

Imaging-anchored multiomics in cardiovascular disease: integrating cardiac imaging, bulk, single-cell, and spatial transcriptomics

Minh H. N. Le¹, Thanh-Huy Nguyen², Tao Li³, Bao Quang Gia Le⁴, Han H. Huynh⁵, Monika Raj⁴, Carl Yang³, Min Xu², Tuan Vinh^{6,*}, Nguyen Quoc Khanh Le^{7,8,9,*}

¹Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06510, United States

²Computational Biology Department, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States.

³Department of Computer Science, Emory University, 201 Dowman Drive, Atlanta, GA 30322, United States

⁴Department of Chemistry, Emory University, 1515 Dickey Drive, Atlanta, GA 30322, United States

⁵Irell and Manella Graduate School of Biological Science, City of Hope, CA, United States

⁶Medical Sciences Division, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU, Oxfordshire, United Kingdom

⁷AI BioMed Research Group, Taipei Medical University, 250 Wu-Hsing Street, Xinyi District, Taipei 11031, Taiwan

⁸In-Service Master Program in Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, 250 Wu-Hsing Street, Xinyi District, Taipei 11031, Taiwan

⁹Translational Imaging Research Center, Taipei Medical University Hospital, 252 Wu-Hsing Street, Xinyi District, Taipei 11031, Taiwan

*Corresponding authors. Tuan Vinh, Medical Sciences Division, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU, Oxfordshire, United Kingdom.

E-mail: tuan.vinh@hertford.ox.ac.uk; Nguyen Quoc Khanh Le, In-Service Master Program in Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan.

E-mail: khanhlee@tmu.edu.tw.

Abstract

Cardiovascular disease arises from interactions between inherited risk, molecular programmes, and tissue-scale remodelling that are observed clinically through imaging. Cardiac MRI (CMR), computed tomography (CT), and echocardiography are integral to routine cardiovascular care, while bulk RNA sequencing, single-cell RNA sequencing, and spatial transcriptomics are providing increasingly detailed molecular characterization of cardiac tissue. Yet, these imaging and molecular data are still analysed in largely separate pipelines. This review examines joint representations that link cardiac imaging phenotypes to transcriptomic and spatially resolved molecular states. An imaging-anchored perspective is adopted in which echocardiography, CMR, and CT define a spatial phenotype of the heart, and bulk, single-cell and spatial transcriptomics provide cell-type- and location-specific molecular context. We define the representation requirements of each modality, compare multimodal fusion strategies, and synthesize integrative pipelines for radiogenomics, spatial alignment, and image-based gene-expression prediction, together with their validation requirements, limitations, and failure modes. Spatial multiomic maps of human myocardium and atherosclerotic plaque, together with single-cell, spatial, and multimodal medical foundation models, are advancing imaging-anchored multiomics; however, cost, scalability, and tissue availability remain substantial barriers to large-scale cardiovascular translation.

Keywords cardiovascular imaging, multiomics integration, spatial transcriptomics, radiogenomics, foundation models, artificial intelligence

Introduction

Cardiovascular disease (CVD) remains the leading global cause of mortality and disability despite major advances in prevention and treatment [1]. Guideline-directed workflows and classical risk models summarize a patient's state using a small number of imaging and clinical measurements that are robust but coarse.

At the same time, modern cardiology increasingly operates in a multiomics environment: bulk RNA-seq and proteomics are scalable to population-level cohorts, whereas sc/snRNA-seq [2] and spatial transcriptomics remain confined to smaller studies in which tissue availability [3], sampling region, and processing infrastructure impose additional constraints.

Received: December 13, 2025. **Revised:** March 16, 2026. **Accepted:** May 4, 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

This review focuses on *imaging-anchored molecular phenotyping*.

Cardiac imaging describes anatomy and function in space and time, whereas transcriptomic readouts, from bulk RNA-seq through single-cell profiling to spatially resolved assays, capture the gene-expression programmes that drive or accompany structural change. These two data layers are still largely analysed in separate pipelines. Linking them could allow imaging features such as late gadolinium enhancement (LGE), regional strain, or plaque morphology to be interpreted in terms of cell states, pathways, and tissue niches, and may ultimately enable images to serve as non-invasive surrogates for molecular assays when biopsies or spatial omics are infeasible. We survey current computational approaches toward this goal, including radiogenomic association, spatial transcriptomics alignment, and image-based prediction of gene expression. Imaging is treated as the anchor modality for three practical reasons.

First, cardiac imaging is ubiquitous in clinical care, providing a shared coordinate system and a common language across sites.

Second, imaging captures the integrated result of genetics, haemodynamics, inflammation, and tissue remodelling, offering holistic but indirect readouts of biology.

Third, imaging can be registered, approximately *in vivo* and more precisely *ex vivo*, to histology and spatial transcriptomics, enabling cross-scale integration from voxel or segment to cell.

Several recent reviews survey machine learning for cardiac imaging [4, 5], multimodal modelling for coronary artery disease, cardiovascular risk prediction [6], spatial technologies in cardiac biology [7], and medical foundation models that integrate imaging, text and clinical data [8, 9]. Here, we position cardiac imaging as the primary spatial reference frame, spatial transcriptomics as the bridge from voxel to cell, and representation learning and multimodal fusion as the computational machinery for building joint latent spaces for mechanistic interpretation, disease subtyping, and cardiovascular translation.

Scope and literature identification strategy

This article is a narrative, methods-focused review rather than a systematic meta-analysis. To increase transparency and reproducibility, we used a structured identification strategy. Searches were performed across PubMed, Web of Science, Scopus, and IEEE Xplore, complemented by screening relevant preprint servers for recent methodological developments. Search terms were organized along four axes: (i) imaging modalities and phenotypes (echocardiography, CMR, CT/CTA, radiomics, fibrosis, strain, and plaque), (ii) molecular modalities (bulk RNA-seq, sc/snRNA-seq, spatial transcriptomics, and proteomics), (iii) integration tasks (radiogenomics, spatial registration/alignment, cross-modal prediction, or “virtual” transcriptomics) and (iv) computational paradigms (fusion strategies, representation learning, contrastive learning, graph neural networks (GNNs), transformers, and foundation models). Backward and forward citation chasing from exemplar studies and existing surveys identified additional relevant work, and seminal machine-learning papers were included where needed for technical completeness. We prioritized studies that (i) link imaging phenotypes to molecular readouts in cardiovascular settings, (ii) introduce multimodal or spatial integration methods directly applicable to imaging-omics problems, and/or (iii) provide clinically oriented validation relevant to biomarker discovery or risk prediction.

Interdependent conditions and multimorbidity

Cardiovascular phenotypes rarely occur in isolation. Heart failure frequently co-exists with atrial fibrillation, chronic kidney disease, diabetes, obesity, and inflammatory conditions, which can share upstream molecular drivers while altering imaging phenotypes through haemodynamic loading, neurohormonal signalling, and systemic inflammation. Imaging-anchored multiomics, therefore, benefits from modelling *interdependent conditions* rather than single labels: multi-task prediction, joint latent spaces shared across related endpoints, and causal frameworks that separate shared biology from condition-specific effects can reduce confounding and improve generalizability across heterogeneous clinical cohorts.

The remainder of the review is organized into four methodological layers, beginning with the modalities and biological context. Spatial transcriptomics bridges imaging-scale phenotypes and cellular molecular states, but it must span a fundamental resolution gap.

In vivo cardiac MRI and CT operate at millimetre-to-centimetre voxel scales, whereas current spatial transcriptomics platforms interrogate tissue at the micron-to-spot level: the widely used 10× Visium captures transcripts at 55 μ m spots with 100 μ m centre-to-centre spacing, next-generation platforms such as Visium HD and Stereo-seq achieve 2–8 μ m bin resolution approaching single-cell scale, and imaging-based systems such as Xenium and MERSCOPE resolve individual cells or subcellular compartments [10, 11].

Bridging this gap requires deliberate experimental design at two levels. At the macroscopic level, the principal strategy is *imaging-guided tissue sampling*: tissue blocks are dissected from regions predefined by *in vivo* imaging findings—such as LGE-positive versus LGE-negative myocardial segments, or histologically characterized infarct core, border zone, and remote myocardium—establishing a coarse but intentional correspondence between the imaging phenotype and the spatial molecular map [12, 13]. Real-time CMR-guided endomyocardial biopsy can further improve targeting accuracy by directing sampling to LGE-defined lesions rather than relying on anatomical landmarks alone [14, 15]. At the microscopic level, where *ex vivo* tissue is available, high-resolution *ex vivo* MRI or micro-CT serves as an intermediate registration bridge between *in vivo* imaging and the histology section on which spatial transcriptomics is performed. Despite these strategies, fine-grained, voxel-level alignment between *in vivo* imaging and spatial transcriptomic spots remains an open methodological challenge, limited by tissue deformation during sectioning, scale mismatch across modalities, and the absence of comprehensive spatial reference atlases of the human heart [11, 12].

Key public resources that combine cardiovascular imaging with molecular, waveform, or rich clinical data are summarized in Table 1, and representative imaging biomarkers that commonly serve as anchors (and their likely molecular correlates) are summarized in Table 2. The overall imaging-anchored multiomics framework is illustrated in Fig. 1.

Foundations: modalities and biological context

Cardiac imaging as a spatial phenotype

Cardiac imaging provides a high-level description of heart structure and function that is naturally indexed in space (and often time). Echocardiography offers real-time 2D and 3D views at the bedside, and resources such as EchoNet-Dynamic and EchoNet-LVH have accelerated deep learning for view classification, segmentation, and disease

Table 1 Public multimodal cardiology datasets relevant to imaging-anchored multiomics, including resources that directly pair cardiovascular imaging with molecular data and adjacent multimodal cardiovascular resources useful for encoder pretraining and transfer learning.

Dataset/study	Type	Cardiovascular focus	Modalities included	Approximate scale and access	Link
UK Biobank imaging extension [17]	Population imaging cohort	Cardiac structure, function, and risk in mid-life adults	CMR, vascular ultrasound and other MRI sequences linked to genome-wide genotyping, exome sequencing, plasma proteomics, metabolomics, EHR, and questionnaires	~100,000 participants with CMR and >500,000 with genomics; application-based access	UKBiobank
Multi-Ethnic Study of Atherosclerosis (MESA) [31]	Population cohort with multimodal imaging	Subclinical atherosclerosis and incident CVD in a diverse community sample	CMR, coronary artery calcium CT, carotid ultrasound, and other imaging linked to genotype, labs, and risk-factor profiling	~6,800 participants; controlled access via NHLBI repositories	BioLINCC
Hypertrophic Cardiomyopathy Registry (HCMR) [32]	Disease registry and biobank	Risk stratification in hypertrophic cardiomyopathy	CMR, echocardiography and electrocardiography (ECG) combined with targeted gene sequencing and circulating biomarkers	~2,700 patients; collaborative access via data-use agreements	HCMR
MIMIC-IV and related PhysioNet modules [33]	Critical-care waveform and EHR resource	Acute CVD and haemodynamic instability in ICU/ED settings	Twelve-lead ECG waveforms linked to high-resolution EHR; additional echo and note modules for subsets	~800,000 ECGs in ~160,000 subjects; credentialed access via PhysioNet	PhysioNet
EchoNet family (EchoNet-Dynamic, EchoNet-LVH) [5, 16]	Curated echocardiography video datasets	LV systolic function, wall motion, and hypertrophy phenotypes	Apical four-chamber echo videos with contour tracings, volumetric measurements, and labels (EF, LVH)	On the order of 10 ⁴ videos; de-identified datasets downloadable for research	EchoNet
Cardiac Atlas Project CMR cohorts [34]	Aggregated CMR repositories	Cardiac structure and function in health and early disease	3D cine CMR studies with segmentations, finite-element meshes and clinical covariates in selected cohorts	Thousands of CMR exams; access under data-use agreements	CardiacAtlas
Spatial multi-omic map of human myocardial infarction [12]	Human tissue atlas with single-cell and spatial multiomics	Early and late remodelling after acute myocardial infarction	snRNA-seq, snATAC-seq, Visium spatial transcriptomics, and matched histology across infarct core, border zone, and remote myocardium	31 samples from 23 patients; processed data via public portals and archives	HCAportal
Coronary plaque spatial transcriptomics studies [29, 35, 36]	Plaque-level spatial-omics studies	Plaque stability, inflammatory niches, and high-risk morphologies	Spatial transcriptomics of human coronary arteries with histology and plaque morphology; some cohorts include CT or invasive imaging	Dozens of plaques; expression matrices; and metadata commonly deposited to GEO/ENA (see articles)	GEO

Table 2 Representative cardiovascular imaging biomarkers that often serve as anchors in imaging-anchored multiomics, with typical computational extraction strategies and plausible molecular correlates.

Biomarker (modality)	Clinical interpretation	Typical computational extraction	Plausible or reported molecular correlates
LGE extent and pattern (CMR) [20]	Focal myocardial scar/fibrosis; arrhythmic and HF risk	Segmentation of LGE-positive regions; regional burden by AHA segments; texture/radiomics	Fibroblast activation, collagen deposition (ECM programmes), immune infiltration in border zones; stress and remodelling programmes in cardiomyocytes
Global longitudinal strain (echo) [21]	Subclinical systolic dysfunction; prognostic across many CVD states	Speckle-tracking strain curves; regional strain maps; video encoders for deformation fields	Energetic stress, sarcomeric dysfunction, inflammatory signalling; chamber-specific cardiomyocyte stress responses
Low-attenuation noncalcified plaque (coronary CTA) [22]	Lipid-rich plaque; higher MI risk	Plaque segmentation and attenuation profiling; radiomics; phenotype classifiers	Macrophage-driven inflammation, necrotic core formation, matrix remodelling; immune niches at plaque shoulders
Napkin-ring sign/high-risk plaque morphology (CTA) [23]	High-risk plaque architecture; potential plaque vulnerability	CNN/ViT-based plaque morphology classification; handcrafted geometry features	Protease and cytokine signalling, macrophage activation, smooth muscle cell phenotypic switching; local hypoxia
Perivascular fat attenuation index (FAI) (CTA) [24]	Non-invasive readout of coronary inflammation	Perivascular fat segmentation; attenuation gradients; radiotranscriptomic mapping	Inflammatory adipose and vascular wall transcripts; cytokine signalling and fibrosis-related programmes

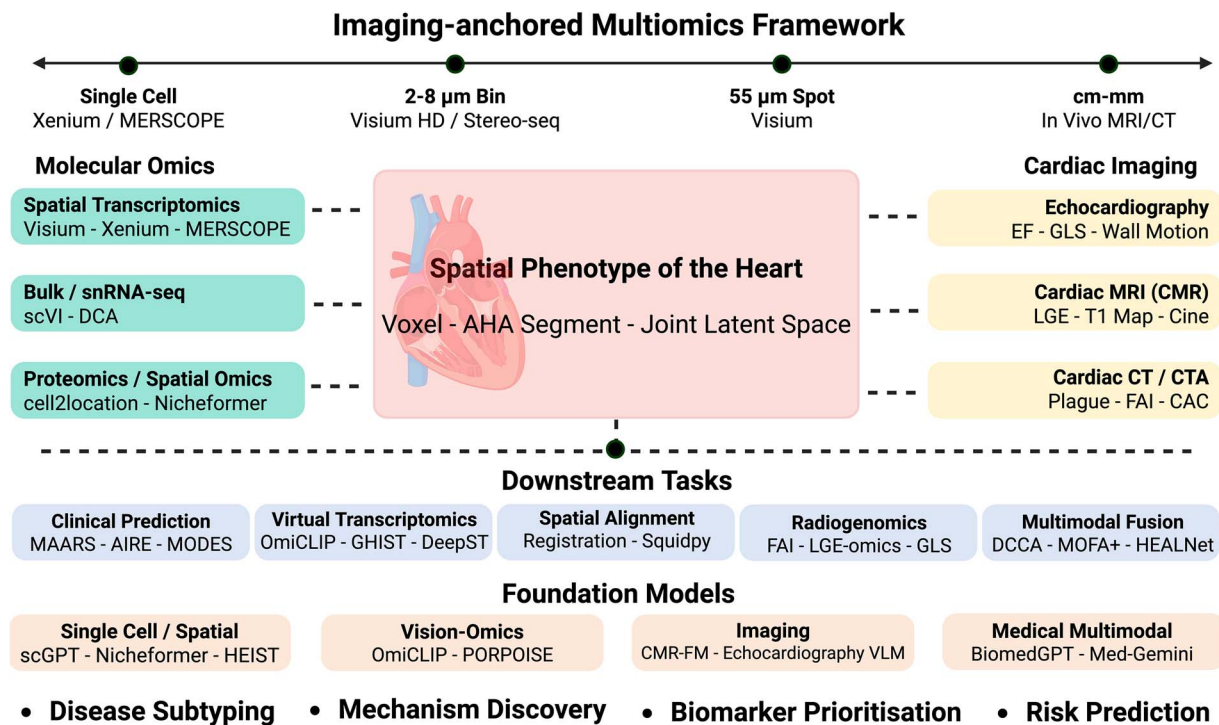


Figure 1 The resolution scale illustrates the gap between in vivo cardiac imaging at centimetre-to-millimetre resolution and molecular spatial profiling, from Visium 55 μm spots and Visium HD/Stereo-seq 2-8 μm bins to single-cell platforms such as Xenium and MERSCOPE. Cardiac imaging modalities, including echocardiography, cardiac magnetic resonance imaging (CMR), and cardiac CT/CTA, define spatial phenotypes of the heart using measures such as ejection fraction, global longitudinal strain, wall motion, late gadolinium enhancement, T1 mapping, cine imaging, plaque features, fat attenuation index, and coronary artery calcium. These phenotypes can be indexed at the voxel, American Heart Association segment, or joint latent-space level and linked to molecular omics layers, including spatial transcriptomics, bulk or single-nucleus RNA sequencing, and proteomics/spatial omics. The resulting joint representations support downstream tasks including clinical prediction, virtual transcriptomics, spatial alignment, radiogenomics, and multimodal fusion. Foundation models spanning single-cell/spatial omics, vision-omics, imaging, and medical multimodal data provide computational support across the framework. Applications include disease subtyping, mechanism discovery, biomarker prioritisation, and risk prediction. Abbreviations: AHA, American Heart Association; CMR, cardiac magnetic resonance imaging; CT, computed tomography; CTA, computed tomography angiography; EF, ejection fraction; GLS, global longitudinal strain; LGE, late gadolinium enhancement; FAI, fat attenuation index; CAC, coronary artery calcium; scVI, single-cell variational inference; DCA, deep count autoencoder; DCCA, deep canonical correlation analysis; MOFA+, multi-omics factor analysis; VLM, vision-language model.

detection in echocardiographic videos [5, 16]. CMR provides cine imaging for ventricular function, parametric mapping for tissue characterization, and LGE for focal fibrosis; population-scale examples are available in the UK Biobank programme [17]. Cardiac CT, particularly coronary CT angiography, visualizes coronary anatomy, and plaque morphology, including low-attenuation plaque and calcification patterns associated with risk [18].

These modalities define a *spatial phenotype*: voxels or segments correspond to physical locations and encode tissue composition, perfusion, and mechanical function. Traditional analyses compress this into global or regional summary metrics, but imaging-omics integration often benefits from intermediate representations that retain spatial detail [19]. These imaging-derived features provide the scaffold onto which molecular context can be mapped.

Imaging biomarkers as anchors and their molecular correlates

Many imaging-omics studies begin with clinically established biomarkers and then ask which molecular programmes best explain

their variation. Table 2 summarizes representative biomarkers used as anchors and plausible cellular and pathway-level correlates.

Omic layers: bulk, single-cell, and proteomic context

Transcriptomic and related omic measurements capture dynamic molecular states shaped by genetic background and environment. Bulk RNA-seq from biopsies or explanted myocardium profiles pathways active in cardiomyopathies, myocarditis, and heart failure [25], while plasma proteomics quantifies circulating inflammatory, fibrotic, and metabolic proteins that can predict events beyond classical risk factors [26]. These bulk measurements average over cell types and niches, offering global molecular context but limited resolution.

Single-cell and single-nucleus RNA sequencing provide cell-level resolution and reference atlases of cardiomyocytes, fibroblasts, endothelial and smooth muscle cells, pericytes, and immune populations [2, 27]. In disease, cardiomyocytes often converge onto stress programmes, whereas fibroblasts, endothelial cells, and immune cells diversify into disease-associated states enriched for inflammatory and fibrotic genes [3, 25]. Although this review emphasizes transcriptomic layers, additional omics (epigenomics, metabolomics, and proteomics)

are frequently available and can be incorporated via multi-view factor models or graph-based encoders.

Spatial transcriptomics as a bridge

Spatial transcriptomics preserves *where* transcripts are located within tissue. Cardiovascular studies have mapped infarct cores, border zones, and remote myocardium [12], delineated microdomains in cardiomyopathies [28], and characterized vulnerable plaque regions in atherosclerotic arteries [29]. Spatial clustering and domain discovery are central for revealing microenvironments and interaction niches, and are typically coupled to marker-gene analysis and deconvolution against single-cell references—using methods such as cell2location [30]—to infer cell-type proportions within each capture spot [12].

For imaging-anchored multiomics, spatial transcriptomics is particularly attractive because it resides close to histology and can be warped toward imaging coordinate systems via intermediate registration steps. Spot-level expression, deconvolved cell composition, and histology features together provide a molecular annotation of tissue architecture that can be linked to imaging-derived phenotypes.

Representation learning across modalities

Imaging encoders

Cardiac imaging data are high-dimensional, often 3D and sometimes time-resolved. Representation learning aims to compress these data into compact embeddings that preserve clinically and biologically relevant information. For imaging-anchored multiomics, the encoder must satisfy two requirements that go beyond standard segmentation or classification: first, it must produce embeddings at spatial granularities that can be aligned with molecular data—from whole-heart summaries matched to bulk transcriptomics, through regional American Heart Association segments paired with biopsy-level profiles, to voxel-level features co-registered with individual spatial transcriptomic spots. Second, it must learn generalizable representations from large unlabelled imaging archives, because the paired imaging-omics samples available for joint training typically number in the tens to low hundreds.

Convolutional architectures

Convolutional neural networks (CNNs) and their 3D or spatiotemporal variants remain the workhorses for echocardiography and CMR segmentation, view classification and disease prediction [4, 5]. Population-scale deployment on the UK Biobank, where fully convolutional networks have segmented cardiac chambers in tens of thousands of subjects [37], illustrates that CNN-derived phenotypes—volumes, mass, wall thickness—can already serve as high-throughput imaging features for genetic association [38] and, by extension, for radiogenomic studies that link imaging variation to transcriptomic programmes—as demonstrated by radiotranscriptomic mapping of perivascular fat [39].

For cine-MRI and echocardiographic videos, 3D CNNs or 2D CNNs combined with temporal convolution or recurrent layers can model wall motion and valve dynamics [4, 5]. Crucially for multiomics integration, intermediate convolutional feature maps retain spatial layout: region- or voxel-level embeddings can be extracted from these maps and paired with molecular measurements at matching anatomical locations [19], providing a direct bridge between the imaging encoder,

and the spatial or regional molecular annotations discussed in the preceding section.

Transformer and self-supervised architectures

Vision Transformers (ViTs) [40] and masked autoencoders (MAEs) [41] offer a complementary design. Images or volumes are decomposed into patches, projected into tokens, and processed via self-attention, which captures long-range dependencies across remote myocardial or coronary segments. This property matters when the imaging phenotype of interest—such as diffuse fibrosis detectable on T1 mapping, global longitudinal strain, or multi-territory plaque burden—reflects spatially distributed biology that local convolutions may underrepresent. Self-supervised objectives such as masked patch prediction or contrastive learning on large unlabeled archives enable these encoders to learn generic cardiac representations that transfer to downstream tasks with comparatively few labels: a CMR foundation model pretrained on 36 million images [42] and a vision-language echocardiography model trained on over one million video-report pairs [43] both demonstrate strong few-shot and zero-shot transfer to clinical cardiac tasks, respectively. For imaging-omics integration, such pretrained encoders can provide high-quality, spatially structured embeddings without consuming the scarce paired imaging-molecular samples for encoder training; those paired data can instead be reserved for the fusion and alignment layers discussed in the “Multimodal fusion: methods, pipelines, and evaluation” section. Bridging from imaging to molecular states at spot or cell resolution, however, requires a different class of encoder that jointly learns from histology and transcriptomics. Vision-omics foundation models such as OmiCLIP, trained on 2.2 million paired haematoxylin and eosin (H&E)–transcriptome patches across 32 organs including heart failure tissue [44], can predict spatial gene expression and infer cell-type composition directly at Visium spot level. Similarly, self-supervised ViTs trained on large histology archives have been shown to predict and spatially localize individual transcript abundances across 23 human tissue types [45]. Although these histology-omics models have been demonstrated primarily in oncology, the architectures are directly adaptable to cardiac tissue sections once sufficiently large cardiac spatial datasets become available—a prospect discussed in the “Future directions” section. Training patient-level imaging-omics models end-to-end from scratch on small paired cohorts meanwhile risks severe overfitting to confounders such as site or scanner batch, a failure mode discussed in the “Bottlenecks, root causes, and failure modes” section.

From handcrafted radiomics to learned features

Before deep encoders became dominant, handcrafted radiomics—shape, intensity histogram, and texture descriptors extracted from segmented regions of interest—provided the primary featurization of cardiac images for association studies [19]. Radiomics remains valuable for interpretability: first-order and texture features have transparent physical meanings and can be mapped onto tissue-level processes such as fibrosis heterogeneity or myocardial oedema [19]. In practice, hybrid strategies that concatenate radiomics with deep embeddings often outperform either alone, because handcrafted features capture domain priors that a data-hungry deep model may miss in small paired cohorts. For imaging-omics integration specifically, radiomic features offer the further advantage that they can be extracted per AHA segment or per plaque, yielding a natural correspondence with

region-level molecular measurements from bulk RNA-seq or deconvolved spatial transcriptomics.

Multi-scale embeddings for molecular alignment

The choice of embedding scale should match the granularity of available molecular data and the biological question. Global (whole-heart) embeddings suffice for patient-level radiogenomic association, where the aim is to correlate an imaging-derived trait such as left ventricular mass with bulk gene-expression or plasma proteomic modules. Regional embeddings at the level of AHA segments or individually segmented plaques are appropriate for linking to biopsy-level profiles or cell-type compositions estimated by deconvolution methods such as cell2location [30]. Voxel-level features are needed to map gene-expression gradients across infarct border zones or plaque shoulders where spatial transcriptomics provides spot-level resolution [12]. Designing encoders that can output embeddings at multiple scales simultaneously, or that can be queried at arbitrary coordinates via spatial interpolation, remains an open challenge that directly determines the resolution at which imaging and molecular data can be fused.

Omics and single-cell encoders

Bulk transcriptomic data are typically high-dimensional, with thousands of genes, substantial measurement noise, and pervasive gene-gene co-expression organized into modules of functionally related genes [46], which makes the data amenable to dimensionality reduction.

Standard analysis toolkits—Seurat [47] in R and Scanpy [48] in Python—implement principal component analysis (PCA)-based dimensionality reduction, graph-based clustering (Leiden/Louvain), differential expression testing and visualization via UMAP or t-SNE, providing the default entry point for both bulk and single-cell workflows. For deeper denoising, DCA [49] trains a zero-inflated negative binomial autoencoder to impute dropout events in count matrices. scVI [50], implemented within the scvi-tools ecosystem [51], uses a variational autoencoder to jointly perform dimensionality reduction, batch correction, and differential expression in a single probabilistic framework. MOFA+ [52] decomposes variation across multiple omic layers into shared factors capturing cross-modality biology and private factors specific to one layer. For imaging-omics integration, the low-dimensional representations produced by these tools—whether PCA components, scVI latent variables, or MOFA+ shared factors—serve as molecular summaries that can be aligned with imaging embeddings.

Single-cell transcriptomics introduces additional complexity: sparse count data, batch effects, and large numbers of cells. The scVI framework described above was designed specifically for these challenges, and its latent spaces support cell clustering, trajectory inference, and integration across donors and conditions [50].

Beyond dataset-specific latent models, single-cell, and multiomics foundation models trained on tens of millions of cells are beginning to appear [53, 54]. Large-scale transformers and state-space models that ingest gene-expression vectors, chromatin accessibility, spatial coordinates, or perturbation labels can provide transferable embeddings that generalize across tissues and species [53, 54]. Early evaluations suggest that cardiac and vascular cell states are represented faithfully in such models, which raises the prospect of using pretrained single-cell foundation models as generic encoders in cardiovascular multiomics.

Spatial transcriptomics encoders

Spatial transcriptomics data reside on grids of capture spots or irregular spatial point sets. Unlike scRNA-seq, where each observation represents a single cell, sequencing-based platforms such as Visium assign each spot a mixture of transcripts from multiple co-located cells [10], and even higher-resolution platforms retain spatial dependencies absent from dissociated-cell data. Applying standard scRNA-seq dimensionality reduction—PCA followed by k-nearest-neighbour graphs constructed purely in expression space—to spots ignores both this within-spot cellular heterogeneity and the tissue-level spatial structure, motivating dedicated spatial encoders.

Graph-based spatial domain identification

GNN methods address this by constructing graphs in which nodes represent spots and edges connect spatial neighbours, so that message passing yields embeddings that respect both expression similarity and physical adjacency. SpaGCN [55] integrates gene expression, spatial coordinates, and histology via graph convolution to identify spatial domains with coherent expression and morphology, and is used for spatially variable gene detection. BayesSpace [56] adopts a Bayesian framework that models spatial neighbourhood priors to achieve sub-spot resolution clustering, enabling finer domain delineation than the native Visium grid. STAGATE [57] introduces an adaptive graph attention auto-encoder that learns the similarity of neighbouring spots through attention weights, improving boundary detection between spatial domains and supporting denoising of expression data. GraphST [58] combines GNNs with self-supervised contrastive learning—minimizing embedding distance between spatially adjacent spots while maximizing it for non-adjacent ones—and is the first method to jointly perform spatial clustering, multi-sample batch integration and cell-type deconvolution in a unified framework. Spatially informed graph attention architectures such as STMSGAL [59] further highlight how multi-scale attention and graph signal processing can be combined to improve spatial domain identification in noisy settings.

Multimodal histology-expression encoders. When paired histology images are available, multimodal encoders that jointly process H&E patches and spot-level gene-expression vectors can learn richer representations than expression-only methods. Contrastive frameworks such as mclSTExp [60] align image features from a DenseNet encoder with spot expression and positional embeddings from a Transformer encoder in a shared latent space, and use the learned alignment to predict spatial gene expression from H&E images alone. A recent benchmark of 11 histology-to-expression prediction methods confirmed that such H&E-based virtual transcriptomics is feasible across tissue types, while identifying batch effects and limited cross-dataset generalizability as key remaining challenges [61]. Foundation-scale architectures trained on hundreds of thousands of spots from diverse tissues—including graph-based models such as HEIST [62] and cross-attention or diffusion approaches such as DeepST [63]—push this further by learning transferable spatial representations in a self-supervised regime.

Relevance to imaging-omics integration

For imaging-anchored multiomics, these spatial encoders serve two complementary roles. First, they compress high-dimensional spot-level expression into low-dimensional spatial feature maps that can be co-registered to *in vivo* imaging via the alignment pipelines described in the “Spatial molecular alignment and tissue-image

correspondence" section. Second, imaging encoders can be designed to output features at the resolution of projected spatial spots—for instance by resampling CMR-derived LGE intensity, T1 values or strain at spot coordinates—enabling feature-level pairing between imaging and molecular embeddings within a shared graph or latent space. Although most current spatial encoder demonstrations are in oncology, the architectures should be evaluated in cardiac and vascular tissue as cardiovascular spatial atlases expand (Future directions).

Self-supervision, batch effects, and limitations

Across modalities, self-supervised learning is becoming the dominant pretraining paradigm, driven by the observation that labelled and paired imaging–omics data are scarce, whereas unimodal archives are large. For cardiac imaging, contrastive and masked-modelling objectives are now producing encoders that rival or exceed supervised baselines in label-limited regimes. In echocardiography, a self-supervised segmentation pipeline trained on only 450 studies without any manual chamber labels produced cardiac chamber measurements across >18,000 echocardiograms that matched both inter-clinician variability and fully supervised learning [64]. At the CMR scale, the foundation model of Jacob *et al.* [42], pretrained with masked image modelling on 36 million CMR images, transfers to segmentation and function quantification with minimal labelled data, and the vision–language echocardiography model of Christensen *et al.* [43] achieves zero-shot view classification from over one million video-report pairs. Together, these models illustrate that self-supervised cardiac imaging encoders can now produce high-quality embeddings without consuming the scarce paired imaging–molecular samples needed for downstream fusion.

For transcriptomics, analogous self-supervised strategies are maturing rapidly. Masked gene modelling, in which a fraction of gene-expression values are withheld and the model is trained to reconstruct them, underpins single-cell foundation models such as scGPT, a generative pretrained transformer trained on over 33 million cells that supports batch correction, cell-type annotation, and perturbation prediction after fine-tuning [65], and scFoundation [53]. Denoising autoencoders such as DCA [49] and variational frameworks such as scVI [50] capture co-expression structure without labels and remain standard preprocessing steps in single-cell pipelines.

Cross-modal self-supervised objectives further enrich representations for integration. A cross-modal autoencoder trained on paired electrocardiography, CMR, and clinical variables at biobank scale demonstrated that a shared latent space can simultaneously support imaging-based prediction, genome-wide association, and counterfactual simulations [66], and CLIP-style contrastive objectives that align imaging–text pairs in radiology [67, 68] provide a template for imaging–omics contrastive pretraining when paired cohorts become available.

In parallel, multimodal medical foundation models that couple imaging, free-text reports, and other clinical data have been proposed for radiology and general biomedical tasks [8, 9, 69, 70]. These architectures typically combine masked modelling and contrastive objectives across modalities and offer a natural starting point for incorporating omic inputs once suitably annotated cohorts are available. Recent large vision–language models for radiology and generalist biomedical AI further underline the need for careful curation,

auditing, and domain-specific fine-tuning when such backbones are repurposed for multimodal bioinformatics [69, 71].

Batch effects and cross-centre variability pose persistent challenges that affect both modalities and their joint analysis. Imaging domain shifts arise from differences in scanners, acquisition protocols, and reconstruction algorithms; a recent survey of deep-learning-based MRI harmonization catalogued a rapidly growing toolkit of GANs, VAEs, normalizing flows, and disentangled-representation methods for reducing scanner-specific bias, while noting that most approaches still lack standardized evaluation on cardiac data [72]. For omics, batch-correction methods such as ComBat [73] and Harmony [74], and joint latent models such as scVI [50], are well established; however, harmonizing across fundamentally different platforms—for instance, Visium versus Visium HD or 10× Chromium versus Smart-seq—remains challenging even with these tools. When imaging and omics batches are confounded, as often happens when specific imaging centres also provide the only tissue samples, the risk of learning spurious associations increases sharply. A practical limitation is sample size: paired imaging–omics datasets remain small, often involving tens to hundreds of patients, which compounds the risk of overfitting and confounding by site or protocol. Representation-learning strategies must, therefore, be coupled to stringent validation and sensitivity analyses, including leave-one-site-out designs and negative controls with permuted pairings, as discussed in the “Bottlenecks, root causes, and failure modes” section.

Multimodal fusion: methods, pipelines, and evaluation

This section examines how radiomic, transcriptomic, and spatial transcriptomic representations can be combined in joint models. We use “fusion” broadly to include architectures, learning objectives, and evaluation frameworks; Fig. 2 summarizes early, intermediate, late, and graph-based strategies.

A prerequisite to any fusion strategy is understanding the data generation pipeline. In clinical cardiovascular practice, echocardiography, CMR, and CT are acquired routinely and cover the entire heart or vascular territory at millimetre-to-centimetre resolution. Radiomics features—shape, intensity histogram, texture, and deep-learned descriptors extracted from segmented regions of interest [19, 75]—constitute the imaging-derived layer of the multiomics stack. Molecular profiling, by contrast, requires physical tissue: endomyocardial biopsies, surgical specimens, explanted hearts or, for circulating analytes, blood draws. Because tissue is scarce and costly, radiomics features play a critical upstream role in *defining which regions will be profiled molecularly*. LGE-positive versus LGE-negative myocardial segments, infarct border zones delineated on parametric maps, or coronary plaque regions classified by CT attenuation can all serve as radiomics-defined targets that guide biopsy or tissue-block dissection [12, 13]. Real-time CMR-guided endomyocardial biopsy can further improve targeting accuracy by directing sampling to LGE-defined lesions rather than relying on anatomical landmarks alone [14, 15]. The resulting molecular data—bulk RNA-seq, sc/snRNA-seq, spatial transcriptomics, plasma proteomics—are then linked back to the radiomics features from which the sampling decision was derived, forming the radiogenomic pairs that fusion models ingest, as exemplified by radiotranscriptomic mapping of perivascular fat [39]. Two consequences follow for model design. First, molecular measurements are spatially sparse relative to radiomics features: omics are available only at sampled regions,

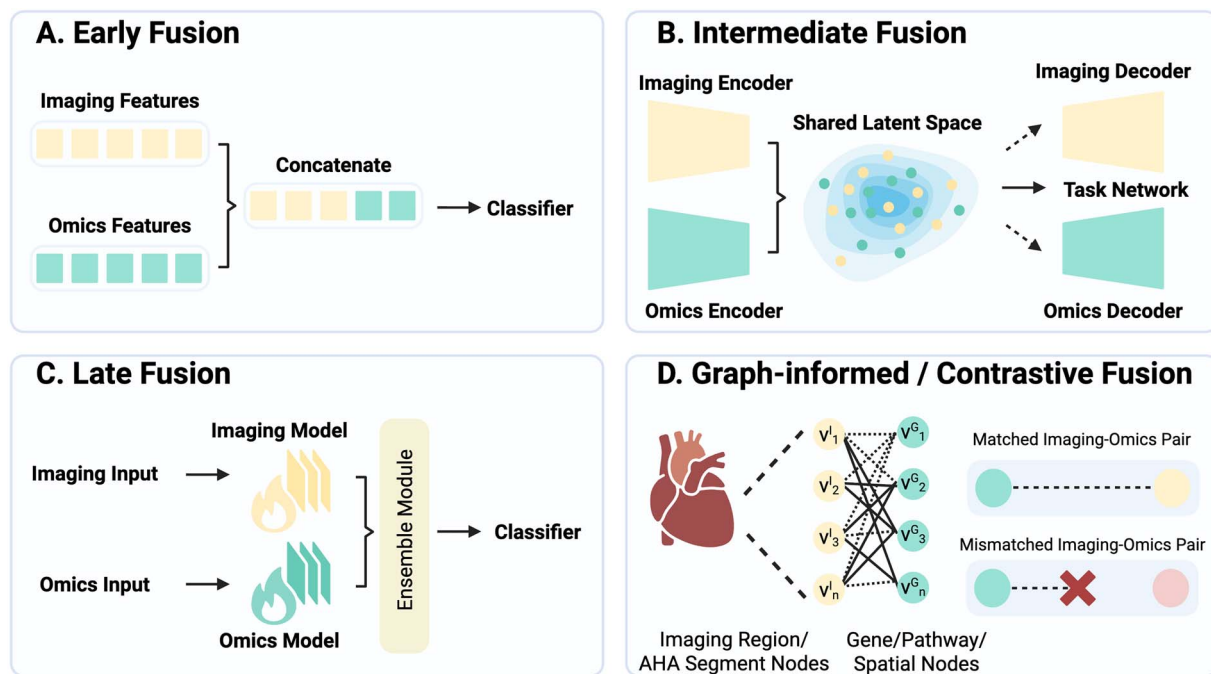


Figure 2 Multimodal fusion architectures for radiomics–omics integration. Panel (A): early fusion concatenates radiomics and omics feature vectors followed by a single predictive model. Panel (B): intermediate fusion uses modality-specific encoders and a shared latent space, often with reconstruction heads; the shared latent captures biology concordant across modalities while private components retain modality-specific information. Panel (C): late fusion combines predictions from separate radiomics and omics models via an ensemble or calibration layer; it is robust to missing modalities but cannot capture cross-modal feature interactions. Panel (D): graph-based and contrastive approaches construct cross-modal latent spaces using biological or spatial graphs and contrastive objectives.

not across the entire imaged field of view, so fusion architectures must tolerate this asymmetry [6]. Second, the radiomics-guided sampling process itself introduces selection bias—tissue is preferentially obtained from regions that are abnormal on imaging—which must be acknowledged when interpreting associations between radiomics features and molecular profiles [11].

Early, intermediate, and late fusion

Early fusion concatenates raw or lightly processed features from all modalities into a single vector per sample before any learning occurs. In an imaging–omics cardiovascular study, this means appending CMR-derived radiomics descriptors (e.g. LGE texture, wall thickness, and strain values) to bulk gene-expression principal components or pathway activity scores, then feeding the combined vector into a standard learner such as gradient boosting, a multilayer perceptron, or a Cox survival model [76]. Early fusion is conceptually the simplest strategy and often serves as a competitive baseline. However, it assumes that all features share a comparable scale and noise structure, which is rarely the case when combining tens of radiomics descriptors with thousands of gene-expression measurements. It is also fragile to missing data: patients without one modality must be imputed or excluded entirely, which can discard a substantial fraction of the cohort given that tissue for omics is available only in selected cases [77].

Intermediate fusion first learns modality-specific embeddings, then combines them in a shared latent space—and it is this shared space that makes the strategy particularly suited to imaging–omics integration. The core idea is that radiomics features and molecular data describe the same underlying biology—tissue composition, cell-state

distributions, pathway activity—but through different measurement processes; an intermediate latent space is designed to capture the biological variation that is *shared* across modalities while retaining *private* variation unique to each. Concretely, modality-specific encoders compress a CMR cine sequence into a vector summarizing ventricular geometry and wall motion, and compress a bulk gene-expression profile into a vector summarizing pathway activation patterns. These vectors are then projected into a common embedding where, for example, a fibrotic imaging phenotype and an activated-fibroblast transcriptomic programme should map to nearby points. Deep canonical correlation analysis (DCCA) achieves this by learning non-linear projections that maximize cross-modal correlation while reconstruction losses preserve modality-specific detail [78, 79]. Multimodal VAEs with product-of-experts or mixture-of-experts inference extend this to a probabilistic setting: the encoder for each modality proposes a posterior distribution over the latent space, and these posteriors are combined via their product or mixture, so the joint latent captures information that is consistent across whichever modalities are present [80, 81]. This formulation offers three practical advantages for imaging–omics cohorts. First, the shared latent explicitly represents which aspects of the imaging phenotype are explained by measured molecular layers—and which are not—enabling mechanistic interrogation via latent traversals or factor-level association tests. Second, cross-modal generation becomes possible: omics embeddings can be decoded into predicted imaging features and vice versa, supporting “virtual transcriptomics” from imaging inputs. Third, the product-of-experts formulation naturally handles missing modalities, because the joint posterior remains well-defined even when only one encoder contributes, allowing patients with imaging alone or omics alone to regularize the shared space without requiring complete

Table 3 Practical comparison of traditional unimodal/low-dimensional modelling versus multimodal fusion strategies for imaging-anchored multi-omics.

Approach	Strengths	Common failure modes	When it is a good default
Classical regression/survival models with hand-crafted radiomics summaries	High interpretability; strong calibration; stable in small samples	Limited ability to capture spatial heterogeneity and cross-modal interactions; features may miss subtle biology	Early translational work; limited paired data; when interpretability and calibration dominate
Early fusion (feature concatenation)	Simple; uses standard ML toolkits; often competitive baselines	Sensitive to missing modalities; scaling/normalization issues; collinearity, and overfitting with high-dimensional omics	Moderate sample sizes with mostly complete modalities; when a strong baseline is needed quickly
Intermediate fusion (joint latent space: DCCA, multimodal VAE, and HEALNet)	Captures non-linear cross-modal interactions; supports “virtual omics” and mechanistic exploration; handles missing modalities via product-of-experts	Latent space collapse; modality imbalance; requires careful regularization and validation	When paired subsets exist but are not large; when cross-modal generation/interpretation is desired
Late fusion (ensembles/stacking)	Robust to missing modalities; reuses validated unimodal models; simplifies regulatory adoption	Cannot learn feature-level synergy; may underperform when cross-modal interactions are important; forfeits opportunity to learn from costly paired data	When missingness is high; when unimodal models are already mature and well-validated; as deployment fallback
Graph/contrastive and transformer-based fusion	Naturally incorporates spatial and biological priors; can scale with pretraining; captures neighbourhood relationships lost by flat vectors	Data-hungry; computationally expensive; prone to spurious shortcuts without strong evaluation	When large weakly labelled data exist for pretraining, or strong spatial/biological graphs are available

pairing. Recent hybrid architectures push intermediate fusion further: HEALNet uses iterative cross-attention between a shared latent bottleneck and modality-specific representations of histopathology and multi-omic data, achieving state-of-the-art survival prediction on four TCGA cancer datasets while gracefully skipping missing modalities at inference [82], and the MODES framework decouples shared and modality-specific latent components from paired ECG and cardiac MRI in the UK Biobank, enabling inference of missing MRI phenotypes from ECG alone [83].

Late fusion trains separate, fully independent models for each modality and combines their predictions at the decision level using averaging, voting, or a meta-learner. Late fusion is sometimes viewed as wasteful for imaging–omics studies, where paired data are expensive to generate and training independent models forfeits the opportunity to learn cross-modal structure from those costly pairings. However, the strategy remains practically important for two reasons. First, it is robust to missingness: predictions can be computed from whichever modalities are available, so a patient with only CMR and no biopsy still receives a calibrated risk estimate from the radiomics arm, while the transcriptomic arm contributes when tissue is obtained—a realistic deployment scenario in which imaging is routine but omics are reserved for selected cases. Second, late fusion reuses validated unimodal models: if a radiomics-only model has already been externally validated for a fibrosis endpoint and an omics-only model has been validated for pathway-level risk scoring, their combined output inherits the credibility of each component, which simplifies regulatory and clinical adoption. Its main limitation is that it cannot capture feature-level interactions—for instance, a specific gene-expression pattern that is prognostically relevant only in the presence of a particular LGE distribution will be invisible to models that never see both modalities jointly. For imaging–omics integration, late fusion is, therefore, best understood as a robust baseline and deployment-stage fallback: it ensures that expensive molecular data are never wasted, while intermediate or graph-based fusion layers can be trained on the smaller

fully paired subset to capture the cross-modal interactions that late fusion misses.

In practice, hybrid designs are common. Modality-specific encoders are often pretrained and fine-tuned on unimodal tasks, supporting late fusion, while intermediate fusion layers are trained on subsets with paired data to capture cross-modal structure. This modularity facilitates reuse and improves robustness: if a radiomics encoder is updated, the fusion and calibration layers can be retrained without discarding the transcriptomics model.

Comparing fusion strategies with traditional approaches

Multimodal fusion should be understood as an extension of long-standing clinical modelling approaches rather than a wholesale replacement. Classical approaches in cardiovascular risk prediction often combine a small number of radiomics-derived summary variables with demographics and labs in regression or Cox models. Fusion models generalize this by (i) allowing radiomics and omics to be represented at higher dimensionality and multiple scales, and (ii) learning non-linear interactions between modalities. Table 3 summarizes trade-offs that are most relevant in practice: interpretability, sample size requirements, robustness to missingness, and suitability for mechanistic inference.

Graph-based fusion and contrastive cross-modal learning

The fusion strategies described above—early, intermediate, and late—treat each sample as a flat feature vector or a fixed-size embedding. This representation discards two types of structure that are central to imaging–omics integration in CVD. First, *spatial structure*: radiomics features are indexed by anatomical location (AHA segments, plaque regions, infarct zones), and spatial transcriptomic spots are connected

by tissue adjacency; collapsing these into unordered vectors loses the neighbourhood relationships that determine, for example, whether a fibrotic border zone abuts viable myocardium or necrotic core. Second, *biological structure*: genes do not act independently but are organized into pathways, protein–protein interaction networks and regulatory modules whose topology encodes prior knowledge about coordinated function. Graph-based and contrastive fusion methods address both limitations by explicitly encoding spatial and biological relationships as edges in a graph, enabling the model to propagate and integrate information along these relationships rather than treating all features as exchangeable.

In *feature graphs*, nodes represent radiomics-defined imaging regions, genes, or pathways, and edges encode known relationships—such as anatomical adjacency in the heart (e.g. neighbouring AHA segments sharing a border zone) or membership in the same biological pathway (e.g. TGF- β signalling linking fibroblast-activation genes). GNNs propagate information along these edges through iterative message passing, allowing imaging nodes to aggregate molecular signals from connected pathway nodes and molecular nodes to incorporate structural context from imaging. A graph might connect LGE-positive segments to fibrosis-related pathways enriched in bulk or spatial transcriptomics, enabling the GNN to learn which molecular programmes best explain regional fibrosis patterns. Zheng *et al.* [84] demonstrated this principle by representing whole-slide pathology images as graphs whose node embeddings are fused with gene-expression signatures via cross-attention, producing spatially resolved survival activation maps that localize which tissue regions and which pathways jointly drive prognosis in non-small cell lung cancer. A comprehensive review of GNN architectures for single-cell and spatial omics—covering graph convolution, graph attention and variational graph autoencoders across more than 100 published models—confirms that graph-based encoders consistently outperform non-spatial baselines for tasks including spatial domain identification, cell-type deconvolution, and multi-omics integration [85].

In *patient graphs*, nodes represent patients and edges reflect similarity in radiomics features, molecular profiles, or both. Message passing across such graphs can regularize predictions, particularly in small cohorts where individual measurements are noisy [86]. For imaging–spatial transcriptomics integration, spatial spots can serve as nodes connected by tissue adjacency and embedded with both expression and local radiomics features, enabling graph-based segmentation of molecularly distinct tissue niches such as infarct border zones or vulnerable plaque shoulders [12, 29].

Contrastive multimodal learning complements graph-based fusion by providing an explicit training objective: pull together paired samples—a radiomics feature vector and its corresponding molecular profile from the same tissue region—and push apart unpaired ones, so that the learned embedding space reflects true biological correspondence rather than spurious batch or modality correlations. CLIP-style objectives have been applied to imaging–text pairs in radiology [67, 68], and analogous objectives can be used for radiomics–omics pairing. Cross-modal autoencoders trained on paired electrocardiography, CMR, and clinical variables at biobank scale have demonstrated how a shared latent space can simultaneously support imaging-based prediction, genotype association, and counterfactual simulations [66]. For radiomics–spatial transcriptomics integration, contrastive objectives can align region-level radiomics embeddings with spot-level molecular embeddings, such that nearby points in latent space correspond to tissue regions that are both visually and molecularly similar. Graph

contrastive learning extends this idea by treating different perturbations of the same spatial or patient graph as positives, which encourages robustness to missing spots or noisy gene counts—a property particularly valuable given the sparsity and technical noise inherent in current cardiovascular spatial transcriptomic datasets.

Cross-modal latent spaces, missing data, and sample size

Learning a cross-modal latent space provides a common coordinate system for downstream tasks such as clustering, subtyping, genetic association, and simulation. Multimodal VAEs, DCCA, and cross-modal autoencoders are typical tools. For imaging–omics applications in CVD, three design challenges are particularly important and interact with one another.

First, *handling of missing modalities* is critical. In practice, imaging is available for nearly all patients, bulk RNA-seq or proteomics may be available for a subset with tissue or blood samples, and spatial transcriptomics for a further subset where fresh-frozen sections were collected—producing a pyramid of decreasing sample size as modality depth increases. Fully paired imaging–bulk–spatial data are therefore rare. Architectures that can ingest any subset of modalities—through modality-specific encoders with masking, product-of-experts or mixture-of-experts combinations—are essential [80, 81]. These models can be trained on all available data: fully paired samples anchor the joint space, while unimodal samples regularize each encoder and reduce the risk of overfitting to small paired subsets. The PORPOISE framework demonstrated this principle at scale in oncology by fusing whole-slide histopathology images with multi-omic profiles across 14 cancer types, tolerating variable molecular availability through gated multimodal fusion and yielding interpretable prognostic markers [87]. An analogous design for cardiovascular imaging–omics would pair a CMR or echocardiographic encoder with molecular encoders and train the fusion layer on the small subset of patients with both imaging and tissue, while the imaging encoder is regularized by the much larger imaging-only cohort—precisely the strategy used by the cross-modal cardiovascular autoencoder of Radhakrishnan *et al.* [66], which leveraged tens of thousands of ECG–CMR pairs to stabilize a latent space that also supports genotype association.

Second, *realistic expectations regarding sample size* are required. High-capacity deep fusion models with millions of parameters are difficult to train on cohorts with tens of paired cases; simpler factor models, linear DCCA, or biologically constrained architectures may perform better in that regime. When the primary goal is mechanistic discovery rather than clinical prediction, modest sample sizes can still be informative if strong biological priors—pathway annotations, protein–protein interaction networks, or cell-type marker lists—are encoded in the model, for example, through pathway-level aggregation in the molecular encoder or graph-structured gene modules in the fusion layer. For patient-level prediction tasks, however, thousands of paired samples may be required to achieve stable performance and support stratified evaluation across demographic and clinical subgroups.

Third, *efficient use of partially paired data* is, therefore, a key strategy. Large imaging-only biobanks such as the UK Biobank (~100,000 CMR studies) and EchoNet (>10,000 echocardiographic videos) can pretrain imaging encoders, while smaller imaging–omics subsets sup-

Table 4 Evaluation metrics and validation designs commonly used (or recommended) in imaging-anchored multiomics.

Task	Example outputs	Useful metrics	Validation emphasis
Clinical prediction/risk modelling	Binary outcome, time-to-event risk, treatment response	AUC/time-dependent AUC, C-index, Brier score; calibration slope/curve; decision-curve net benefit [89]	External validation; leave-one-site-out; calibration and subgroup performance; reporting per TRIPOD+AI/PROBAST+AI [91]
Radiogenomic association	Imaging feature ↔ gene/pathway score	Effect sizes, false discovery rate control; pathway enrichment; replication rate	Confounder adjustment; permutation tests; replication in independent cohorts
Cross-modal alignment/retrieval	Paired embeddings; matched region–spot correspondences	Cross-modal retrieval (Recall@K), embedding correlation/CCA correlation, alignment error	Robustness to missing modalities; negative controls (shuffled pairing); sensitivity to batch
Spatial domain discovery	Spot clusters/spatial domains	Adjusted Rand index (if labels exist), silhouette score; spatial autocorrelation statistics	Stability under downsampling; cross-section reproducibility; biological plausibility via marker genes
Virtual transcriptomics	Predicted gene/module expression maps	Pearson/Spearman correlation, R^2 , RMSE; spatial pattern similarity	Hold-out donors/centres; robustness across staining/scanner protocols; uncertainty estimates

ply the cross-modal signal needed to anchor the joint latent space. Semi-supervised designs that mix paired and unpaired data—through modality dropout during training, pseudo-labelling of unpaired samples using cross-modal predictions, or contrastive objectives that align modalities when both are present and regularize individual encoders when only one is available—extend the effective training set without requiring additional costly tissue profiling. This hierarchical data strategy, using large unimodal data for encoder pretraining and small paired data for fusion fine-tuning, is likely to be the default architecture for cardiovascular imaging–omics until the cost of spatial transcriptomics and single-cell profiling decreases substantially.

Evaluation: metrics, validation, and benchmarking

A recurrent pitfall in multimodal bioinformatics is evaluating only the downstream prediction AUC and ignoring whether a model is calibrated, robust to distribution shift, or biologically plausible. Multimodal histology–genomic models in oncology have been shown to overestimate performance when site-of-origin batch effects are not controlled through site-preserved cross-validation [88], a risk that is equally acute in cardiovascular settings where specific imaging centres also provide the only tissue samples. For imaging-anchored multiomics, evaluation should therefore be *task-aligned*: mechanistic models are judged by different criteria than risk models, and both require specific safeguards against confounding. Table 4 summarizes practical metrics and recommended validation designs across five common task types.

Beyond these task-specific metrics, three overarching evaluation principles apply to imaging–omics models. First, *calibration and clinical utility* should accompany discrimination: a model that achieves high AUC but assigns risk probabilities that deviate systematically from observed event rates offers limited bedside value. Decision-curve analysis quantifies whether a model provides net benefit at clinically relevant thresholds [89], and calibration plots should be reported alongside discrimination metrics. Second, *negative controls* are essential: shuffling the pairing between imaging and molecular data, or permuting spatial coordinates of transcriptomic spots, should abolish any true cross-modal signal; if the model retains apparent performance under permuted pairings, the learned association is likely confounded by batch or population structure. Third, *subgroup and fairness evaluation* should be standard practice: performance should

be stratified by sex, ancestry, disease severity and, where possible, imaging centre, to ensure that the model does not preferentially benefit well-represented subpopulations while failing for others. Emerging AI-specific reporting guidance, including TRIPOD+AI for prediction models [90] and PROBAST+AI for risk-of-bias assessment [91], provides structured frameworks for transparent documentation of these aspects.

Reproducibility and software ecosystem

Reproducible multimodal fusion requires attention to data pre-processing, model specification, evaluation, and software. Pre-processing pipelines for imaging (reconstruction, segmentation, and normalization), bulk, and single-cell data (quality control, normalization, and batch correction) and spatial data (registration, spot filtering, and deconvolution) can strongly influence downstream results, and each choice introduces variation that is rarely quantified systematically. These steps should be documented, version-controlled and, where possible, implemented in reusable workflows with containerized environments and fixed random seeds.

The software ecosystem for multimodal fusion is growing but remains fragmented. General-purpose toolkits such as *scvi-tools*, *MOFA+* and *Seurat/Signac* support multiomics and spatial analyses [51, 92]; spatial-omics libraries such as *Squidpy* [93] and *Giotto* [94] provide graph construction, spatial statistics, and cell–cell interaction analysis. Imaging frameworks such as *MONAI* [95] and *TorchIO* [96] handle medical image processing and model training; and open-source multimodal fusion codebases from the computational pathology community—including *PORPOISE* [87] and *HEALNet* [82]—provide reference implementations for histology–omics integration that can be adapted to cardiac tissue. For imaging–omics integration, no single library yet covers the full pipeline from imaging feature extraction through molecular encoding to cross-modal fusion and evaluation; most teams assemble bespoke pipelines by combining these components with custom code. Designing modular, open-source implementations that separate encoders, fusion layers, and evaluation modules—and that ship with pretrained weights where licensing permits—will be important to make imaging-anchored multiomics analyses transparent and reusable across cardiovascular centres.

Representative frameworks and exemplar studies that instantiate these integration strategies in cardiovascular settings are summarized in Table 5.

Integrating imaging with transcriptomics and spatial transcriptomics

Radiogenomic modelling

Radiogenomics links imaging features with molecular measurements to identify associations between non-invasive phenotypes and underlying biology. In oncology, radiogenomic analyses have associated radiomic signatures extracted from CT and MRI with tumour gene-expression programmes, mutation status and pathway activity, establishing a precedent for non-invasive molecular characterization from imaging [97, 98]. Analogous ideas are now emerging in cardiovascular medicine, where the aim is to connect imaging-derived features—regional LGE burden, wall thickness, strain patterns, plaque attenuation profiles—with transcriptomic, proteomic, or pathway-level molecular readouts [99]. A systematic review of cardiac CT- and CMR-based radiogenomics found that, while radiomic features showed good diagnostic accuracy for plaque and myocardial characterization, very few studies had yet exploited full radiogenomic integration in cardiology, highlighting a substantial gap that imaging-anchored multiomics aims to address [100].

At its simplest, radiogenomic analysis treats imaging-derived features as predictors and gene-expression or pathway scores as outcomes, or conversely uses molecular profiles to predict imaging phenotypes. Regression, correlation, and enrichment analyses can identify genes and pathways whose expression covaries with imaging phenotypes such as myocardial fibrosis burden or plaque vulnerability. The most developed cardiovascular example to date is the radiotranscriptomic mapping of perivascular fat by Oikonomou *et al.* [24, 39], who trained machine-learning models to map coronary CTA radiomic features of perivascular adipose tissue onto tissue-level gene-expression modules obtained from matched biopsies, deriving a radiotranscriptomic signature that improved cardiac risk prediction beyond standard CT and clinical scores. This study illustrates the core radiogenomic workflow: imaging features define a non-invasive phenotype, paired tissue profiling provides the molecular ground truth, and a learned mapping enables molecular inference from imaging alone in subsequent patients without biopsies. The major radiogenomic, spatial-alignment, and virtual-transcriptomics pipelines covered in this section are shown in Fig. 3.

More sophisticated models place imaging and transcriptomics into a joint latent space. DCCA and deep canonical autoencoders learn non-linear transformations that maximize cross-modal correlation [78, 79]. Multimodal VAEs with product-of-experts inference can generate imaging-like representations from transcriptomic inputs and infer transcriptomic profiles from imaging, providing a generative linking model [80, 81]. These approaches are particularly useful when imaging and omics are not perfectly aligned in time—a common situation in cardiology, where late-stage transcriptomic profiles from explanted or surgically sampled hearts must be related back to earlier imaging phenotypes captured months or years prior. The cross-modal autoencoder of Radhakrishnan *et al.* [66] demonstrated this temporal bridging at population scale by learning a shared latent space across ECG, CMR, and clinical variables in the UK Biobank, enabling pheno-

type prediction and genotype association in the latent space despite non-simultaneous data acquisition.

Graph-based radiogenomics introduces biological priors by representing genes, pathways, and imaging features as nodes in a graph, with edges encoding known relationships such as protein–protein interactions, pathway co-membership, or anatomical adjacency. GNNs trained on such graphs can propagate information between molecular and imaging nodes, learning which molecular programmes best explain regional imaging patterns. To illustrate, consider hypertrophic cardiomyopathy (HCM), where the individual components for such a graph already exist in the literature: sarcomeric gene variants in *MYBPC3* and *MYH7* have been associated with distinct patterns of LGE extent and myocardial fibrosis on CMR [32, 101], and single-cell transcriptomics of HCM myectomy tissue has revealed cell-type-specific gene networks—including stress, hypertrophic, and fibrotic programmes in cardiomyocytes and fibroblasts—that drive the remodelling phenotype [102]. A graph-based radiogenomic model could tie together genotype nodes (sarcomeric variant status), imaging nodes (regional LGE distribution, wall thickness, and strain) and transcriptomic nodes (fibroblast activation and cardiomyocyte stress programmes) to explain why specific genotypes produce characteristic imaging patterns and to distinguish pathogenic from benign variants—although such an end-to-end integrative model has not yet been demonstrated in HCM and remains a methodological opportunity.

Causal frameworks such as Mendelian randomization extend radiogenomics beyond association. Genetic variants associated with biomarkers or gene-expression modules can be tested as instruments for imaging traits, probing whether variation in a molecular mediator causally influences an imaging phenotype. Conceptually, this allows chains such as genotype → transcriptomic module → imaging phenotype → outcome to be decomposed and quantified [103]. These analyses are limited by instrument strength and pleiotropy but add a layer of mechanistic interpretation beyond purely correlational radiogenomic studies, and are increasingly feasible at scale as biobank-linked imaging and omics cohorts grow [17, 38]. Recent work demonstrates the potential of such scale: a multitask deep learning framework trained on UK Biobank proteomics and metabolomics data predicted six common CVDs up to 15 years before clinical onset and identified disease-relevant proteins and metabolites as candidate biomarkers [104]. These developments underscore that the computational and data infrastructure for cardiovascular radiogenomics at population scale is maturing rapidly, even as paired imaging–transcriptomic cohorts remain the principal bottleneck.

Spatial molecular alignment and tissue–image correspondence

Linking spatial transcriptomics to *in vivo* imaging requires careful alignment across scales: from MRI voxels or echocardiographic pixels to histology, from histology to array capture spots and from spots to inferred cell types [105]. Experimental design is crucial. In some studies, tissue blocks are sampled from imaging-identified regions such as LGE-positive and LGE-negative segments, which establishes a coarse correspondence between imaging and spatial maps [12, 13]. In others, high-resolution *ex vivo* MRI or micro-CT is acquired on explanted hearts as an intermediate for registration, bridging the resolution gap between *in vivo* imaging and the histology section on which spatial transcriptomics is performed [11].

Table 5 Representative frameworks and studies integrating cardiovascular imaging with molecular or other high-dimensional omic data, including general-purpose frameworks that have been applied or are directly applicable to imaging-anchored multiomics and cardiovascular studies that exemplify concrete imaging–omics applications in cardiology.

Framework/study	Modalities integrated	Integration strategy	Cardiovascular relevance - example use	Key contribution
<i>General-purpose multimodal frameworks applicable to imaging–omics integration</i>				
MOFA+, scVI/totalVI, and scvi-tools [50, 51, 92]	Bulk and single-cell transcriptomics, epigenomics, proteomics; imaging-derived features can be included as an additional view	Probabilistic factor models (MOFA+) and hierarchical VAEs (scVI) that learn joint low-dimensional representations while explicitly modelling batch effects and missing data	Provide molecular latent spaces—shared factors or scVI latent variables—that can be aligned with cardiac imaging embeddings for radiogenomic and spatial integration	Established ecosystem with documented APIs; standard entry point for cardiovascular multiomics before fusion with imaging
Multimodal VAEs [80, 81]	Heterogeneous combinations of imaging, omics, and clinical data	Product-of-experts or mixture-of-experts aggregation of modality-specific encoder posteriors into a shared latent space; joint posterior remains well-defined when modalities are missing	Prototype framework for learning shared imaging–omics latent spaces; supports “virtual omics” generation from imaging inputs alone	Handle missing modalities by design; probabilistic formulation enables uncertainty quantification and cross-modal simulation
PORPOISE [87]	Whole-slide H&E histopathology and multi-omic profiles (RNA-seq, CNV, mutations) across 14 TCGA cancer types	Attention-based multiple instance learning (MIL) for whole-slide images (WSIs) fused with a self-normalizing network for omics via Kronecker-product gated fusion; weakly supervised with survival labels	Architecturally transferable to cardiac histology–omics; interpretable morpho-molecular prognostic markers at pan-cancer scale	First large-scale interpretable histology–genomic fusion; open-source code and interactive atlas; highlighted batch-effect risks in site-unaware splits [88]
HEALNet [82]	Whole-slide histopathology and multi-omic data (four TCGA cancer datasets)	Iterative cross-attention between a shared latent bottleneck and modality-specific token streams; processes raw inputs rather than precomputed embeddings	Directly applicable to cardiac tissue sections with matched omics; handles missing modalities at both training and inference	State-of-the-art survival prediction among end-to-end fusion models; open-source; modular design facilitates addition of imaging modalities
<i>Cardiovascular imaging–omics studies</i>				
Cross-modal autoencoder for cardiovascular state [66]	Twelve-lead ECG, CMR cine images and clinical traits from the UK Biobank	Cross-modal autoencoder reconstructing each modality from a shared latent vector; supports imputation of missing modalities and genotype association in latent space	Phenotype prediction, unsupervised GWAS, and CMR imputation from ECG at population scale ($n > 40,000$)	Scalable cross-modal representation learning linking low-cost ECG signals to rich CMR phenotypes and genetic architecture
MODES [83]	Twelve-lead ECG and CMR from the UK Biobank	Representation fusion framework decoupling shared and modality-specific latent components via orthogonality constraints	Inference of missing MRI phenotypes from ECG alone; shared latent captures interpretable cardiac structure–function relationships	Explicit shared/private decomposition improves cross-modal prediction and clinical interpretability in cardiovascular data
Radiotranscriptomic perivascular fat signature [24, 39]	Coronary CTA radiomics of perivascular fat and microarray gene-expression from perivascular adipose biopsies	Machine-learning mapping of CT radiomic features onto tissue gene-expression modules to derive a radiotranscriptomic risk score	Prediction of cardiac mortality and MI beyond standard CT and clinical risk scores in coronary CTA cohorts	Proof-of-concept that CT radiomics can serve as a non-invasive surrogate for local vascular inflammatory and fibrotic gene-expression states
Spatial multi-omic map of human myocardial infarction [12]	snRNA-seq, snATAC-seq, Visium spatial transcriptomics, and histology of infarcted and control myocardium	Joint embedding of single-cell and spatial omics to annotate infarct core, border and remote zones with dominant cell states and regulatory programmes	Mechanistic dissection of human MI remodelling across regions and time points; tissue blocks sampled from imaging-defined zones	Tissue-scale molecular atlas co-registrable with imaging markers of infarct size, scar distribution and remodelling
Coronary plaque spatial transcriptomics linked to imaging [29, 35, 36]	Spatial transcriptomics of human coronary plaques with matched histology; invasive or CT imaging in selected cohorts	Spatial co-localization of gene-expression modules and immune niches with plaque morphology and CT features such as low-attenuation plaque or thin-cap fibroatheroma	Characterization of high-risk plaque phenotypes and local immune microenvironments in coronary artery disease	Template for connecting coronary imaging markers of vulnerability with spatially resolved arterial wall gene-expression and immune niches

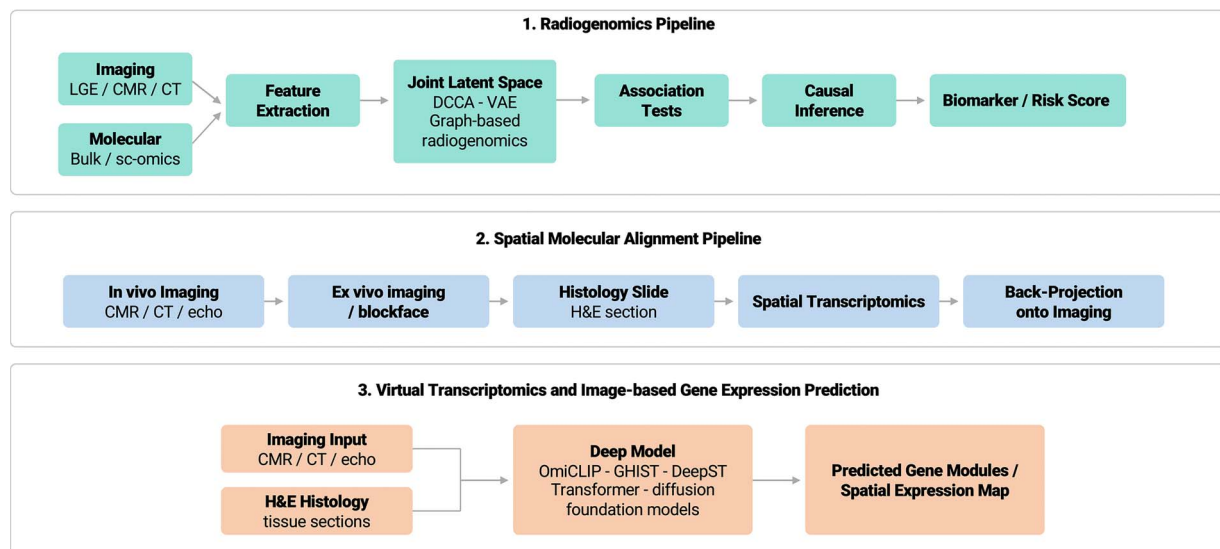


Figure 3 Archetypal pipelines for imaging–omics integration. Radiogenomics workflows link imaging-derived features to bulk or single-cell molecular profiles via feature extraction, joint latent modelling, and association or causal inference; spatial molecular alignment pipelines register *in vivo* imaging, *ex vivo* imaging, histology and spatial transcriptomics into a common coordinate system; virtual transcriptomics approaches use deep models to predict molecular modules or spatial expression maps directly from imaging inputs.

Computationally, tissue–image alignment often proceeds in stages. First, histology slides are registered to block-face photographs or *ex vivo* imaging using affine and non-linear transformations. Second, the spatial transcriptomics capture grid is mapped onto the histology image using fiducial markers or deep-learning-based landmark detection algorithms that can handle nonlinear tissue deformations across modalities [93, 106]. Third, *in vivo* imaging is registered to *ex vivo* imaging or to 3D histology reconstructions via anatomical landmarks and deformable warping. At each step, uncertainties arise from tissue deformation, sectioning artefacts, and differences in imaging physics [11]. Representing these uncertainties explicitly, through probabilistic registration or ensembles of transformations, is important when assessing spatial correlation between imaging and molecular features. Recent work on integrating tissue morphology with spatial transcriptomics further highlights how histology foundation models can enhance alignment and spatial domain identification, although their application to cardiac tissue remains nascent [107].

Once correspondence is established, imaging-derived quantities such as local LGE intensity, strain, or CT attenuation can be sampled at projected spot coordinates, producing paired vectors of imaging and expression features for each spatial location. Spatial graph encoders can then jointly embed these paired features, allowing discovery of spatial domains that are coherent in both imaging and molecular space and aiding identification of structures such as border zones, fibrotic microdomains, or vulnerable plaque shoulders [12, 29].

Image-based gene-expression prediction and virtual transcriptomics

An emerging direction is to predict molecular profiles directly from images, often termed virtual transcriptomics, or image-based gene-expression imputation. In atherosclerosis, models trained on coronary CT angiography and matched plaque gene-expression profiles have predicted expression of inflammatory and matrix-degrading genes from CT alone, effectively performing a non-invasive molecular

biopsy of the plaque [39]. Similar approaches in oncology predict bulk or spatial transcriptomic patterns from radiology or histology images [44, 45, 108].

For cardiac applications, myocardial or plaque gene-expression modules could be predicted from imaging features such as LGE patterns, T1 maps, or strain fields. This would enable large-scale molecular phenotyping in cohorts where only imaging is available. Challenges include the high dimensionality of gene-expression outputs, often addressed by predicting pathway scores or modules rather than individual genes, limited numbers of paired image–omics samples and domain differences between training and deployment data. A recent benchmark of 11 histology-to-expression prediction methods confirmed that H&E-based virtual transcriptomics is feasible across tissue types, while identifying batch effects and limited cross-dataset generalizability as key remaining challenges [61]. Several deep learning methods predict high-dimensional spatial transcriptomic profiles directly from histology, often using transformer or diffusion backbones trained in a semi-supervised manner on large spatial atlases [62, 63]; vision–omics foundation models such as OmiCLIP further demonstrate that spatial gene expression can be predicted at Visium spot level from H&E patches alone after training on millions of paired image–transcriptome samples across 32 organs [44], and GHIST achieves single-cell-resolution gene-expression prediction from histology using deep learning [108]. These developments demonstrate that high-fidelity virtual transcriptomics is technically feasible and provide architectures that can be adapted to cardiac histology, CMR, or CT once joint image–omics datasets reach adequate scale.

Applications: subtyping, mechanism, biomarkers, and early clinical case studies

Integrative imaging–transcriptomic pipelines support four recurring goals: disease subtyping, mechanistic insight, biomarker discovery and virtual transcriptomics, and early clinical translation.

Disease subtyping

Clustering in cross-modal latent spaces allows grouping of patients or tissue regions based on joint imaging and molecular characteristics. In heart failure, integrating echocardiographic phenotypes with transcriptomic or proteomic modules has revealed phenogroups that differ both in structure and in underlying biology; one subgroup may have predominantly inflammatory signatures, whereas another is characterized by fibrotic changes [25]. Spatial transcriptomics of failing human myocardium has further resolved region-specific molecular programmes within hearts that appear homogeneously dilated on imaging, revealing distinct spatial domains of fibrosis, inflammation, and metabolic remodelling [28]. Spatially, clustering of spots based on combined imaging and expression features can identify microdomains, such as stressed border-zone myocardium, that are not evident from histology or gene expression alone [12].

Mechanistic discovery

Radiogenomic and spatial analyses can suggest mechanisms for imaging findings. Spatial transcriptomics of infarct border zones linked to LGE and strain abnormalities has revealed co-localized programmes of fibroblast activation, angiogenesis, and immune-cell recruitment [12, 13]. In atherosclerotic plaques, combining CT features with spatial expression maps has implicated specific macrophage and smooth muscle cell states in regions corresponding to low-attenuation plaque and thin fibrous caps [29, 35]. Such findings help explain which cell types and pathways drive particular imaging phenotypes and how they might be targeted therapeutically.

Biomarker discovery

Imaging phenotypes that predict outcomes can be mined for molecular correlates, yielding candidate biomarkers. When a specific pattern of non-ischaeamic fibrosis on CMR is strongly associated with arrhythmic events, imaging–omics integration can identify gene-expression modules or cell-state proportions enriched in regions with that pattern. These molecular signatures can then be sought in blood-based assays—such as plasma proteomics, where protein-based risk scores have already demonstrated incremental prediction of cardiovascular events beyond classical factors [26]—or used to stratify patients in clinical trials. A recent multitask deep learning framework integrating proteomics and metabolomics from the UK Biobank predicted six common CVDs ~15 years before onset and identified disease-relevant circulating biomarkers [104], illustrating how multiomics-derived molecular signatures can complement imaging phenotypes for early risk stratification. Conversely, known circulating biomarkers can be traced back to spatial imaging patterns, clarifying which structural or cellular changes they reflect.

Early clinical case studies

A notable example of imaging-anchored molecular inference is the perivascular FAI, developed from coronary CTA as a surrogate for local coronary inflammation and shown to predict residual cardiovascular risk [24]. Subsequent work derived a radiotranscriptomic signature by mapping perivascular adipose gene-expression modules to CTA radiomic features, improving risk prediction beyond standard models in coronary CTA cohorts [39]. In a complementary direction, cross-modal representation learning at biobank scale has shown that a shared latent space across ECG and CMR can enable MRI imputation and genetic association studies in the latent space [66], and the MODES framework further demonstrated that explicitly decoupling shared

and modality-specific latent components improves cross-modal prediction of missing MRI phenotypes from ECG alone [83]. Together, these case studies illustrate that (i) imaging can act as a proxy for local biology in specific contexts, and (ii) large-scale paired datasets can support clinically relevant cross-modal phenotyping.

Predictive learning, risk models, and reporting considerations

Imaging-anchored multiomics can be used for mechanistic biology, but many translational applications ultimately require *predictive* models: outcome prediction, treatment response, and risk stratification. Unimodal AI-based predictors have already demonstrated that routine clinical signals carry far more prognostic information than conventional interpretation extracts: deep neural networks trained on resting 12-lead ECGs predict one-year all-cause mortality with an AUC of 0.88, including among tracings read as “normal” by cardiologists [109]; analogous models detect paroxysmal atrial fibrillation from sinus-rhythm recordings [110]; and the AIRE platform extends AI-ECG prediction to multiple actionable cardiovascular endpoints with transnational external validation [111]. In echocardiography, multitask deep learning now automates comprehensive interpretation across diverse pathologies, with a multi-view vision–language foundation model achieving state-of-the-art performance on 23 benchmarks of cardiac form and function [112]. These unimodal successes establish the ceiling that multimodal fusion must surpass to justify its additional complexity.

Multimodal and multi-omic predictive models are beginning to meet this bar. A machine-learning marker for coronary artery disease derived from electronic health records and validated in two independent biobank cohorts quantified atherosclerotic burden on a continuous scale and identified underdiagnosed individuals at elevated mortality risk [113]. In hypertrophic cardiomyopathy, the MAARS model fused electronic health records, echocardiographic and radiology reports, and contrast-enhanced CMR through transformer-based networks to forecast lethal ventricular arrhythmia with an AUC of 0.89 internally and 0.81 externally, outperforming current guideline-based risk calculators [114]. At the multiomics level, integrating genomics, transcriptomics, proteomics and metabolomics into a single predictive framework has demonstrated that each omic layer contributes non-redundant information to cardiovascular risk, with the combined model significantly outperforming any single modality [104]. Collectively, these studies show that multimodal fusion adds genuine prognostic value when validated rigorously.

Four practical points are central to translating these models. First, multimodal models should be evaluated not only for discrimination but also for *calibration and clinical utility*. A model that achieves high AUC but assigns risk probabilities that deviate systematically from observed event rates offers limited bedside value. Calibration determines whether predicted risks are meaningful in clinical decision-making; decision-curve analysis quantifies whether a model provides net benefit at clinically relevant thresholds [89]. External validation across centres and demographic subgroups is essential, because multimodal histology–genomic models in oncology have been shown to overestimate performance when site-of-origin batch effects are not controlled through site-preserved cross-validation [88]—a risk equally acute in cardiovascular settings where specific imaging centres often provide the only tissue samples. Leave-one-site-out designs and negative controls with permuted pairings should be standard practice for

imaging–omics models, as discussed in the “Bottlenecks, root causes, and failure modes” section.

Second, *validation should be stratified*. Performance should be reported across clinically meaningful subgroups—by sex, ancestry, disease severity, and imaging centre—to ensure that the model does not preferentially benefit well-represented subpopulations while failing for others. The radiotranscriptomic perivascular fat signature, for example, was validated in independent coronary CTA cohorts and shown to improve risk prediction beyond standard clinical and imaging scores [39], while the cross-modal cardiovascular autoencoder was evaluated across tens of thousands of UK Biobank participants for phenotype prediction and genotype association [66]. These examples illustrate the scale and rigour of validation needed for clinical translation.

Third, *reporting and risk-of-bias assessment* for predictive models should follow emerging AI-specific guidance. TRIPOD+AI extends the TRIPOD statement to prediction models developed using machine learning [90], and PROBAST+AI adapts risk-of-bias assessment to AI-driven prediction modelling [91]. For multimodal imaging–omics models, these frameworks emphasize transparent reporting of data provenance, missingness mechanisms, preprocessing, model selection, and validation strategy. Emerging guidance for generative clinical AI further highlights the need for detailed documentation of intended use and limitations [115].

Fourth, as multimodal architectures scale from unimodal baselines through pairwise fusion to full imaging–omics integration, the taxonomy of fusion challenges—representation, alignment, translation, and co-learning [77]—provides a useful design checklist. Each added modality increases data requirements and potential failure modes; demonstrating that the multimodal model outperforms each unimodal component, not just a naïve concatenation, remains the minimum standard for claiming added value.

Bottlenecks, root causes, and failure modes

Sample sizes of deeply phenotyped patients with paired imaging, bulk, single-cell, and spatial data remain small—typically tens to low hundreds of patients—which limits statistical power and encourages overfitting. Cohort composition often reflects specific indications, such as transplant or surgical patients, which may restrict generalizability to broader cardiovascular populations. Variability in tissue sampling relative to imaging, such as sampling only a single septal biopsy in a heart with patchy disease, can bias associations; CMR-guided endomyocardial biopsy improves spatial targeting [14] but remains technically demanding and is not standard practice. Spatial misalignment and tissue distortion from fixation, sectioning, and mounting can introduce errors in spot-to-voxel mapping that blur true correlations, a challenge well characterized in the spatial transcriptomics registration literature [105, 106]. These challenges are shared broadly across multimodal biomedical AI, where data heterogeneity, missingness, and privacy constraints compound modelling difficulty at every stage from representation to deployment [116].

Many failure modes trace back to a small set of root causes; below, we pair each with practical mitigations.

(i) Mis-specified correspondence across scales. Even small registration errors—a few voxels in MRI, or rotational offsets in histology—can wash out true imaging–molecular relationships, particularly for spatially heterogeneous phenotypes such as patchy fibrosis or border-zone remodelling. In myocardial infarction studies, infarct core,

border zone, and remote regions lie within millimetres of each other [12], making registration fidelity critical. *Mitigations* include explicit uncertainty modelling through ensembles of registrations, sensitivity analyses that quantify how results change under plausible warps, deep learning-based spatial landmark detection for automated tissue alignment [106], and experimental designs that sample tissue using imaging-defined targets—such as LGE-positive versus LGE-negative segments—whenever possible [14].

(ii) Confounding and shortcut learning. Site, scanner, and processing batch can dominate latent factors, particularly when paired sample sizes are limited. Multimodal histology–genomic deep learning models such as pan-cancer integrative frameworks [87] have been shown to overestimate performance when site-of-origin batch effects are not controlled through site-preserved cross-validation [88]; the same risk applies to cardiovascular cohorts where one centre contributes all paired imaging–spatial data. Imaging domain shifts arise from differences in scanners, field strength, acquisition protocols, and reconstruction algorithms; deep learning harmonization methods for structural MRI are developing but remain less mature than omics batch correction [72]. *Mitigations* include batch-aware latent models such as ComBat [73] and Harmony [74] for omics, domain-adversarial training for imaging, leave-one-site-out evaluation, and negative controls with randomized pairings or permuted spatial coordinates.

(iii) Modality imbalance and missingness. Deep fusion models can collapse onto the strongest modality (typically imaging, which has the largest sample size and richest spatial detail) or exclude patients with missing data, biasing the cohort toward fully profiled subsets. In practice, a cohort may have CMR on thousands of patients, bulk RNA-seq on hundreds, and spatial transcriptomics on dozens; naïve training on the intersection discards most of the imaging data. *Mitigations* include product-of-experts or mixture-of-experts inference that naturally handles variable input subsets [80, 81], modality dropout during training (randomly masking one modality per batch to prevent co-adaptation), explicit missingness indicators, and late-fusion baselines that combine predictions from whichever modalities are available. Leveraging large imaging-only biobanks alongside smaller imaging–omics subsets is a key design strategy [66].

(iv) Over-interpretation of correlational links. Radiogenomic associations can be mistaken for causality. A strong correlation between an imaging feature and a gene-expression module may reflect shared confounders (e.g. disease severity driving both) rather than a mechanistic link. *Mitigations* include replication in independent cohorts, triangulation with genetic instruments via Mendelian randomization where feasible [103], experimental follow-up in model systems, and careful separation of discovery and validation datasets to prevent circular reasoning.

(v) Reproducibility gaps. Heterogeneous preprocessing and analysis choices—segmentation algorithms, normalization strategies, gene-filtering thresholds, spot quality-control criteria—can alter conclusions substantially. Spatial transcriptomics pipelines are particularly sensitive to the choice of deconvolution method and reference atlas [11]. Emerging foundation models that couple imaging, text and structured data across tasks [8] will require particularly rigorous auditing given their scale and opacity. *Mitigations* include workflow tooling with containerized environments, versioned preprocessing, public code and model cards, and adherence to reporting guidance for predictive models such as TRIPOD+AI and PROBAST+AI [90, 91]. A general taxonomy of multimodal learning challenges—representation,

alignment, translation, fusion, and co-learning [77]—provides a useful checklist for identifying where reproducibility investments are most needed.

These root causes are not independent: small sample sizes (the overarching constraint) amplify every other failure mode, making confounding harder to detect, missingness harder to impute, and registration errors harder to average out. Addressing them requires coordinated advances in cohort design (multi-centre, imaging-guided tissue sampling), computational methodology (robust fusion under missingness and batch effects), and community standards (open data, shared preprocessing pipelines, and standardized evaluation protocols).

Multimodal transformers and agentic systems

Multimodal transformers for imaging–omics

Multimodal transformers generalize fusion by representing heterogeneous inputs as tokens processed by self-attention. Imaging patches, gene modules, spatial spots, and clinical variables can each be projected into a shared token space with modality-specific and positional encodings, then integrated via shared attention layers that learn cross-modal interactions without requiring hand-crafted feature alignment [8, 116]. Masking enables training and inference with variable subsets of modalities—a patient with only CMR and bulk RNA-seq receives predictions from the same model that handles a fully profiled case with additional spatial transcriptomics—while attention weights can highlight influential imaging regions, pathways, or cell types, offering a route to *post-hoc* interpretability.

Surveys emphasize these architectures as a core component of emerging medical foundation models that integrate imaging, text and structured data [8, 9, 69, 70]. In oncology, multimodal transformers that tokenise histology patches alongside genomic features have demonstrated improved survival prediction across cancer types [87]; the HEALNet architecture further shows that heterogeneous biomedical modalities can be fused through iterative cross-attention without requiring all modalities at test time [82]. In cardiology, transformer-based networks underpin both the MAARS model for arrhythmic death prediction from CMR, echocardiographic reports, and electronic health records [114], and EchoPrime, a view-informed model that integrates multiple echocardiographic views for comprehensive study-level interpretation [112]. A vision–language foundation model for echocardiography further demonstrates how contrastive pretraining on paired echo videos and clinical text can yield transferable cardiac representations [43].

In parallel, single-cell and spatial foundation models pretrained on tens of millions of cells—such as scGPT [65] and Nicheformer [54]—provide transferable molecular encoders whose embeddings can serve as token inputs to a multimodal transformer alongside imaging tokens, bridging the resolution gap between voxel-level imaging and cell-level transcriptomics.

Despite these advances, multimodal transformers remain data-hungry and computationally intensive relative to typical paired cardiovascular cohort sizes. A pragmatic strategy is to pretrain unimodal transformers (imaging, single-cell) using self-supervision on large unlabelled archives, then train a lighter cross-modal attention or fusion transformer on the available paired data. This modular approach allows each encoder to benefit from the scale of its own modality while limiting the parameter count of the fusion layers to what the paired subset can support.

Agentic and large language model-based orchestration

Large language models (LLMs) are increasingly used to orchestrate complex analysis pipelines by calling specialized tools. In an imaging-anchored multiomics workflow, encoders, fusion models, and statistical tests can be wrapped as callable tools, and an LLM-based agent can coordinate data retrieval, model execution, result interpretation, and report drafting [117]. Such orchestration is particularly attractive for imaging–omics pipelines that span multiple software ecosystems—imaging frameworks such as MONAI [95], single-cell toolkits such as scvi-tools [51], and spatial analysis libraries such as Squidpy [93]—where an agent can automate the handoffs between preprocessing, encoding, fusion, and evaluation.

LLMs should not replace validated integration models: current limitations including hallucinations, brittle multi-step reasoning, and lack of calibrated uncertainty motivate constrained tool use, structured logging, and mandatory human oversight at decision points [118]. Regulatory considerations, including the European Union AI Act and guidance for regulated digital medical products, further emphasize traceability, auditability, and human oversight for clinical-facing agentic systems [119]. The MI-CLAIM-GEN checklist for generative clinical AI provides a reporting framework that is directly applicable to agentic pipelines that generate clinical interpretations [115].

Cross-domain methodological inspirations

Design patterns from adjacent biomedical AI domains can inform imaging–omics integration. Dual-feature attention fusion, developed for anticancer natural product detection, demonstrates how two-sided attention across heterogeneous feature sets can capture cross-modal synergies [120]. Explicit part–whole anatomical priors, introduced for scene parsing, offer a template for encoding the hierarchical organization of cardiac anatomy (chambers, walls, and segments) into fusion architectures [121]. Systematic perspectives on molecular representations and evaluation in generative modelling highlight best practices for assessing the fidelity of predicted molecular outputs [122], which are directly relevant when virtual transcriptomics models generate gene-expression profiles from imaging inputs. The broader taxonomy of multimodal machine learning challenges—representation, alignment, translation, fusion, and co-learning [77]—provides an organizing framework that unifies these cross-domain inspirations and helps identify which design patterns are most transferable to cardiovascular imaging–omics.

Practical considerations

Data cleaning, harmonization, and overall principles

Robust imaging–omics integration starts with rigorous data curation. For imaging this includes quality control, consistent reconstruction, and careful segmentation or landmarking with clinical oversight; frameworks such as MONAI [95] and TorchIO [96] provide standardized pipelines for medical image preprocessing, augmentation, and patch-based sampling that improve reproducibility across sites. For bulk and single-cell data, curation includes alignment and quantification, filtering low-quality cells and genes, and appropriate normalization; toolkits such as Scanpy [48] and scvi-tools [51] offer end-to-end workflows with integrated quality-control metrics. Spatial transcriptomics data add further requirements: tissue coverage checks,

spot-level quality control, accurate registration to paired histology, and careful choice of deconvolution method, all of which can materially alter downstream conclusions [11].

Harmonization is critical across modalities. Omics pipelines commonly use ComBat [73], Harmony [74], or joint latent models such as scVI [50] to reduce technical variation while preserving biological signal. Imaging harmonization is less standardized but may include intensity normalization, resampling, and domain-adversarial training to reduce scanner-specific bias; deep learning-based approaches for structural MRI harmonization are developing but not yet mature [72]. Across all steps, adherence to Findable, Accessible, Interoperable, Reusable principles and rich metadata—acquisition protocols, preprocessing versions, quality-control metrics, and provenance records—improve reproducibility, reuse, and the ability to diagnose failures *post hoc*.

Computational demands, efficiency, and scalability

Compute constraints shape feasible designs and should be reported alongside model performance. Radiomics and regression-style radiogenomics can run on modest hardware, while 3D/4D imaging encoders, multimodal transformers, and large spatial graph models may require multi-GPU training and significant storage for intermediate features and graphs. The gap between the computational envelope of foundation models—pretrained on millions of images or cells [8, 53]—and the resources available to most cardiovascular research groups is substantial.

Practical approaches to bridge this gap include: (i) pretraining unimodal encoders once on large archives and freezing them during fusion, so that only lightweight cross-modal layers are trained on scarce paired data; (ii) operating at clinically meaningful intermediate resolutions—American Heart Association segments or spatial domains rather than individual voxels or spots—when paired data are limited; (iii) predicting pathway or module-level scores rather than thousands of individual genes, which reduces output dimensionality and improves interpretability; and (iv) distilling large models into smaller, deployable surrogates for clinical integration. Reporting training compute and memory footprints, along with lightweight baselines, improves feasibility assessment, and reproducibility [90].

Cross-centre variability and deployment

Multicentre cohorts differ in demographics, disease mix, imaging hardware, and omics platforms; single-centre models often fail to generalize [116]. This is not merely a technical inconvenience: in cardiovascular imaging-omics, a model trained at a single transplant centre may learn features specific to end-stage disease and advanced immunosuppression that do not transfer to community-level screening. Pipelines should incorporate site and batch covariates, domain-adaptation where appropriate, and external validation following TRIPOD+AI guidance [90]. Federated learning can reduce barriers to cross-site training by keeping data local, but must handle statistical heterogeneity across sites and the operational complexity of coordinating preprocessing and model updates.

Cross-modal autoencoders trained at biobank scale have demonstrated that shared latent spaces across ECG, CMR, and clinical variables can scale to tens of thousands of participants when designed for partially paired data [66], providing a template for how large imaging-only cohorts such as UK Biobank [17] can be leveraged

alongside smaller imaging-omics subsets. Deployment additionally favours models that are calibrated, interpretable, efficient, and maintainable. In some settings, simplified surrogate scores distilled from deep models may be necessary for clinical integration, particularly when real-time inference or point-of-care use is envisaged.

Bias, ethics, and equity

Multimodal datasets can reflect and amplify healthcare disparities. Under-representation of women, minority ethnic groups and low-resource settings can lead to biased models or unreliable molecular interpretations—a concern that extends across all of biomedical AI [116] but is particularly acute in imaging-omics, where the most deeply phenotyped cohorts tend to be drawn from well-resourced academic centres in high-income countries. Performance should be assessed across demographic and clinical subgroups, and uncertainty should be reported alongside predictions; multimodal models that appear accurate in aggregate may fail systematically for under-represented groups, as has been documented for batch-confounded histology-genomic models in oncology [88].

Transparent documentation of intended use, data provenance, and limitations is essential. Emerging guidance for AI-driven prediction models—TRIPOD+AI [90] and PROBAST+AI [91]—and for generative clinical AI [115] provide reporting frameworks that are directly applicable to imaging-anchored multiomics. For this field, avoiding over-interpretation of correlational links and ensuring that clinical benefit generalizes across populations are core ethical requirements that should be treated as design constraints rather than afterthoughts.

Future directions

Three priorities define the next phase: continuous spatial modelling, imaging-informed molecular phenotyping, and clinically validated integration with digital twins. Fig. 4 outlines the pathway from imaging-anchored models to decision support and in silico trials.

Spatial modelling and cross-modal generative models

An important next step is to treat the heart as a continuous 3D object with coupled imaging and molecular fields. Spatial statistical models and Gaussian processes defined on anatomical domains can capture smooth gradients of gene expression and imaging phenotypes, enabling interpolation between sampled locations and joint analysis of structure and function. As spatial atlases grow, such models may link mechanical forces, perfusion, and molecular responses at higher resolution.

Conditional cross-modal generative models could simulate plausible molecular profiles from imaging and vice versa, supporting uncertainty-aware imputation and counterfactual experiments. Incorporating physiological constraints and mechanistic priors remains an open challenge but would improve realism and trustworthiness.

Imaging-informed molecular phenotyping and foundation models

A long-term goal is to infer coarse molecular phenotypes from routine imaging with calibrated uncertainty. Achieving this at scale likely requires foundation models that combine large imaging archives with smaller paired imaging-omics cohorts. Self-supervised pretraining on millions of cardiac images followed by fine-tuning on

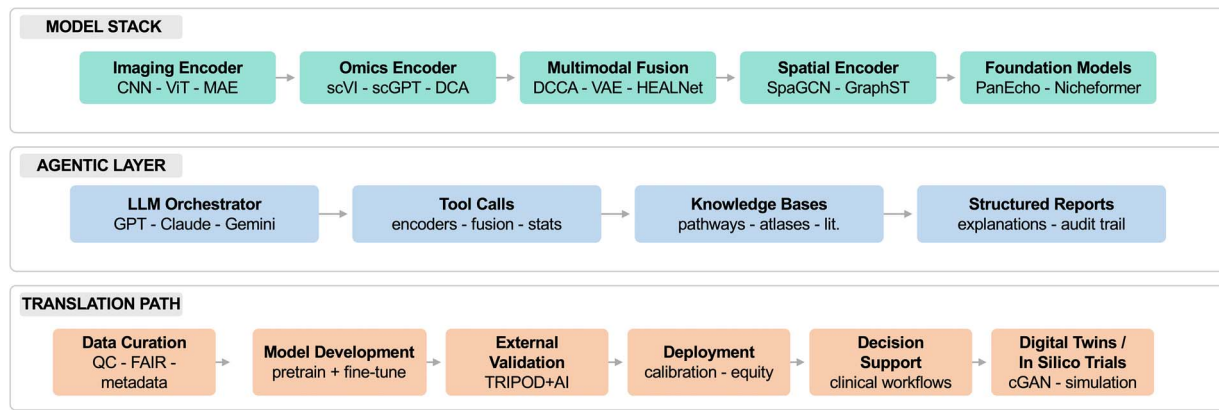


Figure 4 Translational roadmap from imaging-anchored multiomics to digital twins. Top lane: model stack comprising modality-specific encoders, multimodal fusion models and transformer or foundation model architectures. Middle lane: an agentic layer in which large language models orchestrate tool calls to encoders, fusion models and knowledge bases to generate structured reports and explanations. Bottom lane: translational path from data curation and model development through external validation and deployment to decision support and digital twin-enabled *in silico* trials.

imaging–transcriptomic pairs could yield representations aligned to molecular programmes. In parallel, single-cell foundation models trained on tens of millions of cells [53, 54] provide transferable molecular encoders that can plug into cardiovascular fusion pipelines as atlas resources expand.

Digital twins and integrative simulation

Imaging-anchored multiomics is also converging with cardiac digital twins. Personalized computational models of anatomy and physiology derived from imaging can be augmented with patient-specific molecular context to simulate disease progression and therapy response [123–125]. Conversely, digital twin simulations can generate synthetic multimodal data for rare scenarios and support *in silico* trials [126, 127]. A complementary imaging-anchored example builds personalized 3D electrophysiological heart models directly from late-gadolinium-enhancement CMR: applied to post-infarction patients, these virtual-heart models simulate the propensity of each patient’s reconstructed myocardium to sustain re-entrant arrhythmia and stratify sudden-death risk more accurately than established clinical metrics [128]. Integrating such imaging-derived digital twins with imaging-anchored multiomics pipelines could enable *in silico* evaluation of how molecular subtypes identified by imaging–omics integration respond to specific interventions, bridging the gap between observational multiomics discovery and interventional evidence. More broadly, integrating data-driven and mechanistic models will require careful calibration and validation, but offers a path to multiscale precision medicine.

Conclusion

Imaging has long been the visual backbone of cardiovascular medicine. In the era of multiomics, it can also serve as the computational backbone: a spatial and temporal reference frame onto which transcriptomic and spatial transcriptomic data are mapped. This review has framed imaging-anchored molecular phenotyping as a central bioinformatics problem and outlined key methodological components required to address it: modality-specific representation learning, multimodal fusion via factor models, graphs, transformers,

and contrastive learning, and integrative pipelines for aligning imaging with bulk, single-cell, and spatial transcriptomics.

Radiogenomic and spatial transcriptomic studies in cardiology are beginning to show how imaging phenotypes—fibrosis patterns, strain abnormalities, and plaque morphologies—relate to specific cell types, pathways, and niches. Multimodal fusion methods provide cross-modal latent spaces where disease subtypes, mechanistic hypotheses, and candidate biomarkers can be explored systematically. At the same time, practical issues surrounding data quality, harmonization, batch effects, sample size, fairness, and reproducibility remain central and should be treated as core design constraints rather than afterthoughts.

Looking ahead, imaging-anchored multiomics is likely to evolve along three main trajectories: towards richer spatial models that integrate imaging and molecular fields, towards foundation models and generative approaches that scale molecularly informed imaging analysis to broader populations, and towards incorporation into digital twins and agentic systems that embed these capabilities into clinical workflows. Achieving this vision will require close collaboration between cardiologists, data scientists, biologists, and engineers, but the potential payoff is substantial: a transition from predominantly descriptive imaging towards a more mechanistic, molecularly guided understanding of cardiovascular health and disease.

Key Points

- Imaging-anchored multiomics learns joint representations linking cardiac imaging phenotypes to bulk, sc/snRNA-seq, and spatial transcriptomics.
- Cardiac imaging provides a spatial phenotype that supports voxel/region alignment with molecular states and cell-type programmes.
- Fusion methods can accommodate missing modalities and batch effects but remain constrained by small paired cohorts.
- Applications include radiogenomics, virtual transcriptomics, disease subtyping, mechanism discovery, and biomarker prioritization for cardiovascular disease.

Conflicts of interest

None declared.

Funding

This work was supported by the National Science and Technology Council, Taiwan (grant number NSTC114-2221-E-038-015).

Data availability

No new data were generated or analysed in this review. All datasets and software mentioned are from previously published studies or publicly available resources cited in the references.

References

- Global Burden of Cardiovascular Diseases and Risks 2023 Collaborators. Global, regional, and national burden of cardiovascular diseases and risk factors in 204 countries and territories, 1990–2023. *J Am Coll Cardiol* 2025;**86**:2167–2243.
- Litviňuková M, Talavera-López C, Maatz H *et al*. Cells of the adult human heart. *Nature* 2020;**588**:466–72.
- Tucker NR, Chaffin M, Fleming SJ *et al*. Transcriptional and cellular diversity of the human heart. *Circulation* 2020;**142**:466–82.
- Leiner T, Rueckert D, Suinesiaputra A *et al*. Machine learning in cardiovascular magnetic resonance: basic concepts and applications. *J Cardiovasc Magn Reson* 2019;**21**:61.
- Ouyang D, He B, Ghorbani A *et al*. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 2020;**580**:252–6.
- Zhou Z, Wang J. A narrative review of multimodal data fusion strategies for precision risk prediction in coronary artery disease: advances, challenges, and future informatics directions. *Rambam Maimonides Med J* 2025;**16**:e0023.
- Farah EN, Diaz JT, Bloomekatz J *et al*. Charting the cardiac landscape: advances in spatial transcriptomics for heart biology. *Semin Cell Dev Biol* 2025;**175**:103648.
- Moor M, Banerjee O, Abad ZSH *et al*. Foundation models for generalist medical artificial intelligence. *Nature* 2023;**616**:259–65.
- Tu T, Azizi S, Driess D *et al*. Towards generalist biomedical AI. *NEJM AI* 2024;**1**:A10a2300138.
- Lim HJ, Wang Y, Buzdin A *et al*. A practical guide for choosing an optimal spatial transcriptomics technology from seven major commercially available options. *BMC Genomics* 2025;**26**:47.
- Pentimalli TM, Karaïskos N, Rajewsky N. Challenges and opportunities in the clinical translation of high-resolution spatial transcriptomics. *Annu Rev Pathol* 2025;**20**:405–32.
- Kuppe C, Ramirez Flores RO, Li Z *et al*. Spatial multi-omic map of human myocardial infarction. *Nature* 2022;**608**:766–77.
- Calcagno DM, Taghdiri N, Ninh VK *et al*. Single-cell and spatial transcriptomics of the infarcted heart define the dynamic onset of the border zone in response to mechanical destabilization. *Nat Cardiovasc Res* 2022;**1**:1039–55.
- Unterberg-Buchwald C, Ritter CO, Reupke V *et al*. Targeted endomyocardial biopsy guided by real-time cardiovascular magnetic resonance. *J Cardiovasc Magn Reson* 2017;**19**:45.
- Lurz P, Eitel I, Adam J *et al*. Diagnostic performance of CMR imaging compared with EMB in patients with suspected myocarditis. *JACC Cardiovasc Imaging* 2012;**5**:513–24.
- Duffy G, Cheng PP, Yuan N *et al*. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA Cardiol* 2022;**7**:386–95.
- Petersen SE, Matthews PM, Bamberg F *et al*. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank-rationale, challenges and approaches. *J Cardiovasc Magn Reson* 2013;**15**:46.
- Ferencik M, Mayrhofer T, Bittner DO *et al*. Use of high-risk coronary atherosclerotic plaque detection for risk stratification of patients with stable chest pain: a secondary analysis of the PROMISE randomized clinical trial. *JAMA Cardiol* 2018;**3**:144–52.
- Cetin I, Raisi-Estabragh Z, Petersen SE *et al*. Radiomics signatures of cardiovascular risk factors in cardiac MRI: results from the UK biobank. *Front Cardiovasc Med* 2020;**7**:591368.
- Kuruwilla S, Adenaw N, Katwal A *et al*. Late gadolinium enhancement on cardiac magnetic resonance predicts adverse cardiovascular outcomes in nonischemic cardiomyopathy: a systematic review and meta-analysis. *Circ Cardiovasc Imaging* 2014;**7**:250–8.
- Kalam K, Otahal P, Marwick TH. Prognostic implications of global LV dysfunction: a systematic review and meta-analysis of global longitudinal strain and ejection fraction. *Heart* 2014;**100**:1673–80.
- Williams MC, Kwiecinski J, Doris M *et al*. Low-attenuation non-calcified plaque on coronary computed tomography angiography predicts myocardial infarction: results from the multicenter SCOT-HEART trial. *Circulation* 2020;**141**:1452–62.
- Otsuka K, Fukuda S, Tanaka A *et al*. Napkin-ring sign on coronary CT angiography for the prediction of acute coronary syndrome. *JACC Cardiovasc Imaging* 2013;**6**:448–57.
- Oikonomou EK, Marwan M, Desai MY *et al*. Non-invasive detection of coronary inflammation using computed tomography and prediction of residual cardiovascular risk (the CRISP CT study): a post-hoc analysis of prospective outcome data. *Lancet* 2018;**392**:929–39.
- Koenig AL, Shchukina I, Amrute J *et al*. Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nat Cardiovasc Res* 2022;**1**:263–80.
- Ganz P, Heidecker B, Hveem K *et al*. Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA* 2016;**315**:2532–41.
- Wang L, Peng Y, Zhou B *et al*. Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. *Nat Cell Biol* 2020;**22**:108–19.
- Lee SE, Joo JH, Hwang HS *et al*. Spatial transcriptional landscape of human heart failure. *Eur Heart J* 2025;**46**:3098–114.
- Campos J, McMurray JL, Certo M *et al*. Spatial transcriptomics elucidates localized immune responses in atherosclerotic coronary artery. *EMBO Mol Med* 2025;**17**:2827–46.
- Kleshcheynikov V, Shmatko A, Dann E *et al*. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* 2022;**40**:661–71.
- Bild DE, Bluemke DA, Burke GL *et al*. Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol* 2002;**156**:871–81.
- Kramer CM, Appelbaum E, Desai MY *et al*. Hypertrophic cardiomyopathy registry: the rationale and design of an international, observational study of hypertrophic cardiomyopathy. *Am Heart J* 2015;**170**:223–30.

33. Gow B, Pollard T, Nathanson LA *et al*. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset (version 1.0). *PhysioNet* 2023. <https://doi.org/10.1111/10.13026/4nqg-sb35>.
34. Fonseca CG, Backhaus U, Bluemke DA *et al*. The cardiac atlas project—an imaging database for cardiovascular phenotypes. *Bioinformatics* 2011;**27**:2288–95.
35. Sun J, Singh P, Shami A *et al*. Spatial transcriptional mapping reveals site-specific pathways underlying human atherosclerotic plaque rupture. *J Am Coll Cardiol* 2023;**81**:2213–27.
36. Gastanadui MG, Margaroli C, Litovsky S *et al*. Spatial transcriptomic approach to understanding coronary atherosclerotic plaque stability. *Arterioscler Thromb Vasc Biol* 2024;**44**:e264–76.
37. Bai W, Sinclair M, Tarroni G *et al*. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson* 2018;**20**:65.
38. Bai W, Suzuki H, Huang J *et al*. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat Med* 2020;**26**:1654–62.
39. Oikonomou EK, Williams MC, Kotanidis CP *et al*. A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography. *Eur Heart J* 2019;**40**:3529–43.
40. Dosovitskiy A, Beyer L, Kolesnikov A *et al*. An image is worth 16 × 16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations*. Virtual conference: OpenReview.net; 2021.
41. He K, Chen X, Xie S *et al*. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans (LA), USA: IEEE; 2022. p. 15979–88.
42. Jacob AJ, Borgohain I, Chitiboi T *et al*. Towards a cardiovascular magnetic resonance foundation model for multi-task cardiac image analysis. *J Cardiovasc Magn Reson* 2025;**27**:101967.
43. Christensen M, Vukadinovic M, Yuan N *et al*. Vision–language foundation model for echocardiogram interpretation. *Nat Med* 2024;**30**:1481–8.
44. Chen W, Zhang P, Tran TN *et al*. A visual–omics foundation model to bridge histopathology with spatial transcriptomics. *Nat Methods* 2025;**22**:1568–82.
45. Cisternino F, Ometto S, Chatterjee S *et al*. Self-supervised learning for characterising histomorphological diversity and spatial RNA expression prediction across 23 human tissue types. *Nat Commun* 2024;**15**:5906.
46. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.
47. Hao Y, Stuart T, Kowalski MH *et al*. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;**42**:293–304.
48. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15.
49. Eraslan G, Simon LM, Mircea M *et al*. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**:390.
50. Lopez R, Regier J, Cole MB *et al*. Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8.
51. Gayoso A, Lopez R, Xing G *et al*. A Python library for probabilistic analysis of single-cell omics data. *Nat Biotechnol* 2022;**40**:163–6.
52. Argelaguet R, Velten B, Arnol D *et al*. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;**14**:e8124.
53. Hao M, Gong J, Zeng X *et al*. Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481–91.
54. Tejada-Lapueta A, Schaar AC, Gutgesell R *et al*. Nicheformer: a foundation model for single-cell and spatial omics. *Nat Methods* 2025;**22**:2525–38.
55. Hu J, Li X, Coleman K *et al*. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* 2021;**18**:1342–51.
56. Zhao E, Stone MR, Ren X *et al*. Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol* 2021;**39**:1375–84.
57. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* 2022;**13**:1739.
58. Long Y, Ang KS, Li M *et al*. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat Commun* 2023;**14**:1155.
59. Zhou L, Peng X, Chen M *et al*. Unveiling patterns in spatial transcriptomics data: a novel approach utilizing graph attention autoencoder and multiscale deep subspace clustering network. *GigaScience* 2025;**14**:giae103.
60. Min W, Shi Z, Zhang J *et al*. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Brief Bioinform* 2024;**25**:bbae551.
61. Wang C, Chan AS, Fu X *et al*. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nat Commun* 2025;**16**:1544.
62. Madhu H, Rocha JF, Huang T *et al*. HEIST: a graph foundation model for spatial transcriptomics and proteomics data. In: *The Fourteenth International Conference on Learning Representations*. Rio de Janeiro, Brazil: OpenReview.net; 2026.
63. Xu C, Jin X, Wei S *et al*. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res* 2022;**50**:e131.
64. Ferreira DL, Lau C, Salaymang Z *et al*. Self-supervised learning for label-free segmentation in cardiac ultrasound. *Nat Commun* 2025;**16**:4070.
65. Cui H, Wang C, Maan H *et al*. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470–80.
66. Radhakrishnan A, Friedman SF, Khurshid S *et al*. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nat Commun* 2023;**14**:2436.
67. Radford A, Kim JW, Hallacy C *et al*. Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning*. Virtual conference: Proceedings of Machine Learning Research; 2021. 139:8748–63.
68. Zhang Y, Jiang H, Miura Y *et al*. Contrastive learning of medical visual representations from paired images and text. In: *Proceedings of the 7th Machine Learning for Healthcare Conference*. Durham (NC), USA: Proceedings of Machine Learning Research; 2022. 182:2–25.
69. Akinci D'Antonoli T, Bluethgen C, Cuocolo R *et al*. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. *Diagn Interv Radiol* 2026;**32**:259–72.
70. Sun K, Xue S, Sun F *et al*. Medical multimodal foundation models in clinical diagnosis and treatment: applications, challenges, and future directions. *Artif Intell Med* 2025;**170**:103265.
71. Strotzer QD, Nieberle F, Kupke LS *et al*. Toward foundation models in radiology? Quantitative assessment of GPT-4V's multimodal and multianatomic region capabilities. *Radiology* 2024;**313**:e240955.

72. Abbasi S, Lan H, Choupan J *et al.* Deep learning for the harmonization of structural MRI scans: a survey. *Biomed Eng Online* 2024;**23**:90.
73. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 2007;**8**:118–27.
74. Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96.
75. Hassani C, Saremi F, Varghese BA *et al.* Myocardial radiomics in cardiac MRI. *Am J Roentgenol* 2020;**214**:536–45.
76. Bhattacharya A, Sadasivuni S, Chao C-J *et al.* Multi-modal fusion model for predicting adverse cardiovascular outcome post percutaneous coronary intervention. *Physiol Meas* 2022;**43**:124004.
77. Baltrušaitis T, Ahuja C, Morency L-P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell* 2019;**41**:423–43.
78. Andrew G, Arora R, Bilmes J *et al.* Deep canonical correlation analysis. In: *Proceedings of the 30th International Conference on Machine Learning*. Atlanta (GA), USA: Proceedings of Machine Learning Research; 2013. 28:1247–55.
79. Wang W, Arora R, Livescu K *et al.* On deep multi-view representation learning. In: *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France: Proceedings of Machine Learning Research; 2015. 37:1083–92.
80. Wu M, Goodman N. Multimodal generative models for scalable weakly-supervised learning. In: *Advances in Neural Information Processing Systems (NeurIPS)* 2018;**31**:5580–90.
81. Joshi A, Gupta N, Shah J *et al.* Generalized product-of-experts for learning multimodal representations in noisy environments. In: *Proceedings of the 2022 International Conference on Multimodal Interaction*. Bengaluru, India: Association for Computing Machinery; 2022. p. 83–93.
82. Hemker K, Simidjievski N, Jamnik M. HEALNet: multimodal fusion for heterogeneous biomedical data. *Adv Neural Inf Process Syst* 2024;**37**:64479–98.
83. Tonekaboni S, Friedman SF, Zhang X *et al.* A representation fusion framework for decoupling diagnostic information in multimodal learning. *NPJ Digit Med* 2025;**8**:765.
84. Zheng Y, Conrad RD, Green EJ *et al.* Graph attention-based fusion of pathology images and gene expression for prediction of cancer survival. *IEEE Trans Med Imaging* 2024;**43**:3085–97.
85. Li S, Hua H, Chen S. Graph neural networks for single-cell omics data: a review of approaches and applications. *Brief Bioinform* 2025;**26**:bbaf109.
86. Parisot S, Ktena SI, Ferrante E *et al.* Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med Image Anal* 2018;**48**:117–30.
87. Chen RJ, Lu MY, Williamson DFK *et al.* Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* 2022;**40**:865–78.
88. Howard FM, Kather JN, Pearson AT. Multimodal deep learning: an improvement in prognostication or a reflection of batch effect? *Cancer Cell* 2023;**41**:5–6.
89. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**:565–74.
90. Collins GS, Dhiman P, Andaur CL *et al.* TRIPOD+AI: updated guidance for reporting clinical prediction models that use machine learning. *BMJ* 2024;**385**:e078378.
91. Moons KGM, Damen JAA, Kaul T *et al.* PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* 2025;**388**:e082505.
92. Argelaguet R, Arnol D, Bredikhin D *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:111.
93. Palla G, Spitzer H, Klein M *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat Methods* 2022;**19**:171–8.
94. Dries R, Zhu Q, Dong R *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol* 2021;**22**:78.
95. The MONAI Consortium. MONAI: Medical Open Network for AI. *arXiv preprint arXiv:2211.02701*. Ithaca (NY): Cornell University; 2022.
96. Pérez-García F, Sparks R, Ourselin S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput Methods Prog Biomed* 2021;**208**:106236.
97. He W, Huang W, Zhang L *et al.* Radiogenomics: bridging the gap between imaging and genomics for precision oncology. *MedComm* 2024;**5**:e722.
98. Guo Y, Li T, Gong B *et al.* From images to genes: radiogenomics based on artificial intelligence to achieve non-invasive precision medicine in cancer patients. *Adv Sci* 2025;**12**:2408069.
99. Lin M, Guo J, Gu Z *et al.* Machine learning and multi-omics integration: advancing cardiovascular translational research and clinical practice. *J Transl Med* 2025;**23**:388.
100. Infante T, Cavaliere C, Punzo B *et al.* Radiogenomics and artificial intelligence approaches applied to cardiac computed tomography angiography and cardiac magnetic resonance for precision medicine in coronary heart disease: a systematic review. *Circ Cardiovasc Imaging* 2021;**14**:1133–46.
101. Ho CY, Day SM, Ashley EA *et al.* Genotype and lifetime burden of disease in hypertrophic cardiomyopathy: insights from the sarcomeric human cardiomyopathy registry (SHaRe). *Circulation* 2018;**138**:1387–98.
102. Wehrens M, de Leeuw AE, Wright-Clark M *et al.* Single-cell transcriptomics provides insights into hypertrophic cardiomyopathy. *Cell Rep* 2022;**39**:110809.
103. Ahlberg G, Andreassen L, Ghouse J *et al.* Genome-wide association study identifies 18 novel loci associated with left atrial volume and function. *Eur Heart J* 2021;**42**:4523–34.
104. Luo Y, Zhang N, Yang J *et al.* AI-based multiomics profiling reveals complementary omics contributions to personalized prediction of cardiovascular disease. *Nat Commun* 2026;**17**:2269.
105. Khan M, Arslanturk S, Draghici S. A comprehensive review of spatial transcriptomics data alignment and integration. *Nucleic Acids Res* 2025;**53**:gkaf536.
106. Ekval M, Bergenstråhle L, Andersson A *et al.* Spatial landmark detection and tissue registration with deep learning. *Nat Methods* 2024;**21**:673–9.
107. Chelebian E, Avenel C, Wählby C. Combining spatial transcriptomics with tissue morphology. *Nat Commun* 2025;**16**:4452.
108. Fu X, Cao Y, Bian B *et al.* Spatial gene expression at single-cell resolution from histology using deep learning with GHIST. *Nat Methods* 2025;**22**:1900–10.
109. Raghunath S, Ulloa AE, Cerna LJ *et al.* Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med* 2020;**26**:886–91.
110. Attia ZI, Noseworthy PA, Lopez-Jimenez F *et al.* An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm:

- a retrospective analysis of outcome prediction. *Lancet* 2019; **394**:861–7.
111. Sau A, Pastika L, Sieliwonczyk E *et al*. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study. *Lancet Digit Health* 2024; **6**:e791–802.
 112. Vukadinovic M, Chiu I-M, Tang X *et al*. Comprehensive echocardiogram evaluation with view primed vision language AI. *Nature* 2026; **650**:970–7.
 113. Forrest IS, Petrazzini BO, Duffy Á *et al*. Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. *Lancet* 2023; **401**:215–25.
 114. Lai C, Yin M, Kholmovski EG *et al*. Multimodal AI to forecast arrhythmic death in hypertrophic cardiomyopathy. *Nat Cardiovasc Res* 2025; **4**:891–903.
 115. Miao BY, Chen IY, Williams CYK *et al*. The MI-CLAIM-GEN checklist for generative artificial intelligence in health. *Nat Med* 2025; **31**:1394–8.
 116. Acosta JN, Falcone GJ, Rajpurkar P *et al*. Multimodal biomedical AI. *Nat Med* 2022; **28**:1773–84.
 117. Quer G, Topol EJ. The potential for large language models to transform cardiovascular medicine. *Lancet Digit Health* 2024; **6**:e767–71.
 118. Sarraju A, Ouyang D, Itchhaporia D. The opportunities and challenges of large language models in cardiology. *JACC: Advances* 2023; **2**:100438.
 119. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union* 2024; L 2024/1689. Luxembourg: Publications Office of the European Union; 2024.
 120. Norouzi R, Norouzi R, Abbasi K *et al*. DFT_ANPD: a dual-feature two-sided attention network for anticancer natural products detection. *Comput Biol Med* 2025; **194**:110442.
 121. Abbasi K, Razzaghi P. Incorporating part-whole hierarchies into fully convolutional network for scene parsing. *Expert Syst Appl* 2020; **160**:113662.
 122. Abbasi K, Razzaghi P, Gharizadeh A *et al*. Computational drug design in the artificial intelligence era: a systematic review of molecular representations, generative architectures, and performance assessment. *Pharmacol Rev* 2026; **78**:100095.
 123. Corral-Acero J, Margara F, Marciniak M *et al*. The ‘digital twin’ to enable the vision of precision cardiology. *Eur Heart J* 2020; **41**:4556–64.
 124. Coorey GM, Figtree GA, Fletcher DF *et al*. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. *NPJ Digit Med* 2022; **5**:126.
 125. Prakosa A, Arevalo HJ, Deng D *et al*. Personalized virtual-heart technology for guiding the ablation of infarct-related ventricular tachycardia. *Nat Biomed Eng* 2018; **2**:732–40.
 126. Sel K, Osman D, Zare F *et al*. Building digital twins for cardiovascular health: from principles to clinical impact. *J Am Heart Assoc* 2024; **13**:e031981.
 127. Akbarialiabad H, Pasdar A, Murrell DF *et al*. Enhancing randomized clinical trials with digital twins. *NPJ Syst Biol Appl* 2025; **11**:110.
 128. Arevalo HJ, Vadakkumpadan F, Guallar E *et al*. Arrhythmia risk stratification of patients after myocardial infarction using personalized heart models. *Nat Commun* 2016; **7**:11437.