Functional effects of a KCNQ1 mutation associated with the long QT syndrome". In: *Cardiovascular research* 70.3 (2006), pp. 466–474.

[20] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. "Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 7213–7222.

[21] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. "Geometric Deep Learning: Going beyond Euclidean data". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42. DOI: 10.1109/MSP.2017.2693418.

[22] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. "Spectral networks and locally connected networks on graphs". In: *arXiv preprint arXiv:1312.6203* (2013).

[23] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Gil McVean, Stephen Leslie, Peter Donnelly, and Jonathan Marchini. "Genome-wide genetic data on ∼500,000 UK Biobank participants". In: *bioRxiv* (2017). DOI: 10.1101/166298. eprint: https://www.biorxiv.org/content/early/2017/07/20/166298.full.pdf.

[24] Statistics group Ciampi Antonio 8 Greenwood Celia MT (co-chair) 7 8 14 19 Hendricks Audrey E. 1 12 Li Rui 7 13 14 Metrustry Sarah 5 Oualkacha Karim 80 Tachmazidou Ioanna 1 Xu ChangJiang 7 8 Zeggini Eleftheria (co-chair) 1 et al. "The UK10K project identifies rare variants in health and disease". In: *Nature* 526.7571 (2015), pp. 82–90.

[25] Alexander Guldmann Clausen, Oliver Bundgaard Vad, Julie Husted Andersen, and Morten Salling Olesen. "Loss-of-function variants in the SYNPO2L gene are associated with atrial fibrillation". In: *Frontiers in cardiovascular medicine* 8 (2021), p. 650667.

[26] Francis S Collins, Michael Morgan, and Aristides Patrinos. "The Human Genome Project: lessons from large-scale biology". In: *Science* 300.5617 (2003), pp. 286–290.

[27] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), p. 68.

[28] GTEx Consortium. "The GTEx Consortium atlas of genetic regulatory effects across human tissues". In: *Science* 369.6509 (2020), pp. 1318–1330.

[29] Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, et al. "Ensembl 2022". In: *Nucleic acids research* 50.D1 (2022), pp. D988–D995.

[30] Antonio De Marvao, Timothy JW Dawes, and Declan P O'Regan. "Artificial intelligence for cardiac imaging-genetics research". In: *Frontiers in Cardiovascular Medicine* 6 (2020), p. 195.

[31] Raymond N Deepe, Jenna R Drummond, Renélyn A Wolters, Emily A Fitzgerald, Hannah G Tarolli, Andrew B Harvey, and Andy Wessels. "Sox9 Expression in the Second Heart Field; A Morphological Assessment of the Importance to Cardiac Development with Emphasis on Atrioventricular Septation". In: *Journal of Cardiovascular Development and Disease* 9.11 (2022), p. 376.

[32] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering". In: *Advances in Neural Information Processing Systems*. 2016. URL: `https://arxiv.org/abs/1606.09375`.

[33] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. "Convolutional neural networks on graphs with fast localized spectral filtering". In: *Advances in neural information processing systems* 29 (2016).

[34] Tim R. Eijgenraam, Herman H.W. Silljé, and Rudolf A. de Boer. "Current understanding of fibrosis in genetic cardiomyopathies". en. In: *Trends in Cardiovascular Medicine* 30.6 (Aug. 2020), pp. 353–361. ISSN: 10501738. DOI: 10.1016/j.tcm.2019.09.003.

[35] Ramón A Espinoza-Lewis, Ling Yu, Fenglei He, Hongbing Liu, Ruhang Tang, Jiangli Shi, Xiaoxiao Sun, James F Martin, Dazhi Wang, Jing Yang, et al. "Shox2 is essential for the differentiation of cardiac pacemaker cells by repressing Nkx2-5". In: *Developmental biology* 327.2 (2009), pp. 376–385.

[36] Andreas Fischer, Nina Schumacher, Manfred Maier, Michael Sendtner, and Manfred Gessler. "The Notch target genes Hey1 and Hey2 are required for embryonic vascular development". In: *Genes & development* 18.8 (2004), pp. 901–911.

[37] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. "Deep ensembles: A loss landscape perspective". In: *arXiv preprint arXiv:1912.02757* (2019).

[38] Alejandro F Frangi, Wiro J Niessen, and Max A Viergever. "Three-dimensional modeling for functional analysis of cardiac images, a review". In: *IEEE transactions on medical imaging* 20.1 (2001), pp. 2–5.

[39] Derk Frank, Robert Frauen, Christiane Hanselmann, Christian Kuhn, Rainer Will, Johanne Gantenberg, Laszlo Füzesi, Hugo A Katus, and Norbert Frey. "Lmcd1/Dyxin, a novel Z-disc associated LIM protein, mediates cardiac hypertrophy in vitro and in vivo". In: *Journal of molecular and cellular cardiology* 49.4 (2010), pp. 673–682.

[40] Nicolás Gaggion, Benjamin A Matheson, Yan Xia, Rodrigo Bonazzola, Nishant Ravikumar, Zeike A Taylor, Diego H Milone, Alejandro F Frangi, and Enzo Ferrante. "Multi-view Hybrid Graph Convolutional Network for Volume-to-mesh Reconstruction in Cardiovascular MRI". In: *arXiv preprint arXiv:2311.13706* (2023).

[41] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. "A gene-based association method for mapping traits using reference transcriptome data". In: *Nature genetics* 47.9 (2015), pp. 1091–1098.

[42] Michael Garland and Paul S. Heckbert. "Surface simplification using quadric error metrics". en. In: *Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97*. ACM Press, 1997, pp. 209–216. ISBN: 978-0-89791-896-1. DOI: 10.1145/258734.258849.

[43] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli L Yu, HM Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. "The international HapMap project". In: (2003).

[44] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[45] Henk L. Granzier and Siegfried Labeit. "The Giant Protein Titin: A Major Player in Myocardial Mechanics, Signaling, and Disease". en. In: *Circulation Research* 94.3 (Feb. 2004), pp. 284–295. ISSN: 0009-7330, 1524-4571.

[46] Lingjuan Hong, Na Li, Victor Gasque, Sameet Mehta, Lupeng Ye, Yinyu Wu, Jinyu Li, Andreas Gewies, Jürgen Ruland, Karen K Hirschi, et al. "Prdm6 controls heart development by regulating neural crest cell differentiation and migration". In: *JCI insight* 7.4 (2022).

[47] Arnold M Katz. *Physiology of the Heart*. Lippincott Williams & Wilkins, 2010.

[48] Eri Kawakami, Akinori Tokunaga, Manabu Ozawa, Reiko Sakamoto, and Nobuaki Yoshida. "The histone demethylase Fbxl11/Kdm2a plays an essential role in embryonic development by repressing cell-cycle regulators". In: *Mechanisms of development* 135 (2015), pp. 31–42.

[49] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[50] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[51] Matthias Kirchler, Stefan Konigorski, Matthias Norden, Christian Meltendorf, Marius Kloft, Claudia Schurmann, and Christoph Lippert. "TransferGWAS: GWAS of images using deep transfer learning". In: *Bioinformatics* 38.14 (2022), pp. 3621–3628.

[52] Tijana Knezevic, Valerie D Myers, Jennifer Gordon, Douglas G Tilley, Thomas E Sharp, JuFang Wang, Kamel Khalili, Joseph Y Cheung, and Arthur M Feldman. "BAG3: a new player in the heart failure paradigm". In: *Heart failure reviews* 20.4 (2015), pp. 423–434.

[53] Jan Koelemen, Michael Gotthardt, Lars M Steinmetz, and Benjamin Meder. "RBM20-related cardiomyopathy: current understanding and future options". In: *Journal of Clinical Medicine* 10.18 (2021), p. 4101.

[54] Thomas Küstner, Niccolo Fuin, Kerstin Hammernik, Aurelien Bustin, Haikun Qi, Reza Hajhosseiny, Pier Giorgio Masci, Radhouene Neji, Daniel Rueckert, René M Botnar, et al. "CINENet: deep learning-based 3D cardiac CINE MRI reconstruction with multi-coil complex-valued 4D spatio-temporal convolutions". In: *Scientific reports* 10.1 (2020), p. 13710.

[55] Thomas LaFramboise. "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". In: *Nucleic acids research* 37.13 (2009), pp. 4181–4193.

[56] Harald Lahm, Meiwen Jia, Martina Dreßen, Felix Wirth, Nazan Puluca, Ralf Gilsbach, Bernard D Keavney, Julie Cleuziou, Nicole Beck, Olga Bondareva, et al. "Congenital heart disease risk loci identified by genome-wide association study in European patients". In: *The Journal of clinical investigation* 131.2 (2021).

[57] Michael G Levin, Noah L Tsao, Pankhuri Singhal, Chang Liu, Ha My T Vy, Ishan Paranjpe, Joshua D Backman, Tiffany R Bellomo, William P Bone, Kiran J Biddinger, et al. "Genome-wide association and multi-trait analyses characterize the common genetic architecture of heart failure". In: *Nature Communications* 13.1 (2022), p. 6914.

[58] Bin Li, Yongkun Zhan, Qianqian Liang, Chen Xu, Xinyan Zhou, Huanhuan Cai, Yufan Zheng, Yifan Guo, Lei Wang, Wenqing Qiu, et al. "Isogenic human pluripotent stem

cell disease models reveal ABRA deficiency underlies cTnT mutation-induced familial dilated cardiomyopathy". In: *Protein & Cell* 13.1 (2022), pp. 65–71.

[59] Wenjia Li, Lei Yin, Cheng Shen, Kai Hu, Junbo Ge, and Aijun Sun. "SCN5A variants: association with cardiac disorders". In: *Frontiers in physiology* 9 (2018), p. 1372.

[60] Jingyu Liu and Vince D Calhoun. "A review of multivariate analyses in imaging genetics". In: *Frontiers in neuroinformatics* 8 (2014), p. 29.

[61] Feng Lv, Xiaojuan Ge, Peipei Qian, Xiaofeng Lu, Dong Liu, and Changsheng Chen. "Neuron navigator 3 (NAV3) is required for heart development in zebrafish". In: *Fish Physiology and Biochemistry* 48.1 (2022), pp. 173–183.

[62] David H MacLennan, Michio Asahi, and A Russell Tupling. "The regulation of SERCA-type pumps by phospholamban and sarcolipin". en. In: *Annals of the New York Academy of Sciences* 986 (Apr. 2003), pp. 472–480. DOI: 10.1111/j.1749-6632.2003.tb07231.x..

[63] Meenakshi Maitra, Sara N Koenig, Deepak Srivastava, and Vidu Garg. "Identification of GATA6 sequence variants in patients with congenital heart defects". In: *Pediatric Research* 68.4 (2010), pp. 281–285.

[64] Ruairidh IR Martin, Mahsa Sheikhali Babaei, Mun-Kit Choy, W Andrew Owens, Timothy JA Chico, Daniel Keenan, Nizar Yonan, Mauro Santibáñez Koref, and Bernard D Keavney. "Genetic variants associated with risk of atrial fibrillation regulate expression of PITX2, CAV1, MYOZ1, C9orf3 and FANCC". In: *Journal of molecular and cellular cardiology* 85 (2015), pp. 207–214.

[65] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, Yang Luo, Carlo Sidore, Alan Kwong, Nicholas Timpson, Seppo Koskinen, and et al. "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature genetics* 48.10 (2016), pp. 1279–1283.

[66] Kathryn M Meurs, Joshua A Stern, DD Sisson, Mark D Kittleson, SM Cunningham, MK Ames, CE Atkins, T DeFrancesco, TE Hodge, BW Keene, et al. "Association of

dilated cardiomyopathy with the striatin mutation genotype in boxer dogs". In: *Journal of Veterinary Internal Medicine* 27.6 (2013), pp. 1437–1440.

[67] Hannah V Meyer, Timothy JW Dawes, Marta Serrani, Wenjia Bai, Paweł Tokarczuk, Jiashen Cai, Antonio de Marvao, Albert Henry, R Thomas Lumbers, Jakob Gierten, et al. "Genetic and functional insights into the fractal structure of the heart". In: *Nature* 584.7822 (2020), pp. 589–594.

[68] Moni Nader, Shahd Alotaibi, Ebtehal Alsolme, Bariaa Khalil, Ahmed Abu-Zaid, Rahmah Alsomali, Dana Bakheet, and Nduna Dzimiri. "Cardiac striatin interacts with caveolin-3 and calmodulin in a calcium sensitive manner and regulates cardiomyocyte spontaneous contraction rate". In: *Canadian Journal of Physiology and Pharmacology* 95.10 (2017), pp. 1306–1312.

[69] Canan G Nebigil, Pierre Hickel, Nadia Messaddeq, Jean-Luc Vonesch, Marie P Douchet, Laurent Monassier, Katalin György, Rachel Matz, Ramaroson Andriantsitohaina, Philippe Manivet, et al. "Ablation of serotonin 5-HT2B receptors in mice leads to abnormal cardiac structure and function". In: *Circulation* 103.24 (2001), pp. 2973–2979.

[70] Jonas B Nielsen, Lars G Fritsche, Wei Zhou, Tanya M Teslovich, Oddgeir L Holmen, Stefan Gustafsson, Maiken E Gabrielsen, Ellen M Schmidt, Robin Beaumont, Brooke N Wolford, et al. "Genome-wide study of atrial fibrillation identifies seven risk loci and highlights biological pathways and regulatory elements involved in cardiac development". In: *The American Journal of Human Genetics* 102.1 (2018), pp. 103–115.

[71] Denis Noble. "Modeling the heart–from genes to cells to the whole organ". In: *Science* 295.5560 (2002), pp. 1678–1682.

[72] Juan Pablo Ochoa, María Sabater-Molina, José Manuel García-Pinilla, Jens Mogensen, Alejandra Restrepo-Córdoba, Julián Palomino-Doza, Eduardo Villacorta, Marina Martinez-Moreno, Javier Ramos-Maqueda, Esther Zorio, et al. "Formin homology 2 domain containing 3 (FHOD3) is a genetic basis for hypertrophic cardiomyopathy". In: *Journal of the American College of Cardiology* 72.20 (2018), pp. 2457–2467.

[73] Karl Pearson. "On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[74] Ricardo Cardoso Pereira, Miriam Seoane Santos, Pedro Pereira Rodrigues, and Pedro Henriques Abreu. "Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes". In: *Journal of Artificial Intelligence Research* 69 (2020), pp. 1255–1285.

[75] Steffen E Petersen, Nay Aung, Mihir M Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Jane M Francis, Mohammed Y Khanji, Elena Lukaschuk, Aaron M Lee, et al. "Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort". In: *Journal of cardiovascular magnetic resonance* 19.1 (2016), p. 18.

[76] James P Pirruccello, Mark D Chaffin, Elizabeth L Chou, Stephen J Fleming, Honghuang Lin, Mahan Nekoui, Shaan Khurshid, Samuel F Friedman, Alexander G Bick, Alessandro Arduini, et al. "Deep learning enables genetic analysis of the human thoracic aorta". In: *Nature genetics* 54.1 (2022), pp. 40–51.

[77] James P Pirruccello, Paolo Di Achille, Victor Nauffal, Mahan Nekoui, Samuel F Friedman, Marcus DR Klarqvist, Mark D Chaffin, Lu-Chen Weng, Jonathan W Cunningham, Shaan Khurshid, et al. "Genetic analysis of right heart structure and function in 40,000 people". In: *Nature genetics* 54.6 (2022), pp. 792–803.

[78] James P. Pirruccello, Alexander Bick, Minxian Wang, Mark Chaffin, Samuel Friedman, Jie Yao, Xiuqing Guo, Bharath Ambale Venkatesh, Kent D. Taylor, Wendy S. Post, Stephen Rich, Joao A. C. Lima, Jerome I. Rotter, Anthony Philippakis, Steven A. Lubitz, Patrick T. Ellinor, Amit V. Khera, Sekar Kathiresan, and Krishna G. Aragam. "Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy". en. In: *Nature Communications* 11.1 (Dec. 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-15823-7.

[79] Alkes L Price, Michael E Weale, Nick Patterson, Simon R Myers, Anna C Need, Kevin V Shianna, Dongliang Ge, Jerome I Rotter, Esther Torres, Kent D Taylor, et al. "Long-

range LD can confound genome scans in admixed populations". In: *American Journal of Human Genetics* 83.1 (2008), p. 132.

[80] Simon JD Prince. *Understanding Deep Learning*. MIT press, 2023.

[81] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. "Generating 3D Faces Using Convolutional Mesh Autoencoders". In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Vol. 11207. Cham: Springer International Publishing, 2018, pp. 725–741. DOI: 10.1007/978-3-030-01219-9_43.

[82] Nibedita Rath, Zhishan Wang, Min Min Lu, and Edward E Morrisey. "LMCD1/Dyxin is a novel transcriptional cofactor that restricts GATA6 function by inhibiting DNA binding". In: *Molecular and cellular biology* 25.20 (2005), pp. 8864–8873.

[83] Nishant Ravikumar, Ali Gooya, Alejandro F Frangi, and Zeike A Taylor. "Generalised coherent point drift for group-wise registration of multi-dimensional point sets". In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*. Springer. 2017, pp. 309–316.

[84] Louise Reilly and Lee L Eckhardt. "Cardiac potassium inward rectifier Kir2: Review of structure, regulation, pharmacology, and arrhythmogenesis". In: *Heart Rhythm* 18.8 (2021), pp. 1423–1434.

[85] Cristobal Rodero, Marina Strocchi, Maciej Marciniak, Stefano Longobardi, John Whitaker, Mark D O'Neill, Karli Gillette, Christoph Augustin, Gernot Plank, Edward J Vigmond, et al. "Linking statistical shape models and simulated function in the healthy adult human heart". In: *PLoS computational biology* 17.4 (2021), e1008851.

[86] Alessandra Ruggiero, Suet Nee Chen, Raffaella Lombardi, Gabriela Rodriguez, and Ali J Marian. "Pathogenesis of hypertrophic cardiomyopathy caused by myozenin 2 mutations is independent of calcineurin activity". In: *Cardiovascular research* 97.1 (2013), pp. 44–54.

[87] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[88] Karoline B Rypdal, A Olav Melleby, Emma L Robinson, Jia Li, Sheryl Palmero, Deborah E Seifert, Daniel Martin, Catelyn Clark, Begoña López, Kristine Andreassen, et al. "ADAMTSL3 knock-out mice develop cardiac dysfunction and dilatation with increased TGFβ signalling after pressure overload". In: *Communications Biology* 5.1 (2022), p. 1392.

[89] Salvatore Santamaria and Rens de Groot. "ADAMTS proteases in cardiovascular physiology and disease". In: *Open Biology* 10.12 (2020), p. 200333.

[90] Panagiotis I Sergouniotis, Adam Diakite, Kumar Gaurav, Ewan Birney, and Tomas Fitzgerald. "Autoencoder-based phenotyping of ophthalmic images highlights genetic loci influencing retinal morphology and provides epidemiologically informative biomarkers." In: *medRxiv* (2023), pp. 2023–06.

[91] Farah Sheikh, Robert C Lyon, and Ju Chen. "Functions of myosin light chain-2 (MYL2) in cardiac muscle and disease". In: *Gene* 569.1 (2015), pp. 14–20.

[92] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. "Deforming autoencoders: Unsupervised disentangling of shape and appearance". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 650–665.

[93] Nona Sotoodehnia, Aaron Isaacs, Paul IW De Bakker, Marcus Dörr, Christopher Newton-Cheh, Ilja M Nolte, Pim Van Der Harst, Martina Müller, Mark Eijgelsheim, Alvaro Alonso, et al. "Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction". In: *Nature genetics* 42.12 (2010), pp. 1068–1076.

[94] JD Steimle and IP Moskowitz. "TBX5: a key regulator of heart development". In: *Current topics in developmental biology* 122 (2017), pp. 195–221.

[95] Thomas M Stokke, Nina E Hasselberg, Marit K Smedsrud, Sebastian I Sarvari, Kristina H Haugaa, Otto A Smiseth, Thor Edvardsen, and Espen W Remme. "Geometry as a confounder when assessing ventricular systolic function: comparison between ejection

fraction and strain". In: *Journal of the American College of Cardiology* 70.8 (2017), pp. 942–954.

[96] Zhongchan Sun, Guang Tong, Nan Ma, Jianying Li, Xiujuan Li, Shuang Li, Jingyu Zhou, Lize Xiong, Feng Cao, Libo Yao, et al. "NDRG2: a newly identified mediator of insulin cardioprotection against myocardial ischemia–reperfusion injury". In: *Basic research in cardiology* 108 (2013), pp. 1–15.

[97] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. "Lossy image compression with compressive autoencoders". In: *International Conference on Learning Representations*. 2016.

[98] Jop H Van Berlo, John W Elrod, Maarten MG Van Den Hoogenhof, Allen J York, Bruce J Aronow, Stephen A Duncan, and Jeffery D Molkentin. "The transcription factor GATA-6 regulates pathological cardiac hypertrophy". In: *Circulation research* 107.8 (2010), pp. 1032–1040.

[99] Ramachandran S Vasan, Nicole L Glazer, Janine F Felix, Wolfgang Lieb, Philipp S Wild, Stephan B Felix, Norbert Watzinger, Martin G Larson, Nicholas L Smith, Abbas Dehghan, et al. "Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data". In: *Jama* 302.2 (2009), pp. 168–178.

[100] Niek Verweij, Jan-Walter Benjamins, Michael P Morley, Yordi J van de Vegte, Alexander Teumer, Teresa Trenkwalder, Wibke Reinhard, Thomas P Cappola, and Pim van der Harst. "The genetic makeup of the electrocardiogram". In: *Cell systems* 11.3 (2020), pp. 229–238.

[101] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[102] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. "Five years of GWAS discovery". In: *The American Journal of Human Genetics* 90.1 (2012), pp. 7–24.

[103] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. "10 years of GWAS discovery: biology, function, and translation". In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22.

[104] Jens-Uwe Voigt and Marta Cvijic. "2-and 3-dimensional myocardial strain in cardiac health and disease". In: *JACC: Cardiovascular Imaging* 12.9 (2019), pp. 1849–1863.

[105] Milos Vukadinovic, Alan C Kwan, Victoria Yuan, Michael Salerno, Daniel C Lee, Christine M Albert, Susan Cheng, Debiao Li, David Ouyang, and Shoa L Clarke. "Deep learning-enabled analysis of medical images identifies cardiac sphericity as an early marker of cardiomyopathy and related outcomes". In: *Med* 4.4 (2023), pp. 252–262.

[106] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al. "Opportunities and challenges for transcriptome-wide association studies". In: *Nature genetics* 51.4 (2019), pp. 592–599.

[107] Simon G Williams, Dominic JF Byrne, and Bernard D Keavney. "Rare GATA6 variants associated with risk of congenital heart disease phenotypes in 200,000 UK Biobank exomes". In: *Journal of human genetics* 67.2 (2022), pp. 123–125.

[108] Yan Xia, Xiang Chen, Nishant Ravikumar, Christopher Kelly, Rahman Attar, Nay Aung, Stefan Neubauer, Steffen E Petersen, and Alejandro F Frangi. "Automatic 3D+t Four-Chamber CMR Quantification of the UK Biobank: integrating imaging and non-imaging data priors at scale". In: *Medical Image Analysis* (2022), p. 102498.

[109] Fan Xiang, Yasuhiko Sakata, Lei Cui, Joey M Youngblood, Hironori Nakagami, James K Liao, Ronglih Liao, and Michael T Chin. "Transcription factor CHF1/Hey2 suppresses cardiac hypertrophy through an inhibitory interaction with GATA4". In: *American Journal of Physiology-Heart and Circulatory Physiology* 290.5 (2006), H1997–H2006.

[110] Ziqian Xie, Tao Zhang, Sangbae Kim, Jiaxiong Lu, Wanheng Zhang, Cheng-Hui Lin, Man-Ru Wu, Alexander Davis, Roomasa Channa, Luca Giancardo, et al. "iGWAS: image-based genome-wide association of self-supervised deep phenotyping of human medical images". In: *medRxiv* (2022), pp. 2022–05.

[111] Ying-Jia Xu, Xing-Biao Qiu, Fang Yuan, Hong-Yu Shi, Lei Xu, Xu-Min Hou, Xin-Kai Qu, Xu Liu, Ri-Tai Huang, Song Xue, et al. "Prevalence and spectrum of NKX2. 5 mutations in patients with congenital atrial septal defect and atrioventricular block". In: *Molecular medicine reports* 15.4 (2017), pp. 2247–2254.

[112] Mengyao Yu, Catherine Tcheandjieu, Adrien Georges, Ke Xiao, Helio Tejeda, Christian Dina, Thierry Le Tourneau, Madalina Fiterau, Renae Judy, Noah L Tsao, et al. "Computational estimates of annular diameter reveal genetic determinants of mitral valve function and disease". In: *JCI insight* 7.3 (2022).

[113] Manling Zhang, Norimichi Koitabashi, Takahiro Nagayama, Ryan Rambaran, Ning Feng, Eiki Takimoto, Trisha Koenke, Brian O'Rourke, Hunter C Champion, Michael T Crow, et al. "Expression, activity, and pro-hypertrophic effects of PDE5A in cardiac myocytes". In: *Cellular signalling* 20.12 (2008), pp. 2231–2236.

[114] Shaosong Zhang, Ned Watson, Joseph Zahner, Jeffrey N Rottman, Kendall J Blumer, and Anthony J Muslin. "RGS3 and RGS4 are GTPase activating proteins in the heart". In: *Journal of molecular and cellular cardiology* 30.2 (1998), pp. 269–276.

Appendix

This appendix contains supplementary material for chapter 7, with significant or suggestive associations for chambers: right ventricle (RV), two ventricles (BV), left atrium (LA) and right atrium (RA).

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
6	117672972-118963115	PLN	5	6.2×10^{-16}	rs12661338	C	A	44.9	-5.67 ± 0.7
17	43056905-45876022	GOSR2	5	1.1×10^{-13}	rs17677363	A	T	13.2	8.41 ± 1.01
2	178553183-181312739	TTN	5	1.9×10^{-12}	rs10497529	G	A	3.6	-7.78 ± 1.84
12	110336719-113263518	MYL2	4	2.7×10^{-15}	rs35350651	A	AC	51.4	-2.77 ± 0.69
1	14891511-16897730	HSPB7	4	2.1×10^{-12}	rs1627145	C	T	65.7	5.13 ± 0.73
3	99373762-100592217	FILIP1L*	3	1.1×10^{-09}	rs9865713	G	A	37.9	3.37 ± 0.72
2	36122006-38132712	STRN	3	2.2×10^{-10}	rs2013223	T	C	58.7	-4.26 ± 0.67
10	120591353-122407323	BAG3	3	9.2×10^{-11}	rs72840788	G	A	21.5	-5.52 ± 0.85
14	23018665-24905123	MYH6	3	3.7×10^{-11}	rs376439	A	G	39.5	-2.34 ± 0.68
3	157312028-159477890	SHOX2	3	8.2×10^{-12}	rs56148815	G	C	40.1	4.42 ± 0.74
12	27799773-29651255	CCDC91*	3	8.9×10^{-14}	12:28491249_GT_G	GT	G	25.9	6.03 ± 0.81
16	4001196-5118345	SRL	2	3.7×10^{-10}	rs868302	A	G	23.5	4.71 ± 0.83
13	93586455-96087558	DCT*	2	4.3×10^{-10}	rs9524217	G	A	29.6	-4.75 ± 0.76
7	118351581-121045273	WNT16	2	1.2×10^{-09}	rs3779381	A	G	26.3	2.87 ± 0.79
15	48136048-50008043	FBN1	2	6.2×10^{-09}	rs61999107	C	T	10.3	5.31 ± 1.17
18	33861964-35075250	FHOD3	2	6.6×10^{-09}	rs34617432	C	CA	27.1	2.79 ± 0.8
5	171074292-172678327	NKX2-5	1	1.1×10^{-10}	rs6882776	G	A	28.6	-3.14 ± 0.78

Table A1: GWAS results for static variables for RV.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
2	178553183-181312739	TTN	5	4.4×10^{-18}	rs17362588	G	A	8.7	8.61 ± 1.21
14	23018665-24905123	MYH6	5	1.7×10^{-13}	rs365990	A	G	36.9	4.22 ± 0.72
12	23820634-25371083	BCAT1	4	3.8×10^{-23}	rs4963772	G	A	15.0	-5.9 ± 0.97
5	171074292-172678327	NKX2-5	4	3.2×10^{-18}	5:172643468_GC_G	GC	G	16.4	6.15 ± 0.96
20	34960446-36909530	KIAA1755	4	3.9×10^{-17}	rs2881138	A	G	46.2	-3.09 ± 0.7
12	33076989-37856717	SYT10	4	5.4×10^{-15}	rs11052736	T	C	45.6	3.13 ± 0.7
15	73628714-76398624	HCN4	4	1.1×10^{-12}	rs7172796	T	G	15.9	-6.77 ± 0.95
6	6785207-7808936	DSP	4	1.5×10^{-09}	rs72825054	G	C	9.7	7.15 ± 1.18
14	71131957-72889615	RGS6	4	4.0×10^{-12}	rs17180489	G	C	14.5	4.3 ± 0.93
12	1080331-2544786	CACNA1C	4	1.3×10^{-11}	rs2283274	G	C	17.8	-4.32 ± 0.92
7	100196651-101199253	EPHB4	3	2.9×10^{-10}	rs9691107	T	C	40.2	-3.62 ± 0.72
3	157312028-159477890	SHOX2	3	8.9×10^{-09}	rs6792449	T	C	50.2	-2.82 ± 0.71
4	100678360-103221356	SLC39A8	3	1.4×10^{-09}	rs35225200	A	C	7.9	5.68 ± 1.32
10	120591353-122407323	BAG3	3	7.0×10^{-10}	rs2234962	T	C	21.5	-5.49 ± 0.84
14	19002084-21589402	NDRG2	3	3.4×10^{-10}	rs12889267	A	G	16.7	-3.97 ± 0.93
1	14891511-16897730	HSPB7	3	1.0×10^{-10}	rs12744578	A	T	28.6	3.6 ± 0.77
17	43056905-45876022	GOSR2	3	1.5×10^{-10}	rs1358071	C	A	73.5	-4.9 ± 0.78
6	117672972-118963115	PLN	3	6.1×10^{-19}	rs3951016	T	A	47.1	-6.23 ± 0.7
6	120512128-121905676	GJA1	3	1.0×10^{-11}	rs34782269	G	GA	50.1	2.99 ± 0.71
3	38356116-40221298	SCN5A	3	6.4×10^{-14}	rs6795970	A	G	60.3	-5.33 ± 0.71
8	125683719-126410917	MTSS1	2	1.9×10^{-11}	8:125858538_GA_G	GA	G	31.0	3.07 ± 0.76
22	24984204-26791628	MYO18B	2	1.0×10^{-09}	rs133890	C	G	45.2	-3.42 ± 0.7
11	63804569-65898631	LTBP3	2	9.6×10^{-09}	rs2096560	T	C	23.5	4.74 ± 0.82
3	13070799-14816900	TMEM43	2	1.5×10^{-08}	rs73133428	C	T	18.5	4.97 ± 0.88
2	231843389-233550003	HTR2B	2	2.3×10^{-08}	rs12993290	G	A	22.1	-4.14 ± 0.84
9	126971887-129059665	SLC27A4*	2	2.5×10^{-08}	rs473426	G	C	54.0	3.89 ± 0.7
12	110336719-113263518	MYL2	2	3.1×10^{-08}	rs58235019	A	AG	60.5	-3.91 ± 0.71
1	235819436-237555628	ACTN2	1	2.9×10^{-10}	rs4659701	G	A	62.2	2.93 ± 0.73
12	115036602-115503216	TBX3	1	2.2×10^{-12}	rs7309382	T	C	40.2	2.57 ± 0.71

Table A2: GWAS results for dynamic variables for RV.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
6	117672972-118963115	PLN	5	8.2×10^{-27}	rs72967533	T	C	47.7	5.29 ± 0.7
12	110336719-113263518	MYL2	5	8.7×10^{-17}	rs35350651	A	AC	51.4	-2.77 ± 0.69
2	178553183-181312739	TTN	5	2.5×10^{-17}	rs10497529	G	A	3.6	-7.78 ± 1.84
1	14891511-16897730	HSPB7	5	2.8×10^{-14}	rs1763619	A	G	65.7	-4.67 ± 0.73
14	23018665-24905123	MYH6	5	7.9×10^{-13}	rs365990	A	G	36.9	4.22 ± 0.72
10	120591353-122407323	BAG3	5	1.1×10^{-12}	rs72840788	G	A	21.5	-5.52 ± 0.85
3	13070799-14816900	TMEM43	5	6.8×10^{-12}	rs73028848	A	G	34.0	5.08 ± 0.74
3	157312028-159477890	SHOX2	5	1.5×10^{-11}	rs56148815	G	C	40.1	4.42 ± 0.74
4	119933512-120392684	MYOZ2	4	6.3×10^{-10}	rs28634456	A	G	29.9	-4.73 ± 0.77
2	36122006-38132712	STRN	3	4.3×10^{-11}	rs2110944	T	C	52.6	-4.04 ± 0.7
16	4001196-5118345	SRL	3	2.3×10^{-09}	rs868302	A	G	23.5	4.71 ± 0.83
1	226810860-229156248	OBSCN	3	9.0×10^{-09}	rs883748	C	T	47.6	2.94 ± 0.66
7	118351581-121045273	WNT16	3	7.0×10^{-15}	rs3779381	A	G	26.3	2.87 ± 0.79
17	43056905-45876022	GOSR2	3	1.0×10^{-15}	rs17677363	A	T	13.2	8.41 ± 1.01
17	67858770-69387817	KCNJ2	2	6.5×10^{-09}	rs180063	A	G	25.8	-3.46 ± 0.84
6	125424383-127540461	HEY2	2	2.7×10^{-09}	rs373980643	GT	G	46.0	3.22 ± 0.7
3	99373762-100592217	FILIP1L*	2	2.0×10^{-09}	rs7636776	T	C	41.2	4.24 ± 0.71
3	72529329-74321817	PDZRN3	2	6.7×10^{-10}	rs7631905	C	T	47.3	4.14 ± 0.7
6	1452362-2458936	GMDS	2	3.9×10^{-08}	rs6934958	T	C	54.9	2.87 ± 0.7
12	113986709-115036602	TBX5	2	4.2×10^{-10}	rs1895606	C	T	46.5	2.86 ± 0.7
16	74971503-75977954	BCAR1	2	8.4×10^{-11}	rs12446877	T	C	62.7	3.75 ± 0.72
12	27799773-29651255	CCDC91*	2	1.1×10^{-11}	rs10843115	C	T	27.4	5.31 ± 0.78
14	19002084-21589402	NDRG2	1	4.6×10^{-10}	rs12889267	A	G	16.7	-3.97 ± 0.93
17	63148128-64800430	PRKCA	1	2.9×10^{-10}	rs9892651	C	T	58.1	4.46 ± 0.71
15	99244059-100636847	LRRC28*	1	4.6×10^{-11}	rs2871974	C	T	63.5	3.68 ± 0.73
1	182755356-184595513	SMG7	1	7.7×10^{-11}	rs789185	A	G	39.3	-3.42 ± 0.72

Table A3: GWAS results for static variables for BV.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
14	23018665-24905123	MYH6	5	3.7×10^{-17}	rs422068	T	C	35.9	-3.25 ± 0.72
20	34960446-36909530	KIAA1755	5	3.3×10^{-15}	rs2881138	A	G	46.2	-3.09 ± 0.7
3	38356116-40221298	SCN5A	5	8.4×10^{-11}	rs6783110	A	G	59.9	4.64 ± 0.71
5	171074292-172678327	NKX2-5	5	5.8×10^{-17}	5:172643468_GC_G	GC	G	16.4	6.15 ± 0.96
14	71131957-72889615	RGS6	5	9.5×10^{-12}	rs17180489	G	C	14.5	4.3 ± 0.93
10	120591353-122407323	BAG3	5	1.5×10^{-12}	rs72840788	G	A	21.5	-5.52 ± 0.85
12	33076989-37856717	SYT10	5	1.7×10^{-14}	12:33654815_CA_C	CA	C	54.3	-3.74 ± 0.7
12	23820634-25371083	BCAT1	5	9.6×10^{-17}	rs4963772	G	A	15.0	-5.9 ± 0.97
6	117672972-118963115	PLN	5	7.6×10^{-17}	rs72967533	T	C	47.7	5.29 ± 0.7
2	178553183-181312739	TTN	5	7.3×10^{-17}	rs17362588	G	A	8.7	8.61 ± 1.21
18	33861964-35075250	FHOD3	4	8.1×10^{-15}	rs2644262	T	C	27.5	4.67 ± 0.78
12	115036602-115503216	TBX3	4	6.2×10^{-14}	rs7300371	T	C	73.1	4.56 ± 0.79
12	1080331-2544786	CACNA1C	4	6.8×10^{-12}	rs2283274	G	C	17.8	-4.32 ± 0.92
1	14891511-16897730	HSPB7	4	1.0×10^{-09}	1:16141512_CTTT_C	CTTT	C	37.6	3.04 ± 0.69
1	5913893-7247335	PLEKHG5	4	7.0×10^{-10}	1:6283547_GA_G	GA	G	15.2	-5.87 ± 0.99
17	43056905-45876022	GOSR2	3	3.7×10^{-10}	rs1358071	C	A	73.5	-4.9 ± 0.78
14	19002084-21589402	NDRG2	3	2.0×10^{-09}	rs12889267	A	G	16.7	-3.97 ± 0.93
6	6785207-7808936	DSP	3	1.2×10^{-11}	rs6920534	T	C	9.9	8.0 ± 1.18
1	235819436-237555628	ACTN2	3	2.1×10^{-15}	rs12724121	A	T	62.2	-2.56 ± 0.68
13	20686720-22242174	FGF9	3	2.2×10^{-08}	rs4990057	C	T	42.8	2.96 ± 0.71
15	73628714-76398624	HCN4	2	8.4×10^{-10}	rs11630367	A	G	15.8	-4.83 ± 0.95
22	24984204-26791628	MYO18B	2	9.1×10^{-10}	rs133885	G	A	45.2	-3.21 ± 0.7
1	206073265-208410364	CD34	2	1.4×10^{-10}	rs861475	T	C	53.0	2.62 ± 0.71
5	136376050-139265072	SLC23A1	2	1.3×10^{-08}	rs74803899	T	C	5.5	7.46 ± 1.53
12	113986709-115036602	TBX5	2	1.4×10^{-08}	rs883079	C	T	72.9	3.71 ± 0.79
3	168580960-170964909	SAMD7*	1	3.2×10^{-10}	rs11383131	A	AC	60.7	3.16 ± 0.71
4	174264132-176570716	HAND2	1	2.2×10^{-10}	rs10005540	C	T	61.7	2.99 ± 0.72

Table A4: GWAS results for dynamic variables for BV.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
4	111256567-113870102	PITX2	4	7.7×10^{-09}	rs2723321	A	G	69.6	-4.4 ± 0.76
2	178553183-181312739	TTN	3	2.3×10^{-14}	rs17362588	G	A	8.7	8.61 ± 1.21
17	43056905-45876022	GOSR2	3	1.8×10^{-11}	rs533030436	A	G	11.8	5.07 ± 1.12
4	100678360-103221356	SLC39A8	2	1.2×10^{-09}	rs13107325	C	T	7.0	-5.29 ± 1.35
2	54685226-56203345	EFEMP1	2	10.0×10^{-09}	rs147444949	A	G	1.3	11.26 ± 2.86

Table A5: GWAS results for static variables for LA.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
2	178553183-181312739	TTN	5	1.6×10^{-15}	rs17362588	G	A	8.7	8.61 ± 1.21
5	171074292-172678327	NKX2-5	5	5.4×10^{-12}	rs11134777	G	T	16.5	-6.56 ± 0.95
6	117672972-118963115	PLN	5	1.9×10^{-11}	rs72967533	T	C	47.7	5.29 ± 0.7
12	23820634-25371083	BCAT1	3	3.9×10^{-18}	rs4963772	G	A	15.0	-5.9 ± 0.97
7	100196651-101199253	EPHB4	3	2.4×10^{-10}	rs2720392	C	A	49.6	3.11 ± 0.7
16	74971503-75977954	BCAR1	3	2.8×10^{-16}	rs59686216	A	G	60.0	3.79 ± 0.72
14	23018665-24905123	MYH6	3	6.5×10^{-15}	rs365990	A	G	36.9	4.22 ± 0.72
12	33076989-37856717	SYT10	3	3.0×10^{-14}	rs11052736	T	C	45.6	3.13 ± 0.7
12	1080331-2544786	CACNA1C	3	3.6×10^{-13}	rs2283274	G	C	17.8	-4.32 ± 0.92
14	71131957-72889615	RGS6	3	2.8×10^{-12}	rs17180489	G	C	14.5	4.3 ± 0.93
12	65559695-67181144	HMGA2	3	1.5×10^{-08}	rs761210718	AAG	A	36.9	-3.14 ± 0.72
20	34960446-36909530	KIAA1755	3	9.4×10^{-17}	rs2881138	A	G	46.2	-3.09 ± 0.7
1	14891511-16897730	HSPB7	2	1.9×10^{-09}	rs12138117	T	C	32.2	-4.45 ± 0.74
6	120512128-121905676	GJA1	2	9.2×10^{-10}	rs34782269	G	GA	50.1	2.99 ± 0.71
15	73628714-76398624	HCN4	2	7.9×10^{-10}	rs11630367	A	G	15.8	-4.83 ± 0.95
16	14464002-16154060	MYH11	2	4.5×10^{-11}	rs9284324	G	A	32.0	-3.25 ± 0.76
18	33861964-35075250	FHOD3	2	2.3×10^{-10}	rs2848901	C	T	28.5	3.45 ± 0.77
10	120591353-122407323	BAG3	2	1.4×10^{-10}	rs7095308	G	A	27.5	4.42 ± 0.78
17	43056905-45876022	GOSR2	2	1.2×10^{-10}	rs76774446	C	A	13.2	-6.16 ± 1.02
6	6785207-7808936	DSP	2	1.2×10^{-10}	rs6920534	T	C	9.9	8.0 ± 1.18
3	13070799-14816900	TMEM43	2	2.2×10^{-08}	rs73031103	A	G	34.0	4.33 ± 0.74
7	126869221-128778386	FLNC	1	4.7×10^{-12}	rs4472439	G	A	11.2	7.65 ± 1.11

Table A6: GWAS results for dynamic variables for LA.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
17	43056905-45876022	GOSR2	5	1.6×10^{-18}	rs11874	G	A	13.3	4.99 ± 1.01
12	27799773-29651255	CCDC91*	3	1.4×10^{-10}	rs200507341	C	CT	20.1	3.58 ± 0.89
2	178553183-181312739	TTN	2	2.4×10^{-12}	rs151041685	G	T	8.7	-7.0 ± 1.22
6	117672972-118963115	PLN	2	4.4×10^{-10}	rs11153730	T	C	48.6	3.8 ± 0.69
12	65559695-67181144	HMGA2	2	3.7×10^{-09}	rs761210718	AAG	A	36.9	-3.14 ± 0.72
14	23018665-24905123	MYH6	1	1.2×10^{-10}	rs422068	T	C	35.9	-3.25 ± 0.72

Table A7: GWAS results for static variables for RA.

chr.	region	candidate gene	count	min. p-value	SNP	NEA	EA	EAF	$\hat{\beta} \pm \text{se}(\hat{\beta})(\times 100)$
12	23820634-25371083	BCAT1	4	3.3×10^{-21}	rs4963772	G	A	15.0	-5.9 ± 0.97
2	178553183-181312739	TTN	4	4.3×10^{-18}	rs10497529	G	A	3.6	-7.78 ± 1.84
20	34960446-36909530	KIAA1755	4	8.1×10^{-16}	rs2881138	A	G	46.2	-3.09 ± 0.7
5	171074292-172678327	NKX2-5	4	6.2×10^{-15}	rs539079213	C	CT	15.5	-3.83 ± 1.02
12	33076989-37856717	SYT10	4	7.3×10^{-15}	rs11052736	T	C	45.6	3.13 ± 0.7
14	23018665-24905123	MYH6	3	4.9×10^{-12}	rs376439	A	G	39.5	-2.34 ± 0.68
6	6785207-7808936	DSP	3	3.0×10^{-09}	rs112019128	C	T	9.6	-6.23 ± 1.19
17	43056905-45876022	GOSR2	3	2.4×10^{-10}	rs9896243	C	G	21.0	5.37 ± 0.85
12	1080331-2544786	CACNA1C	3	1.5×10^{-10}	rs2283274	G	C	17.8	-4.32 ± 0.92
14	71131957-72889615	RGS6	3	1.8×10^{-12}	rs17180489	G	C	14.5	4.3 ± 0.93
3	38356116-40221298	SCN5A	3	1.4×10^{-19}	rs6801957	T	C	59.7	-3.92 ± 0.67
15	73628714-76398624	HCN4	3	1.2×10^{-12}	rs7172796	T	G	15.9	-6.77 ± 0.95
6	120512128-121905676	GJA1	2	2.8×10^{-10}	rs34782269	G	GA	50.1	2.99 ± 0.71
1	5913893-7247335	PLEKHG5	2	6.8×10^{-10}	rs796136824	GA	G	23.2	4.62 ± 0.93
6	117672972-118963115	PLN	2	1.1×10^{-12}	rs57912492	C	T	41.9	-2.9 ± 0.71
1	206073265-208410364	CD34	2	1.1×10^{-08}	rs2785647	G	A	34.3	3.95 ± 0.73
7	100196651-101199253	EPHB4	1	3.4×10^{-10}	rs9691107	T	C	40.2	-3.62 ± 0.72

Table A8: GWAS results for dynamic variables for RA.

Publications

- **R. Bonazzola**, E. Ferrante, N. Ravikumar, Y. Xia, T. Syeda-Mahmood, A.F. Frangi. *Learning disentangled mesh representations for cardiac genetic discovery: from static to dynamic biomarkers.* Manuscript in preparation.
- **R. Bonazzola**, E. Ferrante, N. Ravikumar, Y. Xia, B. Keavney, S. Plein, T. Syeda-Mahmood, A.F. Frangi. *Unsupervised ensemble-based phenotyping enhances discoverability of genes related to left-ventricular morphology.* Nature Machine Intelligence (2024).
- **R. Bonazzola**, N. Ravikumar, R. Attar, E. Ferrante, T. Syeda-Mahmood, A. F Frangi. *Image-derived phenotype extraction for genetic discovery via unsupervised deep learning in CMR images.* International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 699-708 (2021).
- Y. Deo, **R. Bonazzola**, H. Dou, Y. Xia, T. Wei, N. Ravikumar, A. F. Frangi, T. Lassila. *Learned local attention maps for synthesising vessel segmentations from T_2 MRI.* International Workshop on Simulation and Synthesis in Medical Imaging. 32-41 (2023).
- C. Maldonado García, **R. Bonazzola**, N. Ravikumar, A.F. Frangi. *Predicting Myocardial Infarction Using Retinal OCT Imaging.* Annual Conference on Medical Image Understanding and Analysis, 787-797 (2022).
- N. Cheng, **R. Bonazzola**, N. Ravikumar, A.F. Frangi. *A Generative Framework for Predicting Myocardial Strain from Cine-Cardiac Magnetic Resonance Imaging.* Annual Conference on Medical Image Understanding and Analysis, 482-493 (2022).

Acknowledgements

First and foremost, I would like to express my gratitude to my family, who have always been my safe place and whose unwavering support has been the foundation of this journey. None of this would have been possible without their constant presence.

To my supervisory team: Alejandro Frangi, for his guidance and for establishing an environment in which this research could take place; Enzo Ferrante, for his guidance, insightful discussions, and mentorship –particularly for helping me recognize strengths in my work that I might have otherwise missed; Nishant Ravikumar, for his supervision throughout this process and for stepping in as my main supervisor when Alex moved to Manchester in 2023; and Tanveer Syeda-Mahmood, for her supervision and for securing funding that allowed this journey to begin.

To my colleagues and friends from my lab, the School of Computer Science (then School of Computing), and the University for making my experience in Leeds far more enriching and fun. Especially Cynthia, Nurbanu, and Udayraj for their close friendship and support. Also, to Yash, Lucy, Ning, Soodeh, Haoran and other lab colleagues for numerous shared moments. To Kattia and Alejandro for their support and for bringing homeliness to my stay in Leeds.

To my friends in Santa Fe –Cacho, Juan, Fernando and Miguel– for always being there in spite of the distance, for believing in me, and for their constant encouragement. They have never ceased to show that my friendship with them was more than justified.

To someone far away in space and time, yet close to heart: Mot E. K.