CPAgents: Agentic Composite Phenotype Generation for Cardiac Disease Association

Zuoou Li[1*], Wenlong Zhao[2*], Kelly Yu[1*], Weitong Zhang[3], Paul M. Matthews[4,7,8], Wenjia Bai[3,4,5], Bernhard Kainz[3,6], Mengyun Qiao[1†]

[1] Department of Mechanical Engineering, University College London, London, UK
[2] CSIG Group, Tencent, Beijing, China
[3] Department of Computing, Imperial College London, London, UK
[4] Department of Brain Sciences, Imperial College London, London, UK
[5] Data Science Institute, Imperial College London, London, UK
[6] FAU Erlangen–Nürnberg, Erlangen, DE
[7] UK Dementia Research Institute, Imperial College London, London, UK
[8] Rosalind Franklin Institute, Harwell Science and Innovation Campus, Didcot, UK
lizoleo9344@gmail.com, m.qiao@ucl.ac.uk

Abstract. Identifying robust associations between cardiac imaging phenotypes and clinical diseases is fundamental to population-scale cardiovascular research and reliable risk stratification. However, current phenome-wide association studies rely on pre-defined, single-variable phenotypes or expert-crafted features, which limits their ability to capture clinically meaningful non-linear effects and cross-phenotype interactions. To address this, we propose `CPAgents`, an iterative phenotype-**C**omposition framework for cardiovascular **P**henome-wide association study (PheWAS) that automatically constructs and validates interpretable composite phenotypes (e.g., polynomial, ratio, and interaction forms) from base imaging features. Specifically, our system coordinates three agents: (i) an *Analyst* that identifies statistical pathologies and nominates candidate transformations; (ii) a *Proposer* that generates constrained, medically and statistically motivated expressions under numerical safety rules; and (iii) a *Verifier* that evaluates candidates using multi-stage criteria and produces transparent evidence trails for accepted phenotypes. Evaluated on a population-scale cardiac imaging cohort, the discovered composite phenotypes markedly improve disease discrimination: across 72 classifier–disease–metric combinations, our variants achieve the top rank in 56 cases versus 18 for baselines, with gains observed across all nine clinical disease categories. Our framework yields compact, clinically interpretable phenotype formulas with transparent evidence trails, enabling scalable discovery of stronger phenotype-disease associations beyond expert-driven feature selection.

Keywords: Interpretable Composite Phenotypes, Agentic Phenotype Composition, Phenome-wide Association Study

[*] Equal contribution.
[†] Corresponding author.

1 Introduction

Identifying robust associations between cardiac imaging phenotypes and clinical diseases is central to understanding disease mechanisms and building reliable risk stratification models [1]. Modern cardiovascular imaging cohorts, such as large-scale CMR and echocardiography biobanks, provide rich quantitative phenotypes that describe chamber size and function, ventricular mass and strain, aortic geometry [2]. These imaging-derived traits serve as intermediate phenotypes linking genetic and environmental factors to clinical outcomes, and are increasingly used in mechanistic studies, biomarker discovery, and treatment response assessment [3].

The emergence of population-based cohorts with standardised CMR and echocardiography protocols, coupled with automated image analysis pipelines, has produced large imaging datasets at unprecedented scale [4]. While this enables systematic mapping of the cardiac imaging phenome to diverse clinical conditions and risk factors [5], it also introduces substantial analytic challenges. In practice, the sheer number, redundancy, and complex dependency structure of imaging-derived traits create a substantial analytic bottleneck: it is non-trivial to decide which combinations of phenotypes to test, how to capture potentially non-linear relationships between imaging traits and clinical outputs, and how to ensure that resulting associations are robust and clinically interpretable [6].

Previous work has largely adopted PheWAS-style pipelines [7], in which domain experts predefine a fixed set of single-variable imaging phenotypes and evaluate their associations with a range of diseases under carefully specified statistical models [8, 9] or machine learning algorithms [10, 11]. Subsequent work constructs composite indices based on prior clinical knowledge, for example, by combining related volumetric measures or summarizing regional features into global scores [12, 13]. While these approaches are transparent and clinically grounded, they do not scale well without substantial manual effort, and are constrained by expert-defined features and linear modeling assumptions.

More recent studies have explored automated feature construction and data-driven discovery pipelines, including symbolic composition and agent-assisted data science systems [14–16]. Although such methods broaden the candidate feature space, they are typically developed for general predictive modelling rather than large-scale phenotype discovery with explicit confounder control and numerical safeguards. Greater flexibility can enhance expressiveness, but it may also introduce instability when applied to correlated imaging traits.

Both expert-driven and automated approaches face two major limitations in high-dimensional, population-scale imaging data. First, predefined feature sets restrict the ability to capture clinically meaningful non-linear effects and interactions across phenotypes. Second, expanding the feature space through adaptive composition inflates the multiple-testing burden and increases vulnerability to unstable findings under cohort heterogeneity and residual confounding.

To address these limitations, we propose `CPAgents`, a **C**omposite-**P**henotype discovery framework for disease-wide cardiovascular phenotype association analysis that automatically constructs and validates interpretable composite imaging

phenotypes capable of capturing non-linear effects and cross-phenotype interactions. The main contributions are: (i) **Agentic iterative phenotype discovery.** We formulate phenotype discovery as an iterative analyse-propose-verify loop coordinated by specialised agents, enabling automatic construction of composite phenotypes within a constrained operator set. (ii) **Statistics-informed composition and validation.** We first summarise key statistical structure in the data (e.g., distributions, redundancy, and correlation patterns), and then perform principled screening and verification using cross-validated metrics. Candidate composites are retained only when improvements are consistent across folds and models and supported by statistical tests, mitigating selection-induced overfitting. (iii) **Interpretable and reproducible outputs.** We represent each discovered composite phenotype as a compact, closed-form formula with a clear construction trace and cross-validated evidence summaries, enabling transparent auditing and facilitating reproducible downstream association analyses and external replication. [1]

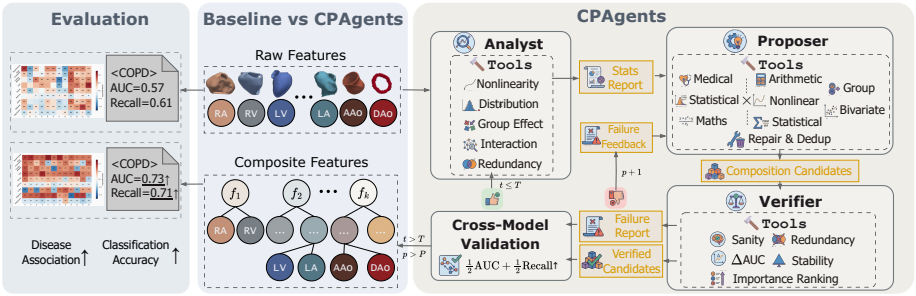

Fig. 1: **Overview of `CPAgents`, an agentic phenotype composition framework.** `CPAgents` iteratively transforms raw cardiovascular phenotypes into highly predictive, hierarchical composite features ($f_1 ... f_k$). An *Analyst* first profiles feature statistics to guide a *Proposer*, which synthesises candidate features using medical, statistical, and exploratory operations. Next, a *Verifier* filters candidates via sanity, stability, and ΔAUC checks. Candidates then undergo cross-model validation; failures trigger a feedback loop for refinement, while successes are retained. Ultimately, these transparent composite features significantly enhance downstream disease association mapping and classification accuracy.

2 Method

Problem Setup. We consider a supervised *composite phenotype* discovery problem for disease-wide cardiovascular imaging phenotype association analysis across multiple target diseases. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote a cohort of N subjects.

[1] Code is available at: https://github.com/LumaLabAI/CPAgents.

Each subject is represented by a comprehensive feature vector $\mathbf{x}_i \in \mathbb{R}^d$ (where $d = p+q$), comprising p base cardiac imaging phenotypes (e.g., atrial/ventricular measures for `RA`/`RV`/`LA`/`LV` and aortic measures for `AAo`/`DAo`) and q potential confounders (e.g., age, sex, alcohol intake, blood pressure). The target $y_i \in \{1, \ldots, K\}$ is a K-class disease label.

Our goal is to automatically discover and validate a compact set of interpretable composite phenotypes $\mathcal{F} = \{f_k\}_{k=1}^{M}$. Each f_k is a closed-form mapping $f_k : \mathbb{R}^d \to \mathbb{R}$ operating on $\mathbf{x}$, such that $f_k \in \mathcal{G}$, where $\mathcal{G}$ denotes the space of valid expressions generated by a predefined operator set (e.g., clinical, statistical, and arithmetic transformations). These derived phenotypes aim to improve the predictive performance of downstream disease classification.

Composite Phenotype Space. While standard linear or shallow models often under-parameterize non-linear effects, higher-order interactions, and group-dependent distributions, composite phenotypes make these underlying structures explicit in a compact, one-dimensional representation.

To formalize the search space $\mathcal{G}$, we define the generation of a composite phenotype as the construction of an Abstract Syntax Tree (AST), denoted as $\mathcal{T}$. Let $\mathcal{V} = \{x^{(1)}, x^{(2)}, \ldots, x^{(d)}\}$ denote the set of all d available scalar features from the input space. In this representation: (i) Leaf Nodes ($\mathcal{L}$) correspond either to a raw scalar feature $x^{(j)} \in \mathcal{V}$ or to domain-specific numerical constants $c \in \mathbb{R}$ organically generated by the Large Language Model (LLM). (ii) Internal Nodes ($\mathcal{N}$) represent allowable operators $o \in \mathcal{O}$ applied to their child nodes.

To guarantee the validity, executability, and safety of the generated expressions on cross-sectional tabular data, the operator space $\mathcal{O}$ is strictly constrained to a predefined set of functions. These operators can be broadly categorised into: (i) Element-wise Non-linearities & Transformations: Mathematical functions (*e.g.*, `log1p`, `sqrt`, `clip`, `power`, `where`) designed to handle skewed distributions and outlier masking. (ii) Arithmetic Interactions: Basic algebraic operations (*e.g.*, `add`, `sub`, `mul`, `div`) to construct interaction terms, ratios, and differences. (iii) Group-aware Reductions & Statistics: Advanced functions (*e.g.*, `groupby`, `transform(mean/std)`, `rank`) that allow features to be normalised or evaluated relative to specific demographic or clinical sub-populations (*e.g.*, stratifying by `Sex`). Numerical safety guardrails, such as adding a smoothing term ($\epsilon = 10^{-9}$) to the denominator of all division operators (`div`), are strictly enforced at the AST parsing level to prevent zero-division.

Clinical Example. Coupled with the LLM's medical priors, this AST space enables the zero-shot discovery of complex, actionable indices. For example, the system can autonomously formulate demographic-adjusted phenotypes like:

$$f_{\text{new}} = \frac{\text{LVM}_{\text{index}} - \mu(\text{LVM}_{\text{index}} \mid \text{Sex})}{\sigma(\text{LVM}_{\text{index}} \mid \text{Sex}) + \epsilon}, \text{LVM}_{\text{index}} = \frac{\text{LVM}}{(\text{Height}/100)^{2.7} + \epsilon} \quad (1)$$

This AST represents the *sex-specific Z-score of allometrically indexed Left Ventricular Mass*, where μ and σ denote group-conditional statistics. Crucially, domain-specific constants like 2.7 (a cardiologic allometric scaling factor) and 100 are elicited directly from the LLM's internal knowledge, bypassing compu-

tationally expensive random search. Such explicit, domain-aware features significantly enhance the performance of interpretable downstream classifiers.

Iterative Search Procedure. We formulate automated feature engineering as an iterative search over a constrained composite-phenotype space, orchestrated by a closed-loop system of three specialised agents: an *Analyst*, a *Proposer*, and a *Verifier*. At each iteration (t), the system seeks to augment the current feature set $\mathcal{F}^{(t)}$ through a structured four-phase procedure: *statistical profiling*, *transformation generation*, *multi-stage filtering*, and *global acceptance testing*.

Phase 1: Statistical Profiling. To prevent blind combinatorial search, the Analyst Agent systematically profiles $\mathcal{F}^{(t)}$ to uncover data characteristics. This yields five comprehensive statistical reports: (i) Univariate shape assessment via ΔAIC between linear and spline logistic models to detect non-linear and monotonicity benefits; (ii) Distributional characterisation through skewness, kurtosis, and outlier rates; (iii) Pairwise interaction screening using mutual information, followed by cross-validated logistic confirmation; (iv) Group-effect analysis via intraclass correlation (ICC) and Cramér's V to assess safe group-level standardisations; and (v) Redundancy clustering on the absolute correlation matrix. Subsequently, a Large Language Model (LLM) compresses these statistical diagnostics into structured, per-feature recommendations, prioritising interaction pairs and safely bounded group operations.

Phase 2: Transformation Generation. Guided by the structured hints from the *Analyst*, the *Proposer* generates $C = \min(|\mathcal{F}^{(t)}|, C_{\max})$ new composite feature candidates $f_{\text{cand}} \in \mathcal{G}$. To maintain a balance between domain validity and empirical discovery, these candidates are produced via three separate LLM calls, corresponding to medical, statistical, and exploratory priorities allocated in a 2:2:1 ratio: (i) Medical Priors (40%): Features derived purely from the LLM's internal clinical knowledge (*e.g.*, specific risk indices or established physiological ratios), irrespective of statistical hints. (ii) Statistical Transformations (40%): Features directly addressing the *Analyst*'s reports, such as applying `log1p` to heavy-tailed distributions or creating interaction terms with a high ΔAIC. (iii) Exploratory Mathematics (20%): Safe, automated mathematical combinations within the operator space $\mathcal{O}$, designed to discover latent, unintuitive patterns.

Each LLM call is conditioned on the current feature set $\mathcal{F}^{(t)}$, forbidden target labels y, safe group operations, and the *Analyst*'s per-feature recommendations, interaction pairs, and redundancy clusters. All generated expressions are constrained to an interpreter that performs AST-level validation, restricting operations to the predefined operator space $\mathcal{O}$. Failed candidates are re-prompted to the LLM with error messages for up to two repair attempts.

Phase 3: Multi-Stage Filtering. To distill the top-ranked proposals, the *Verifier* evaluates the newly proposed and existing features through a multi-stage screening: (i) Marginal Utility Check: Evaluating the isolated ΔAUC of each new feature. (ii) Stability Selection: Applying ElasticNet [17] regularisation to penalize spurious correlations and verify stability. (iii) Deduplication: Removing redundant features via a combination of statistical correlation clustering and AST-based expression textual matching. (iv) Global Importance Assess-

ment: Training a LightGBM model and utilizing SHapley Additive exPlanations (SHAP) [18] values to extract the absolute top-ranked impactful features.

Phase 4: Global Acceptance Criteria. Finally, the top-ranked candidate features, denoted as $\mathcal{F}_{\text{cand}}$, undergo empirical validation. To assess their downstream predictive utility, we define a composite evaluation metric $\mathcal{S} = 0.5 \cdot \text{AUC} + 0.5 \cdot \text{Recall}$. The augmented feature space, $\mathcal{F}^{(t)} \cup \mathcal{F}_{\text{cand}}$, is evaluated on a hold-out validation dataset using a panel of four base classifiers (SVM, LDA, AdaBoost, and MLP). The feature update is governed by a ensemble agreement threshold: if $\mathcal{S}(\mathcal{F}^{(t)} \cup \mathcal{F}_{\text{cand}})$ improves over the previous $\mathcal{S}(\mathcal{F}^{(t)})$ on at least two out of the four models, the new features are accepted if $t <$ max iteration T, yielding $\mathcal{F}^{(t+1)} = \mathcal{F}^{(t)} \cup \mathcal{F}_{\text{cand}}$. Otherwise, the candidates are discarded ($\mathcal{F}^{(t+1)} = \mathcal{F}^{(t)}$), and the *Verifier* passes the performance feedback back to the *Proposer* to refine its output for the next iteration. To bound the computational cost, we uses early stopping, terminating after $P = 3$ consecutive rejections.

3 Experiments and Results

Dataset and Settings. We evaluate the proposed framework on the UK Biobank dataset[9], a population-scale cardiac imaging cohort comprising 26,893 participants. We consider $K = 9$ clinical disease categories, including hypertension, high cholesterol, cardiac disease, PVD, stroke, asthma, COPD, diabetes, and depression. Cardiac imaging phenotypes are extracted using a standardised CMR analysis pipeline [8] and released by the UK Biobank[10]. The following confounders are included in all analyses: sex, age, weight, height, alcohol intake, alcohol intake log10, systolic blood pressure, and diastolic blood pressure.

Evaluation Metrics. We divide the dataset into training, validation, and test sets with a ratio of 6:2:2. Composite phenotype discovery, model selection, and threshold tuning are performed exclusively on the training and validation sets to prevent information leakage. Final evaluation is conducted once on the independent test set. We report the Area Under the ROC Curve (AUC) and Recall as primary performance metrics. Additionally, to interpret why the composite features generated by our method are beneficial for classification, we utilize the Silhouette score. For a given sample i, it is defined as $s(i) = (b(i) - a(i))/\max\{a(i), b(i)\}$, where $a(i)$ represents the mean intra-class distance to all other samples within the same class, and $b(i)$ is the mean inter-class distance to the samples of the nearest neighboring class. The overall score is the average of $s(i)$ across all samples, ranging from -1 to 1.

Implementation Details. The *Analyst* and *Proposer* agents use DeepSeek-V3.2 [19] as the LLM backend via LangChain with the default temperature for a balance between deterministic reasoning and exploration. During the ΔAIC profiling phase, the *Analyst* models non-linear effects using B-splines (degrees of freedom $df = 4$, polynomial degree $d = 3$) with five-fold cross-validation. Clinical definitions of base features (*e.g.*, `LVEDV`: Left Ventricular End-Diastolic

[9] http://www.ukbiobank.ac.uk/register-apply
[10] https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=157

Volume) are injected into the *Proposer* system prompts to ground candidate generation. The *Verifier* performs stability selection with 50 bootstrap runs of an ElasticNet classifier (ℓ_1-ratio $= 0.5$, inverse regularization $C = 0.5$). To prevent data leakage, all feature engineering and selection are restricted to the training set, and final performance is reported on held-out validation and test sets. Each search uses a fixed random seed for reproducibility, and all metrics are averaged over five independent runs with different seeds to ensure robustness.

Baselines. To demonstrate the benefits of our composite features, we employ several baselines: expert-defined features [8], features selected by MESHAgent [12] through debate and consensus. We employ ChatGPT and DeepSeek, two cutting-edge LLMs accessed via the official APIs. All experiments are performed on a single NVIDIA RTX 6000 GPU.

Results. (i) Disease Association Study. The disease–phenotype association heatmap (Figure 2) shows that composite phenotypes discovered by our framework yield clearer, more concentrated association patterns compared to base and expert-defined traits, with stronger effects in key cardiac and other disease groups. (ii) *Diagnostic performance.* Table 1 compares disease discrimination using our composite phenotypes with expert-selected single features [8] and MESHAgents-based features [12] across four classifiers and nine disease endpoints. A ranking analysis over all 72 classifier–disease–metric combinations shows that our two composite-phenotype variants occupy the top rank in **56** cases versus **18** for the baselines, and appear in the top two in **51** settings (compared with 33 and 31 for expert-defined and MESHAgents features), indicating consistently competitive and superior performance across diseases rather than gains concentrated in a few favourable tasks.

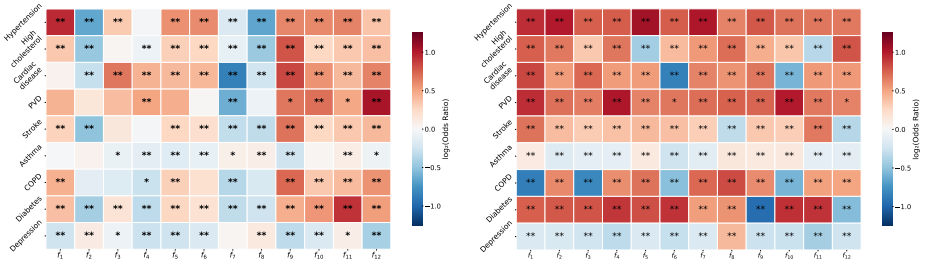

Fig. 2: Disease–phenotype association heatmaps for expert-defined features (left) and our composite phenotypes (right). Except for asthma and depression, composite phenotypes produce more concentrated and consistent association patterns across diseases, supporting improved cluster-level interpretability of the learned features.

Robustness and Ablation Studies. As shown in Table 2, the full framework achieves the highest overall performance, yielding an average AUC of

Table 1: Performance comparison across models and diseases. Best/second-best results are shown in **bold**/underlined. For AUC, $^{\dagger}/^{\ddagger}$ indicate overall significant gains over Experts/MESHAgents by paired Wilcoxon tests ($p < 0.05$).

Methods	Hypertension		High cholesterol		Cardiac disease		PVD		Stroke		Asthma		COPD		Diabetes		Depression	
	AUC ↑	Recall ↑	AUC ↑	Recall ↑	AUC ↑	Recall ↑	AUC ↑	Recall ↑	AUC ↑	Recall ↑	AUC ↑	Recall↑	AUC ↑	Recall ↑	AUC ↑	Recall ↑	AUC ↑	Recall ↑
SVM																		
Experts [8]	0.755 ±0.007	0.779 ±0.009	0.663 ±0.014	0.683 ±0.026	0.737 ±0.010	0.620 ±0.033	0.710 ±0.075	**0.783** ±0.183	0.635 ±0.040	0.615 ±0.070	0.566 ±0.011	0.539 ±0.031	0.720 ±0.052	0.686 ±0.096	0.781 ±0.008	0.715 ±0.028	0.601 ±0.007	0.566 ±0.045
MESHAgents [12]	0.748 ±0.009	0.771 ±0.013	0.664 ±0.014	**0.689** ±0.025	0.734 ±0.009	0.622 ±0.027	0.703 ±0.049	0.710 ±0.133	0.623 ±0.047	0.549 ±0.060	0.565 ±0.009	0.550 ±0.023	0.712 ±0.063	**0.692** ±0.110	0.771 ±0.008	0.717 ±0.022	0.603 ±0.009	**0.586** ±0.016
Ours†‡ GPT-5.1	0.757 ±0.006	0.775 ±0.009	**0.667** ±0.014	0.679 ±0.018	**0.752** ±0.009	**0.646** ±0.030	**0.713** ±0.079	0.603 ±0.199	0.641 ±0.035	0.609 ±0.033	**0.582** ±0.012	**0.567** ±0.021	**0.726** ±0.040	0.654 ±0.100	**0.785** ±0.004	**0.736** ±0.015	0.607 ±0.013	0.564 ±0.016
Ours†‡ DeepSeek V3.2	**0.759** ±0.006	**0.780** ±0.006	0.663 ±0.012	0.687 ±0.039	0.746 ±0.011	0.634 ±0.057	0.712 ±0.088	0.553 ±0.183	**0.652** ±0.040	**0.632** ±0.078	0.580 ±0.009	0.565 ±0.015	0.706 ±0.058	0.641 ±0.100	0.784 ±0.005	0.725 ±0.016	**0.609** ±0.014	0.560 ±0.023
AdaBoost																		
Experts [8]	0.741 ±0.010	0.699 ±0.018	0.667 ±0.006	0.655 ±0.027	0.728 ±0.008	0.637 ±0.023	0.626 ±0.175	0.667 ±0.236	0.615 ±0.028	0.561 ±0.037	0.563 ±0.012	**0.589** ±0.043	0.651 ±0.049	0.604 ±0.082	0.772 ±0.006	0.687 ±0.018	0.619 ±0.006	0.588 ±0.017
MESHAgents [12]	0.738 ±0.011	0.695 ±0.016	**0.669** ±0.007	**0.658** ±0.022	0.724 ±0.009	**0.657** ±0.014	0.682 ±0.074	**0.817** ±0.207	0.611 ±0.026	0.565 ±0.060	0.569 ±0.015	0.582 ±0.044	0.658 ±0.041	0.567 ±0.092	0.770 ±0.010	0.694 ±0.019	0.620 ±0.008	0.594 ±0.017
Ours†‡ GPT-5.1	0.752 ±0.008	0.725 ±0.016	0.667 ±0.007	0.638 ±0.023	**0.732** ±0.009	0.647 ±0.020	0.602 ±0.154	0.577 ±0.211	0.628 ±0.015	**0.588** ±0.043	**0.583** ±0.004	0.565 ±0.025	**0.733** ±0.036	**0.713** ±0.067	0.785 ±0.013	0.713 ±0.027	0.621 ±0.004	0.601 ±0.009
Ours†‡ DeepSeek V3.2	**0.754** ±0.006	**0.733** ±0.015	**0.669** ±0.010	0.648 ±0.028	0.726 ±0.013	0.641 ±0.027	**0.684** ±0.187	0.560 ±0.339	**0.639** ±0.029	0.569 ±0.057	0.574 ±0.013	0.550 ±0.021	0.700 ±0.049	0.654 ±0.095	**0.789** ±0.007	**0.724** ±0.006	**0.624** ±0.011	**0.606** ±0.041
MLP																		
Experts [8]	0.761 ±0.006	**0.775** ±0.017	**0.663** ±0.013	0.701 ±0.011	0.738 ±0.008	0.631 ±0.028	0.628 ±0.133	0.643 ±0.217	0.595 ±0.018	0.621 ±0.166	0.564 ±0.008	0.509 ±0.031	0.669 ±0.010	0.612 ±0.149	0.784 ±0.012	0.705 ±0.030	0.594 ±0.010	0.591 ±0.006
MESHAgents [12]	0.754 ±0.010	0.768 ±0.015	0.658 ±0.012	0.686 ±0.019	0.728 ±0.008	0.597 ±0.020	0.640 ±0.042	0.660 ±0.159	0.576 ±0.036	0.577 ±0.144	0.555 ±0.006	0.470 ±0.060	0.673 ±0.024	0.662 ±0.091	0.772 ±0.012	0.696 ±0.039	0.598 ±0.005	0.599 ±0.039
Ours†‡ GPT-5.1	**0.763** ±0.007	0.770 ±0.020	0.662 ±0.015	0.712 ±0.017	0.735 ±0.010	**0.658** ±0.021	0.637 ±0.150	0.643 ±0.286	**0.619** ±0.022	0.563 ±0.108	0.545 ±0.008	0.521 ±0.044	0.616 ±0.095	0.546 ±0.138	**0.791** ±0.013	**0.716** ±0.035	**0.610** ±0.016	0.599 ±0.027
Ours†‡ DeepSeek V3.2	0.710 ±0.007	0.716 ±0.017	0.657 ±0.010	**0.757** ±0.037	**0.740** ±0.013	0.629 ±0.058	**0.741** ±0.091	**0.707** ±0.301	0.591 ±0.062	**0.637** ±0.409	**0.565** ±0.009	**0.542** ±0.078	**0.730** ±0.064	**0.708** ±0.155	0.784 ±0.010	0.705 ±0.017	0.599 ±0.015	**0.663** ±0.017
LDA																		
Experts [8]	0.760 ±0.006	0.699 ±0.007	**0.675** ±0.010	0.637 ±0.016	0.747 ±0.008	**0.667** ±0.018	0.636 ±0.142	0.610 ±0.292	0.619 ±0.038	0.593 ±0.063	0.583 ±0.013	**0.571** ±0.020	0.712 ±0.041	0.644 ±0.044	0.792 ±0.007	0.716 ±0.019	0.619 ±0.010	0.589 ±0.012
MESHAgents [12]	0.755 ±0.009	0.701 ±0.008	0.673 ±0.011	0.637 ±0.017	0.740 ±0.009	0.661 ±0.023	**0.666** ±0.060	**0.630** ±0.178	0.612 ±0.039	0.589 ±0.072	0.583 ±0.008	0.570 ±0.019	**0.730** ±0.035	**0.669** ±0.079	0.787 ±0.007	0.705 ±0.017	0.616 ±0.009	0.590 ±0.018
Ours†‡ GPT-5.1	**0.762** ±0.006	**0.728** ±0.006	0.671 ±0.010	0.644 ±0.017	0.750 ±0.014	0.654 ±0.024	0.658 ±0.087	0.543 ±0.266	0.604 ±0.031	0.595 ±0.060	**0.590** ±0.008	0.554 ±0.019	0.705 ±0.014	0.633 ±0.079	0.792 ±0.006	**0.726** ±0.032	0.614 ±0.012	0.590 ±0.028
Ours†‡ DeepSeek V3.2	**0.764** ±0.008	0.725 ±0.007	0.667 ±0.008	**0.673** ±0.042	**0.751** ±0.015	0.656 ±0.034	0.529 ±0.115	0.537 ±0.210	**0.621** ±0.044	**0.598** ±0.064	0.588 ±0.013	0.559 ±0.024	0.678 ±0.043	0.576 ±0.068	**0.794** ±0.004	0.720 ±0.023	**0.621** ±0.011	**0.593** ±0.027

0.686 ± 0.076 and a Recall of 0.643 ± 0.069. The ablation experiments identify the *Verifier* as the most critical module in the framework; its removal causes the most severe performance degradation, plummeting the AUC to 0.590 ± 0.006 and Recall to 0.544 ± 0.012, followed by the *Analyst*, Feedback, and SHAP Ranking.

Table 2: Ablation and interpretability analysis. **Left:** Ablation studies of the proposed framework. **Right:** Cluster-level interpretability, showing that composite phenotypes improve classification (silhouette scores $\times 10$ for readability).

Setting	$\overline{\text{AUC}} \pm \text{std}$	$\overline{\text{Recall}} \pm \text{std}$
w/o Analyst	0.672±0.007	0.626±0.021
w/o Verifier	0.590±0.006	0.544±0.012
w/o Feedback	0.679±0.085	0.640±0.157
w/o SHAP Ranking	0.685±0.011	0.637±0.016
CPAgents	**0.686±0.076**	**0.643±0.069**

Feature Set	$\overline{\text{Silhouette}} \times 10$
Raw Feature	0.293
Expert	0.276
MESHAgent	0.113
Constituent	0.094
Composite	**0.413**

Cluster Separability. We employ the silhouette score to quantify cluster separability. As shown in Table 2, our composite phenotype attains the highest value, about **1.5**× that of the expert-defined feature set. The raw base feature has a silhouette score of **0.029**, which increases to **0.041** after applying our composition (an absolute gain of 0.012, roughly **41%** relative), indicating improved

class separation and helping to explain the downstream classification benefits of the learned composite phenotypes.

4 Conclusion

We have presented an agentic, iterative phenotype-composition framework that discovers clinically interpretable composite cardiac imaging phenotypes and strengthens phenotype–disease association signals beyond expert-defined single traits. By combining analyst, proposer, and verifier agents with statistically grounded robustness checks, our approach yields compact, auditable phenotype libraries that are better aligned with underlying pathophysiology. Future work will extend the framework to multi-modal data to evaluate how composite phenotypes can further support clinical decision-making and trial design.

Acknowledgments. Z.L. and M.Q. is supported by the Departmental Funding of the Department of Mechanical Engineering, University College London. W.Z. is supported by the JADS programme and the UKRI Centre for Doctoral Training in AI for Healthcare (EP/S023283/1). P.M.M. acknowledges generous personal support from the Edmond J. Safra Foundation and Lily Safra, an NIHR Senior Investigator Award, Rosalind Franklin Institute, UKRI EPSRC funding, and UK Dementia Research Institute, which is funded predominantly by the UKRI Medical Research Council. W.B. is supported by EPSRC CVD-Net Programme Grant (EP/Z531297/1) and BHF New Horizons Grant (NH/F/23/70013). B.K. HPC resources from NHR@FAU (projects b143dc, b180dc), funded by federal and Bavarian state authorities and Gerhard Wellein's and his team's HPC approach. NHR@FAU hardware is partially funded by DFG 440719683. Additional support was received from ERC projects MIA-NORMAL 101083647, CHARMS 101246053, and ORACLE 101326871, DFG 513220538 and 512819079, and by the state of Bavaria (HTA). This research was conducted using the UK Biobank Resource under Application Number 18545. We thank all UK Biobank participants and staff.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Skelly AC, Dettori JR, Brodt ED. Assessing bias: the importance of considering confounding. Evidence-based spine-care journal. 2012;3(01):9-12.
2. Ruijsink B, Puyol-Antón E, Oksuz I, Sinclair M, Bai W, Schnabel JA, et al. Fully automated, quality-controlled cardiac analysis from CMR: validation and large-scale application to characterize cardiac function. Cardiovascular Imaging. 2020;13(3):684-95.
3. Haupt M, Maurer MH, Thomas RP. Explainable artificial intelligence in radiological cardiovascular imaging—A systematic review. Diagnostics. 2025;15(11):1399.
4. Bustin A, Fuin N, Botnar RM, Prieto C. From compressed-sensing to artificial intelligence-based cardiac MRI reconstruction. Frontiers in cardiovascular medicine. 2020;7:17.

5. Schuijf JD, Shaw LJ, Wijns W, Lamb HJ, Poldermans D, de Roos A, et al. Cardiac imaging in coronary artery disease: differing modalities. Heart. 2005;91(8):1110-7.
6. Lüscher TF, Wenzl FA, D'Ascenzo F, Friedman PA, Antoniades C. Artificial intelligence in cardiovascular medicine: clinical applications. European heart journal. 2024;45(40):4291-304.
7. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010;26(9):1205-10.
8. Bai W, Suzuki H, Huang J, Francis C, Wang S, Tarroni G, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. Nature Medicine. 2020;26(10):1654-62.
9. Allen NE, Lacey B, Lawlor DA, Pell JP, Gallacher J, Smeeth L, et al. Prospective study design and data analysis in UK Biobank. Science translational medicine. 2024;16(729):eadf4428.
10. Huda A, Castaño A, Niyogi A, Schumacher J, Stewart M, Bruno M, et al. A machine learning model for identifying patients at risk for wild-type transthyretin amyloid cardiomyopathy. Nature communications. 2021;12(1):2725.
11. Mukherjee P, Shen TC, Liu J, Mathai T, Shafaat O, Summers RM. Confounding factors need to be accounted for in assessing bias by machine learning algorithms. Nature Medicine. 2022;28(6):1159-60.
12. Zhang W, Qiao M, Zang C, Niederer S, Matthews PM, Bai W, et al. Multi-agent reasoning for cardiovascular imaging phenotype analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2025. p. 429-39.
13. Lan X, Feng J, Liu Y, Li Y, et al. AutoQual: An LLM Agent for Automated Discovery of Interpretable Features for Review Quality Assessment. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track; 2025. p. 1250-64.
14. Abhyankar N, Shojaee P, Reddy CK. Llm-fe: Automated feature engineering for tabular data with llms as evolutionary optimizers. 2025. arXiv preprint arXiv:2503.14434.
15. Wang Z, Wu J, Cai L, Low CH, Yang X, Li Q, et al. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. 2025. arXiv preprint arXiv:2503.18968.
16. Fallahpour A, Ma J, Munim A, Lyu H, Wang B. Medrax: Medical reasoning agent for chest x-ray. 2025. arXiv preprint arXiv:2502.02673.
17. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2005;67(2):301-20.
18. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30:4765-74.
19. Liu A, Mei A, Lin B, Xue B, Wang B, Xu B, et al. Deepseek-v3.2: Pushing the frontier of open large language models. 2025. arXiv preprint arXiv:2512.02556.