

<https://doi.org/10.1038/s41746-025-01676-1>

Digital twin integrating clinical, morphological and hemodynamic data to identify stroke risk factors



Marta Saiz-Vivó¹✉, Jordi Mill¹, Xavier Iriart^{2,3}, Hubert Cochet^{2,3}, Gemma Piella¹, Maxime Sermesant⁴ & Oscar Camara¹

Stroke remains a leading global cause of mortality, with ischemic stroke as the most common subtype. Atrial fibrillation (AF) increases ischemic stroke risk due to thrombus formation in the left atrium (LA), particularly in the left atrial appendage (LAA). Traditional risk assessments, like the CHA₂DS₂-VASc score, focus on clinical factors but often overlook LA morphology and hemodynamics. Existing studies either use mechanistic models with limited cases or rely solely on clinical data, missing hemodynamic insights. This study integrates statistical and mechanistic models within a Digital Twin framework, using unsupervised Multiple Kernel Learning on 130 AF patients. Combining LA morphology, hemodynamics, and clinical data improved patient stratification, identifying three phenogroups. The highest-risk group exhibited larger atrial dimensions, complex LAA structures, and elevated B-type natriuretic peptide levels. This study underscores the potential of Digital Twin models in assessing thrombus risk, emphasizing the need for further research to refine stroke prediction models.

Stroke is one of the leading causes of mortality worldwide, comprising 11.6% of all deaths in 2019¹. Ischemic stroke is the most frequent subtype and occurs when a blood vessel supplying the brain is blocked by a blood clot or thrombus, resulting in neurological death. Atrial fibrillation (AF) is the most common cardiac arrhythmia and one of the major risk factors of ischemic stroke; in AF patients, thrombus are formed inside the left atrium (LA), which may embolize to the brain. The main site of thrombus formation is the left atrial appendage (LAA)², a sack-like structure protruding from the LA cavity. The pathophysiology of thrombus formation in AF is multifaceted and complex. The Virchow's triad indicates that the main factors of thrombus formation are multi-domain, including blood stasis, hypercoagulability states, and endothelial dysfunction³. Classical models relate the ineffective contraction of the atrium during AF to blood stasis, favoring thrombus formation inside the LA, whilst more recent models advocate for an interplay between AF-caused abnormal atrial contractility, endothelial dysfunction, and alterations in LA structure⁴, as well as blood coagulability properties.

The CHA₂DS₂-VASc score is the standard clinical tool for thrombus risk assessment. It focuses on comorbidities and demographic factors such as age and sex. However, it does not consider specific LA morphological and hemodynamic characteristics, which may improve stroke risk

classification^{5–8}. For instance, several studies have reported associations between LA/LAA morphological factors with LA stasis and thrombosis^{9–11}.

Digital twins (DTs) are defined as the dynamic integration and augmentation of patient data using mechanistic models and statistical models^{12,13}. In stroke, DTs can support clinical decisions in different medical fields (e.g., neurology, cardiology) such as a more effective identification of risk factors in relation to thrombus formation. A comprehensive stroke DT should cover multiple organs and systems along the heart-brain axis, including blood flow and thrombosis models in the left heart, carotid arteries, and cerebral circulation with brain perfusion, metabolism, and cerebrospinal fluid models.

While complete stroke DTs are overwhelmingly complex, those focusing only in the left atria^{14,15} have already proven their utility in treatment planning^{15,16} and understanding left atrial blood flow patterns in AF patients with a risk of ischemic stroke. For doing so, mechanistic models based on computational fluid dynamics (CFD) models have been developed^{16–20}, capable of simulating blood flow in patient-specific atrial geometries. From these models, in-silico hemodynamic indices can be extracted to gain deeper insights on atrial hemodynamics^{15,16}. However, most studies have been applied to a limited number of cases, due to the difficulty of obtaining large cohorts of imaging data to personalize the

¹Physense, BCN MedTech, Department of Engineering, Universitat Pompeu Fabra, Barcelona, Spain. ²IHU Lyric, CHU Bordeaux, Université Bordeaux, Inserm, Pessac, France. ³Bordeaux University Hospital, Bordeaux, France. ⁴Inria, Université Côte d'Azur, Epione team, Sophia Antipolis, France.

✉ e-mail: marta.saiz@upf.edu

models. Beyond preventing robust statistical analysis of the simulation results, they are rarely analyzed in conjunction with a thorough analysis of the left atrial morphology^{17,20–23} and blood biomarkers, which have shown associations with thrombus formation^{24–29}.

Some statistical modelling methods, such as machine learning (ML), are capable of capturing non-linear and hidden relationships between input variables, and have been previously employed in medical applications for risk, outcome or treatment prediction^{30,31}. In the context of cardioembolic stroke, previous works have applied supervised ML models for thrombus risk assessment in AF patient cohorts based on multi-domain data. For instance, Zhao et al.³² integrated blood biomarkers and morphological data with consensus models based on Random Forest (RF) and eXtreme Gradient Boosting algorithms. Recently, Pieszko et al.³³ evaluated ML models for thrombus prediction based on clinical features and trans-thoracic echocardiography measurements. Nevertheless, these studies did not include parameters characterising the complex hemodynamics in the left atria. On the other hand, Pons et al.³⁴ jointly analysed clinical demographics (but not blood biomarkers), LA morphology (from low resolution 3D rotational angiographies) and in-silico hemodynamic indices, in a limited number of cases ($n = 30$), with a supervised machine learning model based on the RF algorithm.

The main limitations of supervised learning models are the need for hand-crafted features and the dependence on the quality and quantity of labelled data to guide the training phase. On the other hand, unsupervised ML methods can learn the inherent structure of the data and find patient similarities without being conditioned by a specific label. This allows the identification of ‘natural’ patient groups or phenogroups, from which exploratory variable analyses can be performed. Typically, dimensionality reduction techniques are employed to integrate medical variables describing a patient into a low-dimensional space. Subsequently, clustering algorithms are applied to group patients with similar characteristics. These methods have commonly been employed for cardiovascular risk and outcome prediction^{35–38}. Moreover, recent research has emphasized the value of cluster analysis in AF for adverse outcome prediction^{39–42}, although only considering clinical demographics and blood biomarkers. Furthermore, they mainly focus on analyzing future cardiovascular or neurological events rather than characterizing thrombogenesis. In a previous preliminary work⁴³, we performed an initial phenogroup study integrating morphological and in-silico hemodynamic data to analyse history of thrombus formation, although without considering clinical information.

The aim of our work was to develop a Digital Twin integrating multi-domain information to identify stroke risk factors in patients with non-valvular atrial fibrillation. Clinical data, morphological parameters and in-silico hemodynamic indices were jointly analysed in a cohort of 130 patients with unsupervised machine learning techniques to cluster patients into different phenogroups in relation with their history of thrombus. We hypothesized that integrating all available patient information into a digital twin would aid in the exploration of patient phenogroups associated with cardioembolic thrombus.

Figure 1 depicts a schematic overview of the developed methodology, including the construction of the digital twin of the left atria, and the analysis of multi-modal data using unsupervised machine learning. Given a computed tomography (CT) image, LA morphological parameters were automatically extracted after segmentation of the left atrial geometry. A 3D mesh was created from the LA binary mask to serve as the domain for patient-specific fluid simulations. Several in-silico hemodynamic indices were then derived from the fluid simulations, including the total number of particles remaining at the LAA after the simulation. Clinical data was provided but in cases with missing data, imputation techniques were employed. Feature selection techniques were also applied to obtain a condensed set of the features for the clustering step. Following this, unsupervised machine learning techniques were used to create reduced dimensionality spaces, and then cluster patients into different phenogroups based on feature similarities. Several domain-specific analyses were performed, where the input feature domain varied, including clinical, morphological, hemodynamics, or

a combination of these. Finally, the results were clinically interpreted to evaluate their association with a history of thrombus formation, and the phenogroups were named according to the increasing proportion of thrombus history (i.e., Phenogroup 0–2).

Results

The population of study consisted of 130 AF patients with severely diseased left atria and at high thromboembolic risk, including medical images and clinical data. For further details on the database, the reader is referred to the Methods section. Table 1 shows the parameters – clinical, structural and hemodynamic – with statistical significant differences between patients with and without history of thrombus. Non-significant variables can be found in Supplementary Tables 6–8. As observed in Table 1, dyslipidemia was the only clinical parameter associated with thrombogenic history. Regarding LA and LAA anatomical characteristics, patients with a history of thrombus had larger ostium dimensions. Significant differences were found in the proportion of multi-lobe LAAs (≥ 2 lobes), more common in patients with history of thrombus. Regarding the extracted in-silico hemodynamic indices, the total number of LAA particles was found significantly higher in the group with history of thrombus.

Distinct combinations of input data domains were integrated via multiple kernel learning (MKL) and analyzed with clustering algorithms (Table 4). Significant differences in thrombus formation history were observed in clusters (or phenogroups) obtained by combining morphological and in-silico blood flow features (Analysis 5), as well as those incorporating clinical, morphological, and blood flow features (Analysis 6). These phenogroups were further examined, including an intra-cluster analysis to characterize thrombus and non-thrombus patients within each phenogroup. Of the remaining domain combinations, only the clusters based solely on morphological features (Analysis 2) were further explored to assess their relationship with modeling hemodynamic descriptors. For all analyses, the 3-cluster configuration in a 2-dimensional space produced the highest silhouette score. To enhance the interpretability of the results, a Random Forest classifier was trained with the cluster class as labels. Then, a SHapley Additive exPlanations (SHAP) method⁴⁴ ranked the most important features for phenogroup stratification. For further details, refer to Methods section.

Clustering on digital twins based on morphological factors

Patient stratification based solely on LA anatomical features identified phenogroups with significant differences in the number of in-silico fluid particles remaining inside the LAA, indicative of LAA stasis. The SHAP analysis (Supplementary Fig. 6) revealed that the ostium area was the primary driver of phenogroup formation, especially for the classification of Phenogroup 0 patients. The phenogroup with the highest number of remaining fluid particles exhibited large ostium dimensions, along with larger LA and LAA sizes, and a high alignment between the left superior pulmonary vein (LSPV) and LAA. For further details on phenogroup attributes, refer to Supplementary Information-Section II.3.1.

Clustering on digital twins based on morphological and hemodynamic factors

Including the number of in-silico LAA particles as an input feature resulted in phenogroups with significant differences in the number of cases with previous thrombus formation. The phenogroup with the highest proportion of thrombus history (64%) presented the largest number of fluid particles in the LAA, alongside larger LA, LAA, and ostium areas as well as greater LAA/LSPV alignment. This group also showed significant differences in the LAA bending angle, with less bending or chicken-wing (CW)-like morphologies, and higher B-type natriuretic peptide (BNP) levels. For further details on phenogroup characteristics, refer to Supplementary Information-Section II.3.2. The ostium area was the most important feature for predicting phenogroups (Supplementary Fig. 7). The analysis of thrombus cases belonging to the phenogroup with lowest thrombus history revealed a significantly larger dyslipidemia proportion, increased fluid particle count, and steeper RSPV entry, among other factors (Supplementary Table 12).

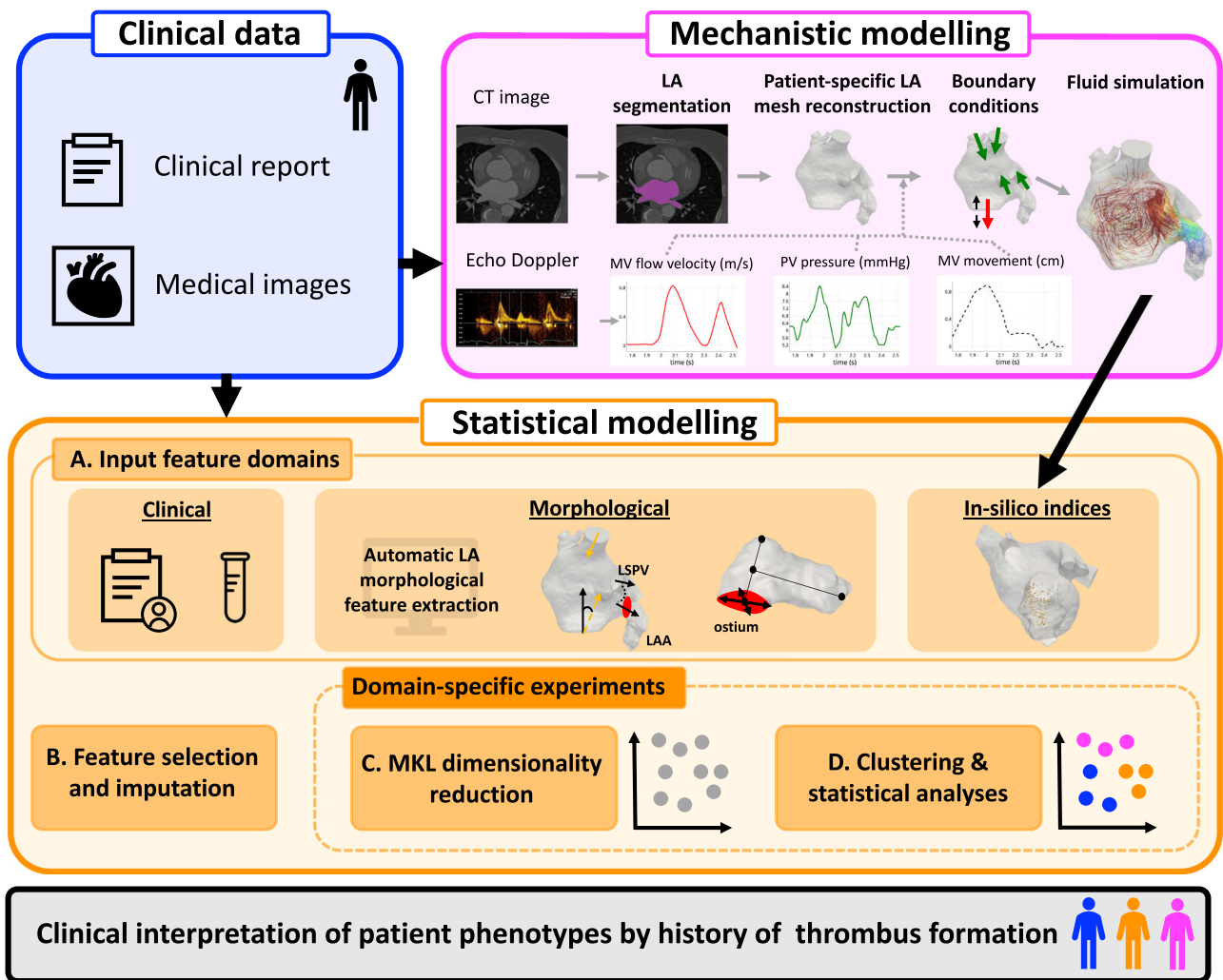


Fig. 1 | Overview of the methodological pipeline. Structured framework for integrating statistical and mechanistic modeling. AF atrial fibrillation, LA left atrium, LAA left atrial appendage, PV pulmonary vein, MKL multiple kernel learning algorithm, CT computed tomography.

Table 1 | Statistically significant clinical, morphological and in-silico hemodynamic characteristics of non-valvular atrial fibrillation (AF) patients with and without history of thrombus

Variable	No history of thrombus (n = 72)	History of thrombus (n = 58)	p-value
Clinical data - Comorbidities			
Dyslipidemia (%)	30/48 (62)	35/40 (88)	0.015
Morphological data			
Maximal ostium diam. (mm)	26.95 (23.63–32.08)	29.47 (25.71–34.14)	0.045
Minimal ostium diam. (mm)	19.42 (17.64–22.95)	22.52 (18.48–25.26)	0.009
Average ostium diam. (mm)	23.17 (21.22–27.13)	26.45 (21.90–28.78)	0.024
Ostium perimeter (mm)	74.27 (67.09–88.20)	85.10 (69.57–93.32)	0.023
Ostium area (mm ²)	412.08 (338.72–565.77)	543.58 (371.33–646.57)	0.020
Number of lobes (%)			
1	31 (43)	13 (22)	0.048
2	23 (32)	25 (43)	
≥3	18 (25)	20 (34)	
In-silico hemodynamic data			
Total LAA particles	130.0 (83.0–180.3)	207.5 (157.5–253.8)	<0.01

Variables are presented as, median (Q1–Q3) or n of patients (%). In variables with missing values, data presented as n/non-missing total (%). Statistical significance was determined by p value <0.05; see Supplementary Tables 6–8 for all analysed variables.

n number of patients, h/o history of, LAA left atrial appendage.

The bold values indicate variables with statistical significant differences (p < 0.05)

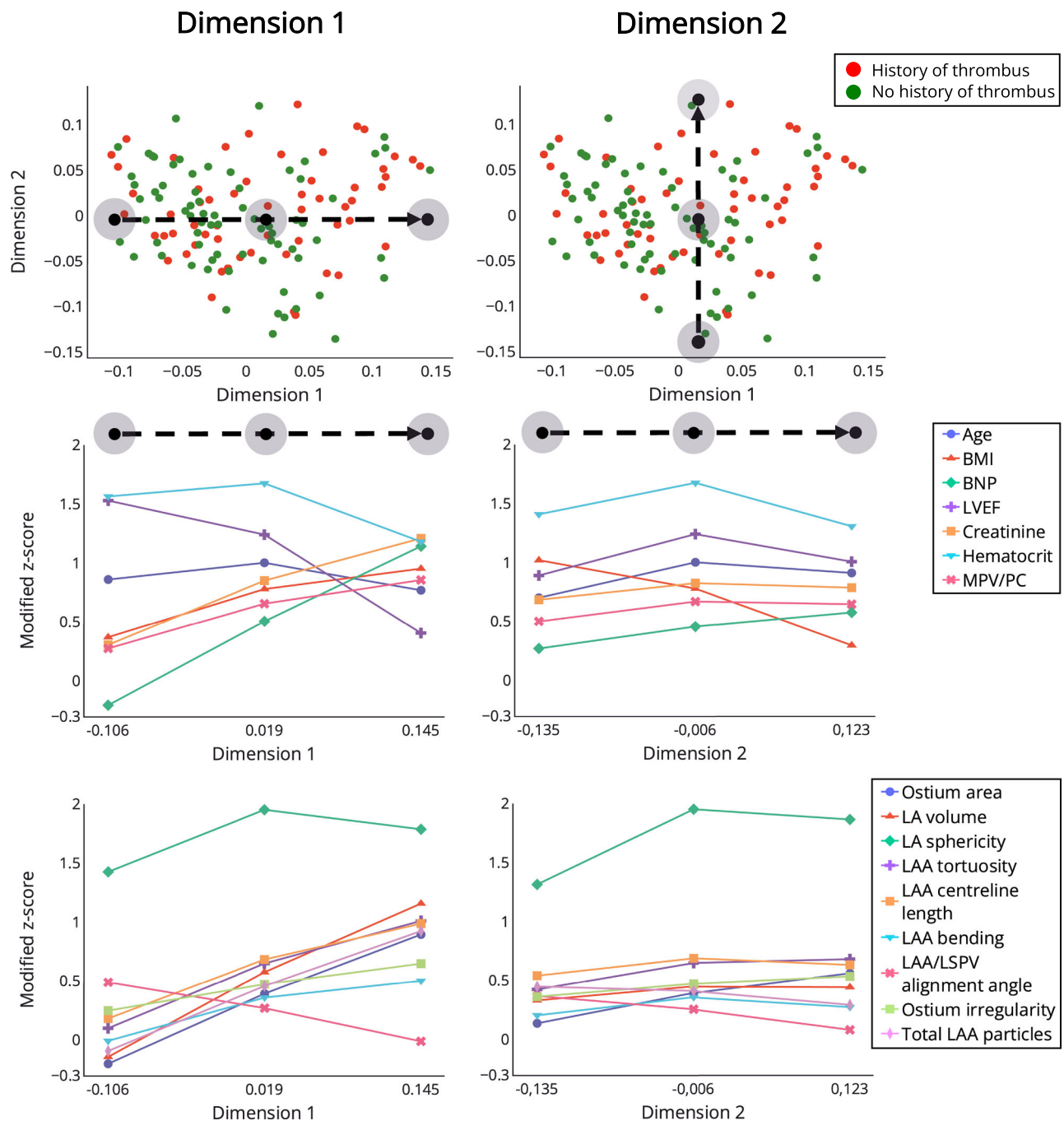


Fig. 2 | Exploration of low-dimensional outputs from clinical, morphological and hemodynamic inputs. Multiple kernel regression of three sampled points along dimension 1 (1st column) and dimension 2 (2nd column). Sampled points based on maximum, middle and minimum dimension value among the patient distribution. First row: low-dimensional output space showing cases with a history of thrombus (red circle) and without (green circle). Second row: regression of clinical features.

Third row: regression of morphological and hemodynamic features. Variables normalized with modified z-score. AF atrial fibrillation, BMI body mass index, BNP B-type natriuretic peptide, LA left atrium, LAA left atrial appendage, LSPV left superior pulmonary vein, LVEF left ventricular ejection fraction, MPV mean platelet volume, PC platelet count.

Clustering on digital twins based on morphological, hemodynamic and clinical factors

Combining the clinical domain into the in-silico circulatory and morphological domains improved patient stratification based on thrombus history, with one phenogroup showing a 70% proportion of prior thrombus formation.

To interpret the ML output space derived from multi-domain parameters, the low-dimensional space was analysed with a multiple kernel regression of input features along each dimension (Fig. 2). Specifically,

starting from the center of mass of the output representation, we move along each dimension and back-project the output representation to the corresponding input descriptors. Expressing the input descriptors as a function of the output dimensions makes it easier to interpret how the variance is embedded in the output representation. Dimension 1 (Fig. 2, 1st column) captured a steep increase in BNP levels from left to right, along with moderate increases in creatinine, mean platelet volume to platelet count (MPV/PC) ratio, and body mass index (BMI), whilst encoding a decrease of left ventricular ejection fraction (LVEF) and hematocrit levels (Fig. 2, 1st

column, 2nd row). Morphological and circulatory features such as the ostium area, LA volume, LAA tortuosity and the number of in-silico LAA particles showed a steep increase from left to right, whilst the LAA/LSPV misalignment (including CW morphologies) steeply decreased (Fig. 2, 1st column, 3rd row). Dimension 2 (Fig. 2, 2nd column) showed less marked differences in opposing sides, especially with regards to clinical variables. Only BNP showed an increasing trend from bottom to upper regions, whilst BMI decreased. We observe greater differences in shape-related characteristics, especially regarding the ostium area, the LA sphericity and the LAA/LSPV alignment, increasing from bottom to upper regions.

The clustering analysis led to 3 phenogroups, with differing characteristics (Fig. 3 and Table 2). Phenogroup 2 patients ($n = 30$) exhibited a significantly higher number of cases with history of thrombus proportion compared to both Phenogroup 0 (70% vs 33%, p -value = 0.016) and 1 (70% vs 40%, p -value = 0.03). As seen in Table 2 and Fig. 3, patients in Phenogroup 2 showed significantly larger LA volume and ostium area (although not regarding the ratio ostium area/LAA volume), LAA complexity, LA sphericity and LAA/LSPV alignment in non-CW morphologies (Fig. 3d and f), as well as the highest number of fluid particles remaining in the LAA (Fig. 3b), compared with other patients. Figure 4a provides a visualization of the residual fluid particles within the LAA for three sample phenogroup patients, with representative particle counts. Figure 4b summarizes the main clinical, anatomical and blood flow properties of each identified patient phenogroup.

Clinically, Phenogroup 2 patients had the highest BNP levels, ratio MPV/PC, creatinine levels (within physiological range after gender adjustment) and proportion of non-paroxysmal AF cases, and lower LVEF and hematocrit levels (both within physiological range). Both phenogroups 0 ($n = 42$) and 1 ($n = 58$) had a significantly lower number of history of thrombus cases, whilst exhibiting differences in their clinical characteristics (Fig. 3c and e). On one hand, Phenogroup 1 patients were characterized by significantly smaller BNP levels and ratio MPV/PC, compared to Phenogroup 2, whereas few clinical differences were observed between Phenogroup 0 and Phenogroup 2 after adjusted pairwise comparisons. Nonetheless, Phenogroup 0 patients presented significantly smaller ostium areas, LA sphericity index, LAA volume and ratio LAA/LA volume compared with Phenogroup 2, coupled with a significantly lower proportion of fully aligned LAA/LSPV non-CW morphologies. The intra-cluster analysis revealed that Phenogroup 2 patients without a previous history of thrombus formation ($n = 9$) had smaller LA volumes, less steep LIPV entry angles, and lower BMI, hematocrit levels, and hypertension proportions (Supplementary Table 13). In contrast, thrombus cases in Phenogroup 0 showed significantly higher MPV (Supplementary Table 14).

The SHAP feature importance analysis (Fig. 5) showed that BNP, LA volume, ostium area, and, to a lesser extent, the total number of LAA particles were the most critical features for differentiating between clusters. Specifically, BNP played a key role in clustering Phenogroups 0 and 1, while ostium area was the primary feature for clustering Phenogroup 2. A Random Forest (RF) classifier, trained using the most important features from the clustering model to predict thrombus formation, achieved 84% precision, 81% recall, and an 81% F1-score. For further details on the supervised model, refer to Supplementary Information-Section I.5. Interestingly, SHAP analysis of the RF model identified the number of LAA particles as the most influential feature, followed by LAA centerline length, ostium area, and BNP (Supplementary Fig. 8).

Robustness and generalization of MKL and clustering model

The robustness of the MKL and clustering model was confirmed by training on different data subsets, yielding similar phenogroups (Supplementary Fig. 9). Stability analysis of the K-means clustering, assessed using a bootstrapping approach, resulted in a stability score of 0.86, suggesting a reliable clustering solution⁴⁵. Internal validation (Supplementary Information-Section II.7.2) revealed consistent trends in phenogroups when comparing a random training subset (75%) and an unseen validation subset (25%) projected into the output space, underscoring the strong generalization

capability of the model. For further information on the internal analyses, refer to Supplementary Information-Section I.6.

Phenogroup stability analysis

To understand the incremental value of clinical markers, we performed a phenogroup stability analysis (Methods section), examining patient reclassifications after the inclusion of clinical variables. Phenogroup 2 displayed the highest stability (i.e., lowest number of reassignments) and neighboring patient similarities (Supplementary Fig. 10); however, reclassifications contributed to an increased proportion of thrombus history within this phenogroup. In contrast, phenogroups 0 and 1 exhibited significant instability. BNP levels emerged as a key factor influencing patient reclassifications.

Discussion

The development of left atrial digital twins is vital for improving stroke treatment and prevention in patients with AF. In this paper, we integrate multi-modal data (i.e., demographics, blood biomarkers and LA morphology) with indices from mechanistic models based on CFD simulations of LA blood flow, to develop digital twins of the left atrium in a cohort of severely diseased AF patients with a large proportion of history of thrombus. We explored patient phenogroups across different combinations of clinical, morphological, and circulatory domains to uncover their interrelations. A clinical interpretation of the resulting phenogroups was conducted, focusing on the proportion of patients in each group with a history of thrombus formation.

Our findings demonstrate that integrating all three domains enhances patient stratification based on past thrombus formation, highlighting factors that may contribute to an increased risk of thrombus formation in severely diseased patients. Not using the clinical label of past thrombus formation to guide patient stratification, allowed the model to explore the underlying data structure and consider subclinical thrombus formation within the obtained phenogroups. Although strong relationships were found between morphological factors and simulated hemodynamic parameters (Supplementary Table 9), incorporating in-silico hemodynamic indices was essential for stratifying patients by thrombus history, and underscores the importance of physics-based models to understand blood stasis in AF patients. While prior studies have endeavoured to identify risk factors of thrombus formation, we propose that the analysis of thrombus formation must be approached from a multi-domain perspective, considering clinical patient characteristics, LA shape and anatomical information, and blood flow information.

Considering LA morphology, in-silico blood flow, and clinical input domains, we identified 3 patient phenogroups with significant differences on history of thrombus formation. Patients in the high history of thrombus group (Phenogroup 2) presented large atrial geometries and complex-shaped LAAs with enlarged LAA volumes and ostium dimensions, coupled with the highest number of in-silico LAA particles. Indeed, tortuous LAAs may lower blood flow velocities at the ostium entrance, which could hinder blood washout²⁰, and larger LAA volumes have been related to increased LAA stasis²¹. Most Phenogroup 2 patients were diagnosed with non-paroxysmal or chronic AF, which may result in enlarged LA and LAAs due to adverse atrial remodeling^{46,47}, and warns of the LAA stasis risk in non-paroxysmal AF patient subtypes, as previously reported in Liu et al.⁴⁸. Similar patterns were observed in cluster analyses driven solely by morphological and hemodynamic factors, with the exception of a higher LA sphericity or balloon-like LAs shown in Phenogroup 2 patients, highly related to atrial remodelling and ischemic risk⁹.

On the other hand, upon normalizing the ostium area by LAA volume we observed the opposite trend (small normalized ostium area in Phenogroup 2 compared with Phenogroup 0, a group with low history of thrombus cases). This trend was similarly observed in other domain combinations (i.e., employing only morphological or a combination of morphological and hemodynamic factors as input). Although larger ostium dimensions have been previously related to increased thrombus risk and LAA stasis^{34,49}, recent studies suggest that a reduced ostium relates with

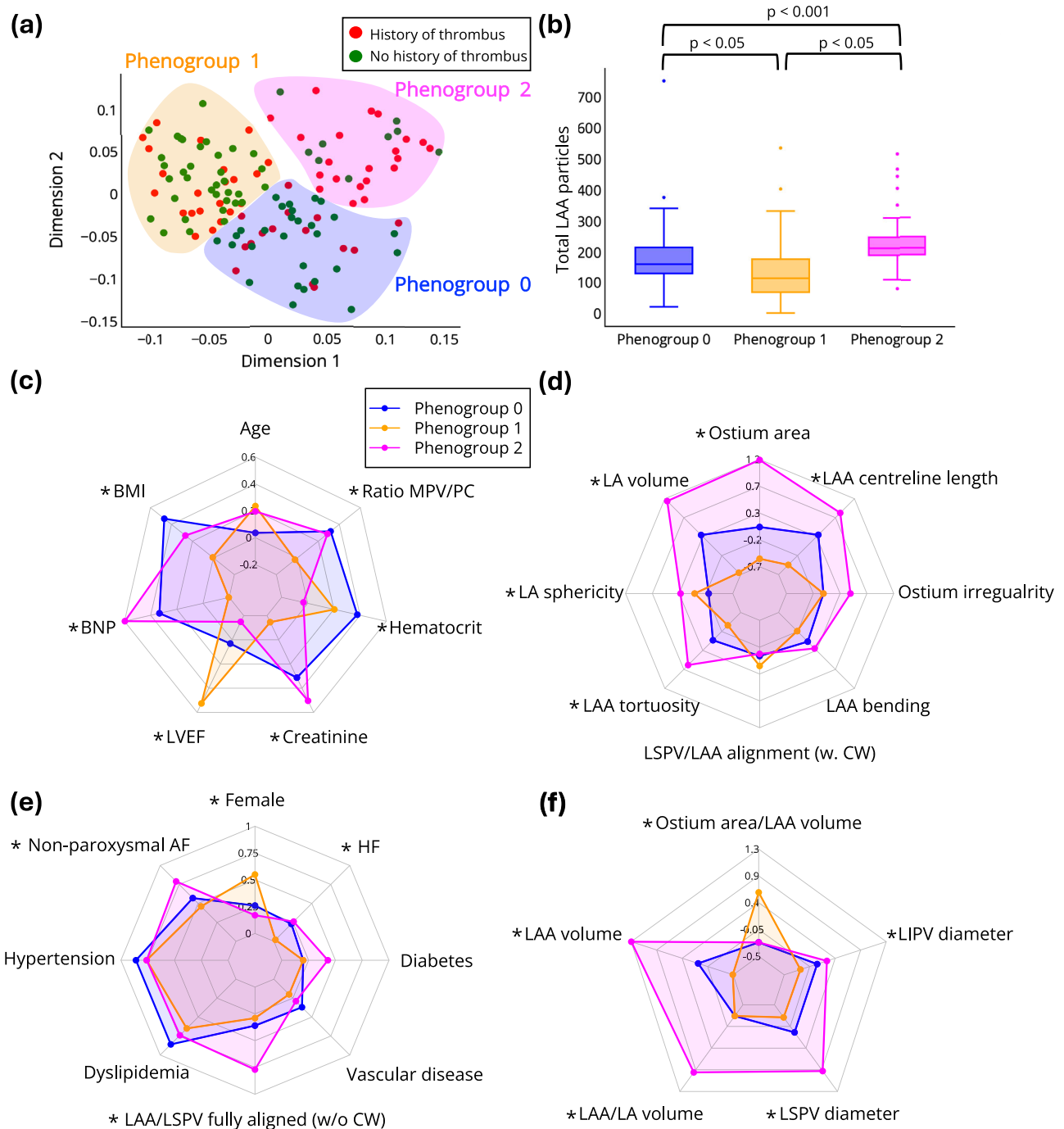


Fig. 3 | Clustering analysis results using clinical, morphological, and hemodynamic inputs. The figure shows (a) the output cluster distribution showing cases with a history of thrombus (red circle) and without (green circle). **b** Distribution of in-silico LAA particle counts. Radar charts showing median normalized values for (c) input demographic and laboratory markers, **d** input shape-related markers, **e** descriptive categorical variables, and **f** descriptive continuous variables, of each cluster. The blue line represents Phenogroup 0, the orange line represents

Phenogroup 1, and the magenta line represents Phenogroup 2. Categorical variables shown as a proportion (%). Variables with statistical significant differences (<0.05) shown with *. AF atrial fibrillation, BMI body mass index, BNP B-type natriuretic peptide, CW chicken wing, MPV mean platelet volume, PC platelet count, LVEF left ventricular ejection fraction, LA left atrium, LAA left atrial appendage, LIPV left inferior pulmonary vein, LSPV left superior pulmonary vein.

higher LAA stasis due to washout impairment; however, this effect saturates in remodeled atria due to the correlation between LAA volume and ostium dimensions²¹. Figure 4b shows a clear example of a Phenogroup 2 patient with a remodelled LAA with a large ostium and, correspondingly, a large number of remaining fluid particles. Our findings align with the results reported by Dueñas-Pamplona et al.²¹ and highlight the importance of considering atrial remodeling and AF subtype when evaluating the blood stasis risk of specific ostium dimensions.

When considering morphological and hemodynamic input feature domains, the phenogroup with increased LAA stasis and a history of thrombus also displayed reduced LAA bending and a significantly lower prevalence of CW morphologies, akin to the findings in Di Biase et al.¹⁰. The effect of a pronounced LAA bending on LAA stasis, as seen in CW morphologies, requires further study. We hypothesize that an increased LAA bending might offer protection by hindering blood flow from reaching the tip of the LAA, though this was not clearly observed in the sample

Table 2 | ML training and inference variables for clinical, morphological and blood flow domains

Variables	Phenogroup 0 (n = 42)	Phenogroup 1 (n = 58)	Phenogroup 2 (n = 30)	p-value
Input ML variables				
Ostium area (mm ²)	462.72 (398.53–546.42) ^{b,c}	349.76 (290.37–465.40) ^{a,c}	701.63 (598.81–784.30) ^{a,b}	<0.01
LA Volume (ml)	171.23 (152.20–181.26) ^b	126.27 (101.77–139.75) ^{a,c}	211.64 (164.40–247.94) ^b	<0.01
LAA tortuosity	0.39 (0.35–0.46) ^c	0.36 (0.32–0.41) ^c	0.44 (0.40–0.48) ^{a,b}	<0.01
LAA centerline length (mm)	47.95 (42.14–51.28) ^b	41.39 (36.13–45.40) ^{a,c}	52.73 (44.0–55.63) ^b	<0.01
LAA/LSPV alignment angle (2D) (deg)	30.9 (17–51)	35 (19–47)	30 (14–37)	0.33
Ostium irregularity	0.02 (0.01–0.05)	0.02 (0.01–0.04)	0.03 (0.01–0.05)	0.53
LA sphericity index	0.75 (0.73–0.79)	0.76 (0.73–0.78)	0.77 (0.76–0.80)	0.0470
LAA bending angle (deg)	127 (106–137)	121.5 (98.66–136.79)	130.6 (117.6–142)	0.45
Total LAA particles	159 (130–214) ^{b,c}	114.5 (70.25–177.25) ^{a,c}	214 (193–246.75) ^{a,b}	<0.01
Age	73.50 (70–78)	76 (71–82)	75.5 (68.75–79.75)	0.3171
BMI	28.08 (23.59–32.20) ^b	24.91 (23.75–26.50) ^a	26.70 (23.00–29.20)	<0.01
BNP (pg/ml)	198 (157–255) ^b	92.18 (59.50–129.06) ^{a,c}	250.81 (211.78–318.94) ^b	<0.01
LVEF (%)	56.61 (54.33–60)	59.81 (55.46–60) ^c	55.45 (53.06–59.45) ^b	<0.01
Creatinine (μmol/l)	91.97 (83–106.11) ^b	77.50 (70–92.5) ^{a,c}	98 (85.06–112.39) ^b	<0.01
Hematocrit (%)	40.94 (38.74–42.38) ^c	40.25 (38.92–42.42)	39.35 (38.30–39.98) ^a	0.027
Ratio MPV / platelet count	4.93 (4.43–5.49) ^b	4.37 (3.76–4.94) ^{a,c}	4.88 (4.59–5.30) ^b	0.01
Descriptive variables – clinical				
Female (%)	11 (26) ^b	32 (55) ^{a,c}	5 (30) ^b	<0.01
Heart failure (%)	8/35 (23) ^b	1/50 (2) ^{a,c}	6/23 (26) ^b	<0.01
Non-paroxysmal AF (%)	21/37 (57)	23/50 (46) ^c	19/24 (79) ^b	0.024
Descriptive variables – morphological				
Maximal ostium diameter (mm)	28.91 (25.94–32.32) ^{b,c}	24.40 (21.75–28.05) ^{a,c}	34.74 (31.95–37.28) ^{a,b}	<0.01
Minimal ostium diameter (mm)	20.45 (18.44–22.58) ^c	18.51 (16.67–21.60) ^c	25.73 (23.38–27.79) ^{a,b}	<0.01
Average ostium diameter (mm)	24.70 (22.32–26.83) ^{b,c}	21.43 (19.28–24.94) ^{a,c}	30.35 (27.71–31.78) ^{a,b}	<0.01
Ostium perimeter (mm)	78.29 (73.62–87.66) ^{b,c}	68.15 (61.57–78.66) ^{a,c}	97.40 (88.62–102.86) ^{a,b}	<0.01
Ostium eccentricity	0.30 (0.23–0.35) ^b	0.24 (0.17–0.29) ^a	0.25 (0.19–0.30)	<0.01
LAA height (mm)	18.52 (14.88–22.19) ^{b,c}	15.05 (12.49–17.25) ^{a,c}	23.57 (20.62–25.07) ^{a,b}	<0.01
LAA depth (mm)	36.98 (32.42–41.96)	33.78 (29.33–37.72) ^c	41.03 (34.79–44.86) ^b	<0.01
LAA volume (ml)	11.12 (9.61–15.76) ^{b,c}	7.57 (5.65–9.96) ^{a,c}	17.97 (13.92–20.30) ^{a,b}	<0.01
LAA surface area (cm ²)	34.2 (31.2–43.5) ^{b,c}	27.9 (23.1–33.4) ^{a,c}	46.9 (39.9–52.5) ^{a,b}	<0.01
LA centreline length (mm)	101.71 (94.94–106.96)	90.60 (84.01–96.98) ^c	107.57 (102.54–116.59) ^b	<0.01
LA surface area (cm ²)	203.2 (194.9–216.2) ^b	168.3 (151.9–184.4) ^{a,c}	235.0 (202.8–256.0) ^b	<0.01
Ratio ostium area / LAA volume	0.04 (0.04–0.05)	0.05 (0.04–0.06) ^c	0.04 (0.03–0.05) ^b	<0.01
Ratio LAA / LA volume	0.06 (0.05–0.08) ^c	0.06 (0.05–0.08) ^c	0.09 (0.07–0.10) ^{a,b}	<0.01
LSPV average ostium diam. (mm)	18.15 (15.67–20.48)	16.90 (15.18–19.48) ^c	21.37 (16.39–22.89) ^b	<0.01
LIPV average ostium diam. (mm)	16.37 (13.35–17.33)	15.39 (12.72–17.73)	16.94 (15.13–20.90)	0.046
Shape & alignment label (%)	^c	^c	^{a,b}	<0.01
LAA/LSPV not aligned (w/o CW)	10 (24)	17 (29)	3 (10)	
LAA/LSPV moderately aligned (w/o CW)	5 (12)	6 (10)	0 (0)	
LAA/LSPV fully aligned (w/o CW)	15 (36) ^c	17 (29) ^c	23 (77) ^{a,b}	
CW	12 (29)	18 (31)	4 (13)	
Outcome variable				
History of Thrombus (%)	14 (33) ^c	23 (40) ^c	21 (70) ^{a,b}	0.006

Variables are presented as median (Q1–Q3) or *n* of patients (%). In variables with missing values, data presented as *n*/non-missing total (%). *n* number of patients,

LA left atrium, LAA left atrial appendage, LSPV left superior pulmonary vein, LIPV left inferior pulmonary vein, LVEF left ventricular ejection fraction, MPV mean platelet volume, PC platelet count, AF atrial fibrillation, BNP B-type natriuretic peptide, CW chicken-wing

^aStatistically different from cluster 0 (*p* value <0.05).

^bStatistically different from cluster 1 (*p* value <0.05).

^cStatistically different from cluster 2 (*p* value <0.05).

The bold values indicate variables with statistical significant differences (*p* < 0.05)

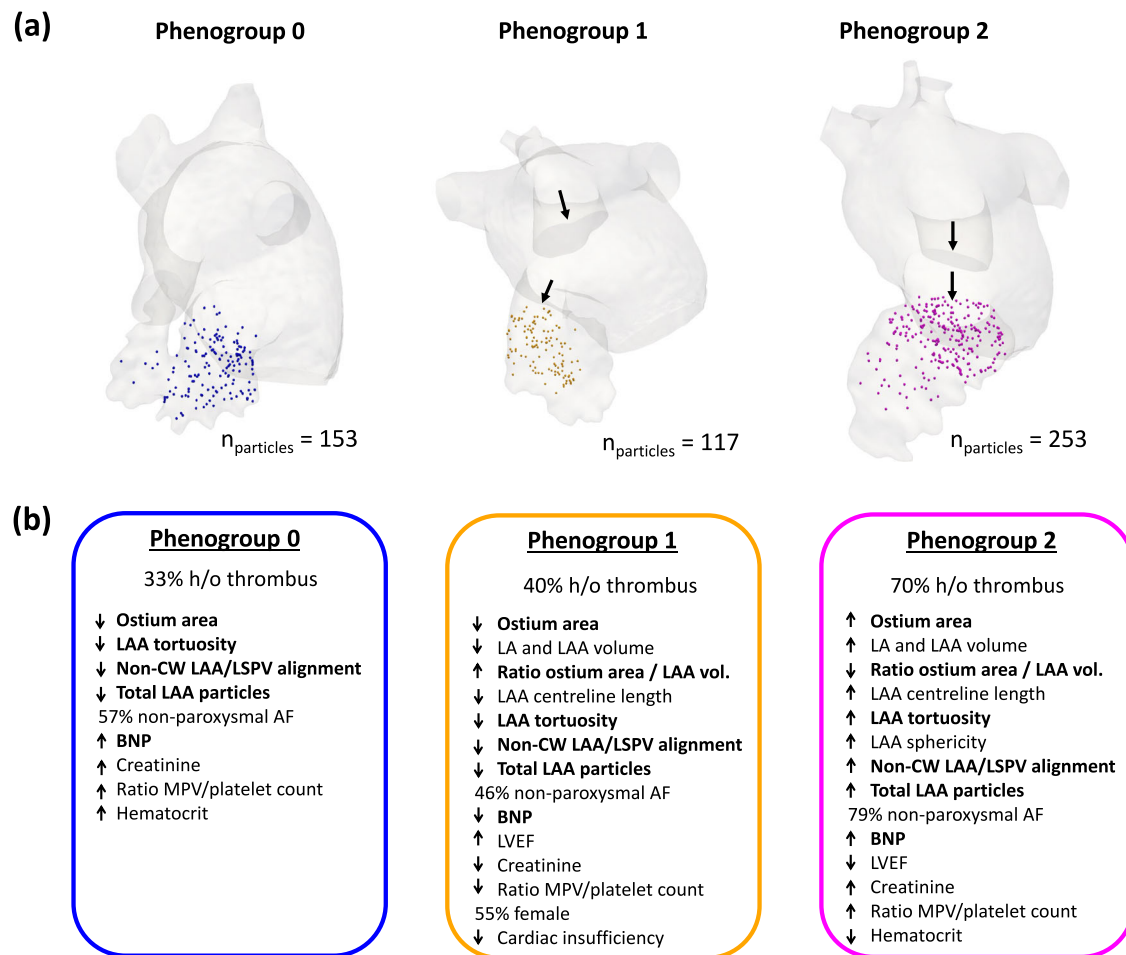


Fig. 4 | Phenogroup comparison using clinical, morphological, and hemodynamic inputs. The figure shows (a) a visualization of computational fluid dynamics particles remaining inside the left atrial appendage (LAA) at the last time-step (end-systolic phase) in phenogroup patients showcasing a representative number of LAA particles ($n_{\text{particles}}$). Black arrows indicate degree of LAA and left superior pulmonary

vein (LSPV) alignment. **b** Characterization of the 3 patient phenogroups. AF atrial fibrillation, BNP b-type natriuretic peptide, CW chicken-wing, LA left atrium, LAA left atrial appendage, LSPV left superior pulmonary vein, LVEF left ventricular ejection fraction, MPV mean platelet volume.

patients from Fig. 4b. Nevertheless, the reduced difference in patients with CW morphologies between phenogroups after including clinical variables as input (Table 2) suggests that adverse clinical and demographic factors may diminish this protective effect.

The number of LAA lobes showed significant differences in the univariate analysis, with the proportion of multi-lobe LAAs (≥ 2 lobes) found higher in patients with history of thrombus, similarly reported Yamamoto et al.⁴⁹. However, no statistically significant differences were observed between the obtained phenogroups. We hypothesize that the included morphological input features could not capture these shape variations, and that integrating this information through mixed-feature analysis could prove beneficial in the study of thrombus formation.

Regarding the pulmonary veins configuration in phenogroups with high LAA stasis or history of thrombus formation, we consistently observed a high alignment of LAA and LSPV, especially in non-CW morphologies, in phenogroups with large LAA stasis across different domain combinations (Table 2 and Fig. 4b, Supplementary Tables 9 and 10). We posit that the high alignment between the LAA and LSPV may lead to increased blood flow into the LAA. This effect appears diminished in CW morphologies, possibly due to different blood flow patterns at the ostium entrance¹⁷. Larger LSPV and left inferior pulmonary vein (LIPV) ostium dimensions were also observed, likely a consequence of pulmonary vein (PV) remodelling changes in non-paroxysmal AF subtypes⁵⁰. Regarding the PVs orientation or entry

angle to the LA, trends were observed relating a higher steepness of the right superior pulmonary vein (RSPV) and LIPV entry angle with increased LAA stasis, and history of thrombus formation, although further studies should be performed to fully ascertain their role in left atrial hemodynamics.

The clinical characteristics of Phenogroup 2 patients (morphological, blood flow and clinical variables) include elevated BNP levels (median 214 pg/ml). BNP is a cardiac neurohormone secreted in response to elevated left ventricular (LV) filling pressures²⁷. Pathophysiologically, the loss of atrial contraction and chronic elevation of LA pressure may lead to adverse atrial remodelling and AF persistence⁵¹, causing a disruption in normal LV filling pressures and subsequent BNP secretion⁵². High BNP has been linked to LA stasis and LAA thrombus risk³². Similar patterns were observed when phenogroups were identified by morphological and hemodynamic variables only, highlighting the relationship between morphological traits and stasis with high-risk clinical markers. LVEF was lower in Phenogroup 2 but within normal ranges³³. Similarly creatinine, a marker for kidney disease and LAA thrombus risk²⁴, was higher but within physiological limits. Elevated mean platelet volume (MPV) has been linked to increased platelet activation and pro-inflammatory states, potentially contributing to thrombus development^{54,55}. Additionally, a higher MPV/PC ratio has been recently associated with LA stasis²⁵. In our study, MPV/PC ratio was higher in Phenogroup 2 patients, indicating a more pro-thrombotic state. In contrast, hematocrit levels, linked to LAA stasis²⁶, were lower in Phenogroup 2

although within physiological range. Dyslipidemia correlated strongly with thrombus history in uni-variate analyses (Supplementary Table 6), aligning with Chen et al.⁵⁶, although no significant differences were observed between phenogroups.

Phenogroups 0 and 1 exhibited a significantly lower incidence of thrombus formation compared to Phenogroup 2, despite their differing characteristics, particularly in relation to high-risk blood biomarkers (i.e., high MPV/PC in Phenogroup 0). This suggests that specific LA morphological and hemodynamic features, such as a smaller ostium area, LAA tortuosity, and low LAA/LSPV alignment, may offer protection against thrombus formation, even in patients with a high pro-thrombotic risk (Fig. 3). Nonetheless, the significantly higher MPV observed in Phenogroup 0 patients with a history of thrombus, emphasizes the importance of considering clinical markers when evaluating thrombus risk.

The SHAP analysis revealed individual feature importance in the clustering model, with the most influential factors spanning morphological, hemodynamic, and clinical domains (i.e., blood biomarkers) (Fig. 5). The similar feature importance trends observed in the thrombus prediction

model (Supplementary Fig. 8b) underscores the value of a multi-domain approach for the prediction of thrombus formation. As observed in Fig. 5, BNP level was the key feature in identifying phenogroups, especially at lower-risk, while ostium area and LA volume were critical in identifying high-risk phenogroups. Indeed, ostium area ranked as the third most important predictor in the thrombus prediction model, after total particle count and LAA centerline length, highlighting the critical role of morphological and hemodynamic factors in thrombus risk assessment.

The phenogroup stability analysis revealed that including clinical variables resulted in multiple reclassifications within the lower-risk phenogroups (Supplementary Fig. 10), likely due to the absence of significant differences in thrombus history among them. Consequently, these groups can be considered distinct low-risk patient signatures. In contrast, Phenogroup 2 (high-risk phenogroup) exhibited high stability after incorporating clinical variables, underscoring the importance of morphological and hemodynamic characteristics in identifying phenotypes with increased thrombus risk. Nonetheless, the reclassifications resulting from the inclusion of clinical variables enhanced patient stratification by thrombus history, highlighting their incremental value in assessing thrombus risk. Therefore, clinical variables such as demographics and blood markers should be included in stroke digital twins, with BNP levels playing a major role in increasing thrombus risk stratification.

When comparing our results with previous joint analysis studies, we find agreement with our previous study on a different database³⁴, which highlighted the importance of large ostium diameters, LAA centerline length, and hemodynamic indices in relation to history of thrombus or transient ischemic attack. However, in the former analysis we did not consider clinical characteristics. On the other hand, Zhao et al.³² similarly found high BNP levels and LAA width being related to LAA thrombosis, although detailed morphological aspects and LA in-silico hemodynamic parameters were not assessed. Pieszko et al.³³ reported that LA volume and LVEF had the highest contribution in their supervised thrombus prediction model. The importance of LA volume coincides with our findings, as it was crucial in discriminating between phenogroups. However, while LVEF was statistically different among clusters, it remained within the physiological range in our study. Previous AF phenogroup studies^{39–42} rely solely on clinical characteristics (i.e., age and comorbidities) to relate patient clusters with upcoming major adverse cardiovascular events such as stroke. This differs from our findings, possibly attributed to variations in the outcome label (upcoming adverse cardiovascular events vs history of thrombus). Additionally, it is likely influenced by the advanced age and severe baseline disease status of most patients in our cohort.

While our study provides valuable insights into the interplay between LA morphology, in-silico hemodynamic features, and clinical/demographic factors in relation to thrombus formation, it is important to acknowledge several limitations. First, the sample size was relatively modest, and the studied population only included patients with severely diseased left atria who were candidates for LAAO implantation. Therefore, the findings of this study may not be directly generalizable to the broader AF patient

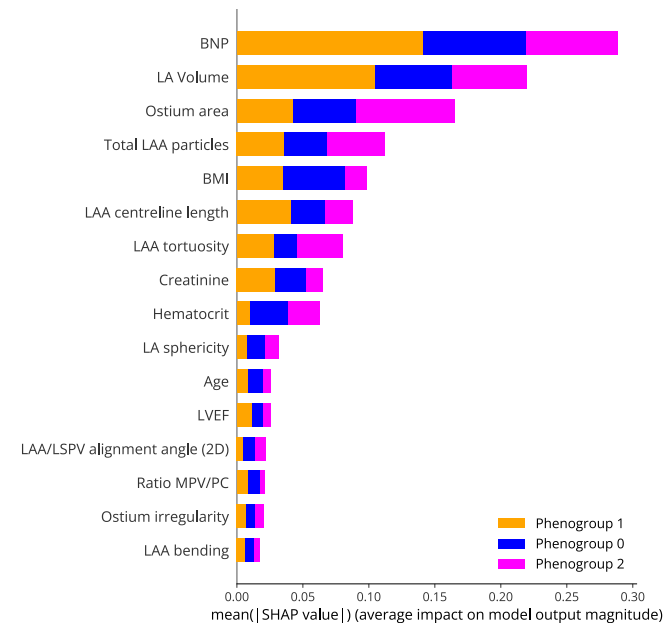


Fig. 5 | SHAP-based feature importance using clinical, morphological and hemodynamic inputs. Bar plot showing the average absolute SHAP values per feature, with taller bars indicating a greater contribution to predicting specific phenogroups. Features are ranked from highest to lowest importance. BNP b-type natriuretic peptide, BMI body mass index, LA left atrium, LAA left atrial appendage, LSPV left superior pulmonary vein, LVEF left ventricular ejection fraction, MPV mean platelet volume, PC platelet count.

Table 3 | Set of extracted features and their respective domains, categorized into machine learning (ML) input features and post-ML descriptive features

Domain	Input ML variables	Descriptive variables
Clinical	Age, BMI*, BNP levels*, creatinine levels*, hematocrit levels*, ratio MPV/PC*, LVEF*	AF type, gender, HF, HT, diabetes type 2, vascular disease, CRP, ratio neutrophyl/lymphocyte dyslipidemia, HAS-BLED, troponin, fibrinogen, urea, GFR
Morphological	Ostium area, ostium irregularity, LA volume, LAA centerline length, LAA tortuosity, LA sphericity, LAA bending angle, LAA/LSPV alignment angle (2D)	LAA volume/area/height/depth, ostium diameter/eccentricity/ perimeter, number of LAA lobes, LA area/centerline length, LAA shape/alignment label, ratio ostium area/LAA volume, ratio LAA/LA volume, LAA/LSPV alignment angle (3D), PVs diameter and orientation
In-silico hemodynamic	Total LAA particles	Particle age

AF atrial fibrillation, BMI body mass index, BNP B-type natriuretic peptide, CRP c-reactive protein, GFR glomerular filtration rate, HF heart failure, HT hypertension, LA left atrium, LAA left atrial appendage, LSPV left superior pulmonary vein, LVEF left ventricular ejection fraction, MPV mean platelet volume, PC platelet count.

*Missing data imputation was performed.

population. Another limitation was the absence of data regarding the ethnicity of the studied population. A larger sample size, including lower-risk patients, could enhance understanding of thrombus formation mechanisms. Additionally, incorporating external databases to validate the identified phenogroups is essential. While preliminary internal analyses highlighted the robustness and generalization capability of the model, further external validation is necessary to confirm these findings. Moreover, LA size may be influenced by variations in CT scan acquisition timing as well as by variability introduced during segmentation and preprocessing. Finally, the clinical features had missing values, some up to 40%, which led to the use of data imputation techniques with the potential introduction of bias. Additionally, more sophisticated mechanistic models of left atrial hemodynamics are available in the literature¹⁵, which should be benchmarked against our approach, although the high computational cost associated to more complex models would prevent to process large databases.

Due to limited availability of clinical data, we were unable to incorporate fluid boundary conditions, such as PV pressure profiles, MV flow velocity or LA wall motion, in a patient-specific manner. While this limitation allowed us to focus on studying hemodynamic patterns driven solely by the geometry of the left atrium, we aim to integrate patient-specific boundary conditions into the mechanistic model in the future. This would enable us to account for individual variations in blood flow, enhancing the assessment of thrombus risk⁵⁷.

As future work, we aim to validate the phenogroup clusters obtained with larger and external patient cohorts, and explore methods to incorporate categorical variables into the analysis. Moreover, mechanistic models need to follow strict verification and validation guidelines, such as the V&V40⁵⁸, to establish the required credibility for using the models to support clinical decisions. In this context, performing sensitivity analysis studies to assess the impact of CFD modelling configurations on hemodynamic indices associated to thrombus risk^{57,59,60} is crucial and will be prioritized as part of future work. Further investigations are needed to elucidate the morphological parameters that influence in-silico hemodynamic indices, reducing the large computational requirements of fluid simulations. To achieve this, we aim to develop synthetic LA geometry data that enables a systematic exploration of morphology (i.e., ostium dimensions) and in-silico fluid hemodynamics. Finally, we aim to analyse the morphological, hemodynamic and clinical markers in patient cohorts from follow-up studies, to develop thrombus risk prediction models.

This descriptive study underscores the importance of integrating clinical, morphological and hemodynamic data to gain deeper insights into the combination of factors linked to thrombus formation in severely diseased patients. It also highlights the potential of unsupervised machine learning techniques in the analysis of thrombus formation, by facilitating multi-domain data integration and phenogroup exploration. Identifying individuals at higher risk of thrombus formation (i.e., high risk phenotypes) before the LAAO intervention can support clinical decision-making, by implementing closer patient monitoring and guiding the personalization of anticoagulation therapy, particularly given the contraindications to oral anticoagulants in this population. Morphological and hemodynamic information is crucial in the study of thrombus formation. Our initial results show that certain morphological factors (low ostium dimensions, atrial remodelling and LAA tortuosity), including PV configuration (low alignment of LAA to LSPV) influence LA hemodynamics by reducing LAA stasis, and could potentially be protective in patients with pro-thrombotic clinical phenogroups. Further work should be performed on exploiting the presented multi-domain data integration pipeline to develop novel thrombus risk scores or predictive models.

Methods

Ethics statement

This study was approved by the Comité de Protection des Personnes Sud-Ouest et Outre-mer III ethical committee. Written informed consent was obtained from all patients, and all experiments were conducted in accordance with relevant guidelines and regulations.

Study population

The clinical data was provided by Hospital Haut-Lévêque (Bordeaux, France), including high-quality pre-procedural computed tomography (CT) scans and medical records from 130 non-valvular AF patients scheduled for a left atrial appendage occlusion (LAAO) procedure. The CT scans were acquired on a 64-slice dual source CT system (Siemens Definition, Siemens Medical Systems, Forchheim, Germany). The image acquisition was performed using a bi-phasic injection protocol: 1 mL/kg of Iomeprol 350 mg/mL (Bracco, Milan, Italy) at the rate of 5 mL/s, followed by a 1 mL/kg flush of saline at the same rate. The provided CT images had pixel matrix dimensions of 512 x 512 and ranged from 169 to 979 slices, with an approximate pixel spacing of 0.5 x 0.5 x 0.5 mm. The CT scans were acquired at different phases of the cardiac cycle, with scans performed at end-diastole ($n = 67$) and end-systole ($n = 63$). In patients with non-paroxysmal or permanent AF ($n = 63$), an arrhythmogenic rhythm may be assumed during CT acquisition.

The population of study was characterized by a high risk of thromboembolic events (mean CHA₂DS₂-VASc score of 4, $n = 108$). History of stroke/TIA or LAA thrombus detected on the pre-LAAO CT was observed in 58 out of 130 patients (44.6%). Most patients had a history of bleeding (92 out of 102, 90.2%), the majority under administration of oral anticoagulants (69 out of 92, 75%). Due to contraindications to oral anticoagulants, the anti-coagulation therapy was short-termed, with stroke/TIA generally occurring after its cessation. See Supplementary Tables 3–5 for a clinical, anatomical, and blood flow baseline description of the population.

Clinical data and laboratory markers

History of thrombus, encompassing history of ischemic stroke/TIA, and/or history of peripheral embolism and/or LAA thrombus detected in the pre-procedural CT, was the clinical adverse label used to stratify the population. Baseline characteristics involving age, body mass index (BMI), gender, hypertension (HT), diabetes type 2, vascular disease, dyslipidemia, type of atrial fibrillation (categorized in paroxysmal vs non-paroxysmal) and HAS-BLED score was extracted from electronic health records. Echocardiography data such as left ventricular ejection fraction (LVEF) was provided. Heart failure was defined as presence of heart failure diagnosis (obtained from patient medical record) or LV systolic function severely ($\leq 40\%$ LVEF) or moderately (41–49% LVEF) reduced. Laboratory data including C-reactive protein (CRP), B-type natriuretic peptide (BNP), hematocrit levels, creatinine levels, glomerular filtration rate (GFR), urea, ratio of neutrophil/lymphocyte levels, mean platelet volume (MPV), platelet count (PC), troponin and fibrinogen levels were collected from pre-procedural electronic health records.

Image analysis and morphological parameters

For the automatic extraction of LA and LAA geometrical parameters, an in-house computational pipeline was employed. The algorithm computes patient-specific LA and LAA morphological parameters from a CT chest scan. Table 3 – 2nd row shows the extracted shape-related parameters, the most important ones described in the following. For instance, the ostium irregularity measures the deviation of the ostium contour from a circle or ellipsis; in the range [0,1], an index of 1 represents an asymmetrical ostium. Additionally, the LAA tortuosity measures the LAA bending and shape irregularities, and was computed following the definition in Mill et al.¹⁷; index values closer to 1 indicate a higher LAA complexity. The LA sphericity index quantifies the ‘balloon-like’ appearance of the LA, similar to the definition in Moon et al.⁶¹; higher indices (closer to 1) indicates the LA is more similar to a sphere. The morphological parameter extraction algorithm also enabled the extraction of PV-based parameters, including the PV diameter and orientation angle relative to the mitral valve plane, to analyze the steepness of the PV entrance into the LA. The alignment between the LAA and LSPV, defined as the variation of LSPV direction with respect to the LAA from a frontal view, was measured by a 2D angle in all patients. Classification labels of chicken-wing (CW) morphologies, and LAA and LSPV level of alignment in non-CW morphologies were provided by an

expert following a qualitative examination on CT, including high, middle and low grade alignment. The distinction between CW and non-CW morphologies was made based on the clinical relevance of CW in thrombus risk¹⁰.

Construction of left atrial digital twins and mechanistic models of hemodynamics

The LA geometry was extracted from the CT images using the semi-automatic region-growing segmentation tools available in Slicer 4.10.112, enabling the generation of a binary mask for the LA. Geometrical surface meshes were then obtained through the application of the Marching Cubes algorithm. Mesh post-processing involved smoothing and the correction of mesh irregularities. Additionally, cuts were made on the pulmonary veins (prior to the first bifurcation) and the mitral valve. Subsequently, volumetric meshes were generated and optimized with the Netgen software⁶². The final mesh resolution comprised from 7 to 9 × 10⁵ elements, subject to the LA volume. For further details on the segmentation and mesh processing, refer to Supplementary Information-Section I.1.

The computational fluid dynamic simulations were run with the ANSYS Fluent Solver 19.2 (ANSYS Inc, United States) to solve the Navier-Stokes equations, utilizing the finite volume method. Blood was considered as an incompressible Newtonian fluid, with a density of 1060 kg/m³ and a dynamic viscosity of 0.0035 Pa · s. Each simulation encompassed three cardiac cycles, the first two cycles served as the warm-up stage and the third cycle underwent the analysis. The time step was configured to 0.01 seconds, and to ensure absolute convergence, the residuals for the continuity equations were set at 0.001. The pipeline presented in Mill et al.¹⁷ was followed.

Two distinct boundary conditions (BC) were applied to all simulation models: (1) pressure inlet at the PVs – the pressure curve profile was obtained through catheterization of an AF patient in sinus rhythm; and (2) blood flow velocity outlet at the MV – the velocity curve was derived from an Doppler echocardiography acquisition of the studied population. The corresponding electrocardiography data was employed to determine the duration of each cardiac cycle. Additionally, LA movement was incorporated as a passive response to the longitudinal motion of the mitral valve annulus ring, as defined by Veronesi et al.⁶³, based on a spring-based dynamic mesh approach.

It must be noted that the BCs were not tailored to individual patients due to the limited availability of echocardiographic patient-specific data. The ultimate goal is to integrate as much patient-specific data as possible for the construction of digital twins; however, the maintained homogeneity of the BCs also allowed the study of complex blood flow patterns purely as a response to LA morphology. The BC profiles can be found in Supplementary Information-Section I.2.

Particle-based in-silico indices were obtained from the CFD models to provide insights into LAA stasis, as described by Mill et al.¹⁷. Massless particles were released at the pulmonary veins and their trajectories followed the simulated velocity vector fields originating from their respective seed points. These particles were tracked throughout the cardiac cycle. At each time step, 100 particles were injected, with 50 particles introduced from each side of the LA, evenly distributed across each pulmonary vein. Two hemodynamic indices were computed: the total number of fluid particles remaining in the LAA at the end of the cardiac cycle (final time step) and the particle age, defined as the duration the particle remains within the LA.

Phenogroup analysis

In total, 47 clinical, morphological and in-silico blood flow variables were extracted (Table 3). Among these, 12 clinical variables had missing values (Supplementary Table 1). To impute the missing values, the Random Forest for mixed-type data⁶⁴ ('missForest' package v1.4 in R) was employed. Regarding the ML input feature selection (Table 3, 2nd column), clinical variables with more than 40% missing values in the raw data were not considered for ML input. The missing data threshold was deemed acceptable after a review of previous literature^{65–67}, although the potential bias introduced by missing data imputation was evaluated (see Supplementary

Information-Section I.3 for further details). Shape-related features were selected to characterize the ostium, LAA and LA dimensions and shape, as well as the spatial relation of LAA and LSPV. Only continuous variables were included as ML input to reduce feature weight imbalances in the MKL algorithm. Moreover, variables with high bi-variate Pearson correlations (> 0.7, Supplementary Fig. 5) were excluded. Furthermore, variables not considered for ML input (Table 3, 3rd column) were analyzed after clustering to fully characterize the resulting phenogroups. The in-silico index particle age did not yield significant findings related to thrombus formation in previous studies¹⁷ and was therefore not considered as ML input. For a detailed overview on the missing data imputation and input feature selection process, refer to Supplementary Table 2.

The unsupervised Multiple Kernel Learning (MKL) algorithm⁶⁸ was used. This validated machine learning technique has been applied previously to explore cardiac disease in patients^{38,69}. MKL uses multiple kernels, each associated with a specific feature, to integrate heterogeneous features into a low-dimensional space. In this transformed space, patients are positioned based on the resemblance of their input features. For this purpose, each feature-based kernel was implemented as a Gaussian kernel, with a bandwidth set to $\sqrt{N}$ (where N represents the total number of samples), to compute the average distance over the $\sqrt{N}$ closest subjects. Since all input parameters were continuous, the Euclidean distance was used for all kernels.

Consequently, individuals sharing similar characteristics are located closer together, enhancing interpretability and facilitating clustering analyses. K-means++ was selected as the clustering algorithm, characterized as an easily interpretable approach with an efficient centroid initialization. To increase its robustness, each run involved 20 iterations with different initial conditions. In total, six cluster analysis were performed with different combinations of clinical, shape-related and circulatory input features, to evaluate the added value of each class of parameters. For each analysis, the selection of the number of clusters and the number of clustering dimensions was driven by the highest average silhouette score of K-means++ and a hierarchical clustering algorithm. The average silhouette score was computed for different combinations of number of clusters and number of output space dimensions, in a grid search approach. The number of clusters considered ranged from 3 to 4, balancing the analysis of multiple risk groups with the constraints of sample size. The number of dimensions assessed ranged from 2 to 5. Different input feature domains combinations were considered for MKL dimensionality reduction and clustering analysis (see Table 4).

To identify the most important attributes determining cluster membership, a Random Forest (RF) classifier was trained with the cluster labels as outcomes. The SHapley Additive exPlanations (SHAP) method⁴⁴ ranked attribute importance for each cluster and assessed the impact of individual attribute values on cluster assignment. Additionally, to improve the interpretability of the phenogroups, a RF classifier was trained to predict 'history of thrombus' using the most important features identified by the unsupervised clustering model. For further details on the feature importance analysis and the comparison with supervised learning methods, refer to Supplementary Information-Section I.4 and I.5, respectively.

We assessed the robustness of the MKL and clustering solution by comparing phenogroups from the full dataset model with those from different training subsets. For this, the dataset was split into four folds, with one fold held out for projection and the remaining three used for training the MKL algorithm and generating clusters. This process was repeated iteratively. We also checked the stability of the K-means clustering through a bootstrapping approach ('bootcluster' package in R). Finally, an internal validation was conducted to evaluate the generalizability of our findings. The dimensionality reduction MKL algorithm was first trained and fine-tuned on a randomly selected 75% of the dataset, followed by the implementation of the clustering algorithm. The remaining unseen data was then projected and classified into phenogroups. Patient characteristics across phenogroups were then analyzed and compared between the training and validation sets. For further details on the internal evaluation of the model, refer to Supplementary Information-Section I.6.

Table 4 | Multiple Kernel Learning dimensionality reduction performed with different combinations of clinical, morphological and in-silico hemodynamic input features

Analysis	ML domain combinations
1	Clinical
2	Morphological
3	Morphological + clinical
4	In-silico hemodynamic + clinical
5	Morphological + in-silico hemodynamic
6	Morphological + in-silico hemodynamic + clinical

ML machine learning.

We conducted a sensitivity analysis to evaluate whether adding clinical variables to anatomical and blood flow data would alter patient placement in the output space and affect the resulting phenogroups. First, we examined the nearest neighbors of each patient in the output spaces before and after including clinical variables and compared their similarities. Next, we analyzed the characteristics of patients reclassified into a different phenogroup, particularly those with a different proportion of thrombus history. Further details are available in Supplementary Information-Section I.7.

Statistics

Continuous variables were tested for normality using the Shapiro-Wilk test and for homoscedasticity using the Levene test. If the null hypothesis was rejected at a significance level of $\alpha = 0.05$, indicating non-normality or heteroscedasticity, the non-parametric Wilcoxon Mann-Whitney test was used for univariate analysis, and variables were reported as median and interquartile range. For parametric distributions, the Student’s t-test was applied, with results presented as mean and standard deviation. Differences between clusters were analyzed using ANOVA (parametric) or the Kruskal-Wallis test (non-parametric). Post-hoc pairwise comparisons between clusters were conducted using Dunn’s test with Bonferroni corrections (non-parametric) or Tukey’s HSD test (parametric). Categorical variables were analyzed with the chi-square test, applying FDR correction for multiple comparisons. A p-value of less than 0.05 was considered statistically significant. All analyses were performed using R.

Data availability

The data supporting the findings of this study is available upon request from CHU Bordeaux-IHU Lyric. However, due to restrictions related to patient confidentiality and data licensing agreements, it is not publicly accessible. Interested parties seeking access may contact the IHU-Lyric department at dpo@ihu-lyric.fr or Xavier Bouteiller, Data Protection Officer (DPO) and Biostatistician - Data Manager at CHU Bordeaux (Cardiology Department, Rhythmology Unit), via email at xavier.bouteiller@chu-bordeaux.fr for further inquiries.

Code availability

The code underlying this study is not publicly accessible but may be provided to qualified researchers upon reasonable request to the corresponding author.

Abbreviations

AF	atrial fibrillation
BC	boundary conditions
BMI	body mass index
BNP	B-type natriuretic peptide
CFD	computational fluid dynamics
CRP	C-reactive protein
CT	computed tomography
CW	chicken-wing

DT	digital twin
GFR	glomerular filtration rate
HF	heart failure
HT	hypertension
LAA	left atrial appendage
LAEO	left atrial appendage occlusion
LA	left atrium
LIPV	left inferior pulmonary vein
LSPV	left superior pulmonary vein
LVEF	left ventricular ejection fraction
MKL	multiple kernel learning
ML	machine learning
MPV	mean platelet volume
MV	mitral valve
PC	platelet count
PV	pulmonary vein
RIPV	right inferior pulmonary vein
RSPV	right superior pulmonary vein
SHAP	Shapley Additive exPlanations

Received: 4 October 2024; Accepted: 24 April 2025;
Published online: 17 June 2025

References

1. Feigin, V. L. et al. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019. *Lancet Neurol.* **20**, 795–820 (2021).
2. Cresti, A. et al. Prevalence of extra-appendage thrombosis in non-valvular atrial fibrillation and atrial flutter in patients undergoing cardioversion: a large transoesophageal echo study. *EuroIntervention* **15**, e225–e230 (2019).
3. Watson, T., Shantsila, E. & Lip, G. Y. Mechanisms of thrombogenesis in atrial fibrillation: Virchow’s triad revisited. *Lancet* **373**, 155–166 (2009).
4. Escudero-Martinez, I., Morales-Caba, L. & Segura, T. Atrial fibrillation and stroke: A review and new insights. *Trends Cardiovascular Med.* **33**, 23–29 (2023).
5. Zingaro, A. et al. A comprehensive stroke risk assessment by combining atrial computational fluid dynamics simulations and functional patient data. *Sci. Rep.* **14**, 9515 (2024).
6. Huang, K., Wen, H. & Liu, C. Association between ischemic stroke and left atrial appendage morphology in patients with atrial fibrillation and low cha2ds2-vasc scores. *Tex. Heart Inst. J.* **49**, e207544 (2022).
7. Segan, L. et al. Identifying patients at high risk of left atrial appendage thrombus before cardioversion: the clots-af score. *J. Am. Heart Assoc.* **12**, e029259 (2023).
8. Sun, Y. et al. Finding low cha2ds2-vasc scores unreliable? why not give morphological and hemodynamic methods a try? *Front. Cardiovascular Med.* **9**, 1032736 (2023).
9. Bisbal, F. et al. Left atrial sphericity: a new method to assess atrial remodeling. impact on the outcome of atrial fibrillation ablation. *J. cardiovascular Electrophysiol.* **24**, 752–759 (2013).
10. Di Biase, L. et al. Does the left atrial appendage morphology correlate with the risk of stroke in patients with atrial fibrillation? results from a multicenter study. *J. Am. Coll. Cardiol.* **60**, 531–538 (2012).
11. Genua, I. et al. Centreline-based shape descriptors of the left atrial appendage in relation with thrombus formation. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers* 9, 200–208 (Springer, 2019).
12. Corral-Acero, J. et al. The -digital twin- to enable the vision of precision cardiology. *Eur. heart J.* **41**, 4556–4564 (2020).

13. Sel, K. et al. Building digital twins for cardiovascular health: From principles to clinical impact. *J. Am. Heart Assoc.* **13**, e031981 (2024).
14. Aguado, A. M. et al. In silico optimization of left atrial appendage occluder implantation using interactive and modeling tools. *Front. Physiol.* **10**, 237 (2019).
15. Guerrero-Hurtado, M. et al. Hemodynamics affects factor xi/xii anticoagulation efficacy in patient-specific left atrial models. *bioRxiv* 2024–08 (2024).
16. Albers, C. et al. Impact of occluder device configurations in in-silico left atrial hemodynamics for the analysis of device-related thrombus. *PLOS Computational Biol.* **20**, e1011546 (2024).
17. Mill, J. et al. The role of the pulmonary veins on left atrial flow patterns and thrombus formation. *Sci. Rep.* **14**, 5860 (2024).
18. García-Villalba, M. et al. Demonstration of patient-specific simulations to assess left atrial appendage thrombogenesis risk. *Front. Physiol.* **12**, 596596 (2021).
19. Quarteroni, A. & Formaggia, L. Mathematical modelling and numerical simulation of the cardiovascular system. *Handb. Numer. Anal.* **12**, 3–127 (2004).
20. Masci, A. et al. The impact of left atrium appendage morphology on stroke risk assessment in atrial fibrillation: a computational fluid dynamics study. *Front. Physiol.* **9**, 417579 (2019).
21. Dueñas-Pamplona, J. et al. Reduced-order models of wall shear stress patterns in the left atrial appendage from a data-augmented atrial database. *Appl. Math. Model.* **130**, 713–727 (2024).
22. García-Isla, G. et al. Sensitivity analysis of geometrical parameters to study haemodynamics and thrombus formation in the left atrial appendage. *Int. J. Numer. Methods Biomed. Eng.* **34**, e3100 (2018).
23. Qureshi, A. et al. Imaging and biophysical modelling of thrombogenic mechanisms in atrial fibrillation and stroke. *Front. Cardiovasc. Med.* **9**, 1074562 (2023).
24. Olesen, J. B. et al. Stroke and bleeding in atrial fibrillation with chronic kidney disease. *N. Engl. J. Med.* **367**, 625–635 (2012).
25. Liu, X., Yu, S., Liang, T., Chen, L. & Zhang, H. Mean platelet volume to platelet count ratio predicts left atrial stasis in patients with non-valvular atrial fibrillation. *Int. J. Gen. Med.* **16**, 847–858 (2023).
26. Black, I. W. et al. Hematologic correlates of left atrial spontaneous echo contrast and thromboembolism in nonvalvular atrial fibrillation. *J. Am. Coll. Cardiol.* **21**, 451–457 (1993).
27. Pant, R. et al. Impact of b-type natriuretic peptide level on the risk of left atrial appendage thrombus in patients with nonvalvular atrial fibrillation: a prospective study. *Cardiovascular Ultrasound* **14**, 1–7 (2015).
28. Cianfrocca, C. et al. C-reactive protein and left atrial appendage velocity are independent determinants of the risk of thrombogenesis in patients with atrial fibrillation. *Int. J. Cardiol.* **142**, 22–28 (2010).
29. Yalcin, M. et al. Neutrophil-lymphocyte ratio may predict left atrial thrombus in patients with nonvalvular atrial fibrillation. *Clin. Appl. Thrombosis/Hemost.* **21**, 166–171 (2015).
30. Shah, D., Patel, S. & Bharti, S. K. Heart disease prediction using machine learning techniques. *SN Computer Sci.* **1**, 345 (2020).
31. Chekroud, A. M. et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* **3**, 243–250 (2016).
32. Zhao, Y. et al. Medical record data-enabled machine learning can enhance prediction of left atrial appendage thrombosis in nonvalvular atrial fibrillation. *Thrombosis Res.* **223**, 174–183 (2023).
33. Pieszko, K. et al. Artificial intelligence in detecting left atrial appendage thrombus by transthoracic echocardiography and clinical features: the left atrial thrombus on transoesophageal echocardiography (lattee) registry. *Eur. Heart J.* **45**, 32–41 (2024).
34. Pons, M. I. et al. Joint analysis of morphological parameters and in silico haemodynamics of the left atrial appendage for thrombogenic risk assessment. *J. Interv. Cardiol.* **2022**, 9125224 (2022).
35. Kwon, D. H. et al. Cardiac mri-enriched phenomapping classification and differential treatment outcomes in patients with ischemic cardiomyopathy. *Circulation: Cardiovasc. Imaging* **17**, e016006 (2024).
36. Verdonschot, J. A. et al. Phenotypic clustering of dilated cardiomyopathy patients highlights important pathophysiological differences. *Eur. Heart J.* **42**, 162–174 (2021).
37. Loncaric, F. et al. Automated pattern recognition in whole-cardiac cycle echocardiographic data: Capturing functional phenotypes with machine learning. *J. Am. Soc. Echocardiogr.* **34**, 1170–1183 (2021).
38. Garcia-Canadilla, P. et al. Machine-learning-based exploration to identify remodeling patterns associated with death or heart-transplant in pediatric-dilated cardiomyopathy. *J. Heart Lung Transplant.* **41**, 516–526 (2022).
39. Inohara, T. et al. Association of atrial fibrillation clinical phenotypes with treatment patterns and outcomes: a multicenter registry study. *JAMA Cardiol.* **3**, 54–63 (2018).
40. Ogawa, H. et al. Characteristics and clinical outcomes in atrial fibrillation patients classified using cluster analysis: the fushimi af registry. *EP Europace* **23**, 1369–1379 (2021).
41. Saito, Y. et al. Phenotyping of atrial fibrillation with cluster analysis and external validation. *Heart* **109**, 1751–1758 (2023).
42. Romiti, G. F. et al. Patterns of comorbidities in patients with atrial fibrillation and impact on management and long-term prognosis: an analysis from the prospective global gloria-af registry. *BMC Med.* **22**, 151 (2024).
43. Saiz-Vivó, M. et al. Unsupervised machine learning exploration of morphological and haemodynamic indices to predict thrombus formation in the left atrial appendage. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, 200–210 (Springer, 2022).
44. Lundberg, S. A unified approach to interpreting model predictions. In *Proceedings Advances in Neural Information Processing Systems* (ed. Guyon, I. et al.) https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf (Curran Associates, Inc., 2017).
45. Yu, H. et al. Bootstrapping estimates of stability for clusters, observations and model selection. *Comput. Stat.* **34**, 349–372 (2019).
46. Levy, S. & Sbragia, P. Remodelling in atrial fibrillation. *Arch. des. Maladies du Coeur et. des. Vaiss.* **98**, 308–312 (2005).
47. Shirani, J. & Alaeddini, J. Structural remodeling of the left atrial appendage in patients with chronic non-valvular atrial fibrillation: implications for thrombus formation, systemic embolism, and assessment by transesophageal echocardiography. *Cardiovasc. Pathol.* **9**, 95–101 (2000).
48. Liu, H. et al. The risk factors of thrombus formation and the effect of catheter ablation on repetitive thrombus formation in patients with atrial fibrillation: a single center retrospective study in china. *BMC Cardiovasc. Disord.* **23**, 28 (2023).
49. Yamamoto, M. et al. Complex left atrial appendage morphology and left atrial appendage thrombus formation in patients with atrial fibrillation. *Circulation: Cardiovascular Imaging* **7**, 337–343 (2014).
50. Al Housari, M. et al. Pulmonary vein remodeling between atrial fibrillation subtypes: A matched comparison cardiac computed tomography-based study between patients with paroxysmal and persistent atrial fibrillation. *Am. J. Cardiol.* **207**, 100–107 (2023).
51. Park, J. et al. High left atrial pressures are associated with advanced electroanatomical remodeling of left atrium and independent predictors for clinical recurrence of atrial fibrillation after catheter ablation. *Heart rhythm* **11**, 953–960 (2014).
52. Kerr, B. & Brandon, L. Atrial fibrillation, thromboembolic risk, and the potential role of the natriuretic peptides, a focus on bnp and nt-probnp—a narrative review. *IJC Heart Vasculture* **43**, 101132 (2022).
53. Harkness, A. et al. Normal reference intervals for cardiac dimensions and function for use in echocardiographic practice: a guideline from the british society of echocardiography. *Echo Res. Pract.* **7**, G1–G18 (2020).

54. Choi, D.-H., Kang, S.-H. & Song, H. Mean platelet volume: a potential biomarker of the risk and prognosis of heart disease. *Korean J. Intern. Med.* **31**, 1009 (2016).
55. Providência, R. et al. Mean platelet volume is associated with the presence of left atrial stasis in patients with non-valvular atrial fibrillation. *BMC cardiovascular Disord.* **13**, 1–8 (2013).
56. Chen, Y.-Y. et al. Effect of metabolic syndrome on risk stratification for left atrial or left atrial appendage thrombus formation in patients with nonvalvular atrial fibrillation. *Chin. Med. J.* **129**, 2395–2402 (2016).
57. Mill, J. et al. Sensitivity analysis of in silico fluid simulations to predict thrombus formation after left atrial appendage occlusion. *Mathematics* **9**, 2304 (2021).
58. Viceconti, M. et al. In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. *Methods* **185**, 120–127 (2021).
59. Kjeldsberg, H. A. et al. Impact of left atrial wall motion assumptions in fluid simulations on proposed predictors of thrombus formation. *Int. J. Numer. Method. Biomed. Eng.* **40**, e3825 (2024).
60. Khalili, E. et al. On the importance of fundamental computational fluid dynamics toward a robust and reliable model of leftatrial flows. *Int. J. Numer. Method. Biomed. Eng.* **40**, e3804 (2024).
61. Moon, J. et al. Prognostic implication of left atrial sphericity in atrial fibrillation patients undergoing radiofrequency catheter ablation. *Pacing Clin. Electrophysiol.* **40**, 713–720 (2017).
62. Schöberl, J. Netgen an advancing front 2d/3d-mesh generator based on abstract rules. *Comput. Vis. Sci.* **1**, 41–52 (1997).
63. Veronesi, F. et al. Quantification of mitral apparatus dynamics in functional and ischemic mitral regurgitation using real-time 3-dimensional echocardiography. *J. Am. Soc. Echocardiogr.* **21**, 347–354 (2008).
64. Stekhoven, D. J. & Bühlmann, P. Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
65. Penone, C. et al. Imputation of missing data in life-history trait datasets: which approach performs the best? *Methods Ecol. Evolution* **5**, 961–970 (2014).
66. Pimor, A. et al. Predictors of post-operative cardiovascular events, focused on atrial fibrillation, after valve surgery for primary mitral regurgitation. *Eur. Heart J.-Cardiovascular Imaging* **20**, 177–184 (2019).
67. Zhang, S. et al. Imputation of gps coordinate time series using missforest. *Remote Sens.* **13**, 2312 (2021).
68. Sanchez-Martinez, S. et al. Characterization of myocardial motion patterns by unsupervised multiple kernel learning. *Med. image Anal.* **35**, 70–82 (2017).
69. Cikes, M. et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *Eur. J. heart Fail.* **21**, 74–85 (2019).

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the grant agreements No 101016496 (SimCardioTest) and No 101136438 (GEMINI), as well as by the Spanish Ministry of Science, Innovation and Universities under the GENERALITAT grant, (PID2022-143239OB-I00). GP's contribution was supported by the ICREA Academia program.

Author contributions

M.S.V. and O.C. designed the research and wrote the manuscript. X.I. and H.C. collected the data. M.S.V., J.M., X.I., H.C., G.P., M.S. and O.C. did the research and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-01676-1>.

Correspondence and requests for materials should be addressed to Marta Saiz-Vivó.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025