Medical Image Analysis 102 (2025) 103538

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Graph-based prototype inverse-projection for identifying cortical sulcal pattern abnormalities in congenital heart disease

Hyeokjin Kwon [a,b,d,e], Seungyeon Son [c], Sarah U. Morton [d,e], David Wypij [e,f,g], John Cleveland [h], Caitlin K Rollins [i], Hao Huang [j], Elizabeth Goldmuntz [k], Ashok Panigrahy [l], Nina H. Thomas [m,n], Wendy K. Chung [o], Evdokia Anagnostou [p], Ami Norris-Brilliant [q], Bruce D. Gelb [r], Patrick McQuillen [s], George A. Porter Jr. [t], Martin Tristani-Firouzi [u], Mark W. Russell [v], Amy E. Roberts [b,e,d,w], Jane W. Newburger [e,g], P. Ellen Grant [b,d,e,x], Jong-Min Lee [a,c,y,1,*], Kiho Im [b,d,e,1,**]

[a] Department of Electronic Engineering, Hanyang University, Seoul, South Korea
[b] Fetal Neonatal Neuroimaging and Developmental Science Center, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA
[c] Department of Artificial Intelligence, Hanyang University, Seoul, South Korea
[d] Division of Newborn Medicine, Boston Children's Hospital, Boston, MA, USA
[e] Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[f] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[g] Department of Cardiology, Boston Children's Hospital, Boston, MA, USA
[h] Department of Surgery and Department of Pediatrics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[i] Department of Neurology, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA
[j] Department of Radiology, Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA
[k] Division of Cardiology, Department of Pediatrics, Children's Hospital of Philadelphia, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[l] Department of Pediatric Radiology, Children's Hospital of Pittsburgh, University of Pittsburgh Medical Center, Pittsburgh, PA, USA
[m] Department of Child and Adolescent Psychiatry and Behavioral Sciences and Center for Human Phenomic Science, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[n] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA
[o] Department of Pediatrics and Department of Medicine, Columbia University Medical Center, New York, NY, USA
[p] Department of Pediatrics, Holland Bloorview Kids Rehabilitation Hospital, University of Toronto, Toronto, Ontario, Canada
[q] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[r] Mindich Child Health and Development Institute and Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[s] Department of Pediatrics and Department of Neurology, University of California, San Francisco, CA, USA
[t] Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA
[u] Division of Pediatric Cardiology, University of Utah School of Medicine, Salt Lake City, UT, USA
[v] Department of Pediatrics, C.S. Mott Children's Hospital, University of Michigan, Ann Arbor, MI, USA
[w] Department of Pediatrics, Boston Children's Hospital, Boston, MA, USA
[x] Department of Radiology, Harvard Medical School, Boston, MA, USA
[y] Department of Biomedical Engineering, Hanyang University, Seoul, South Korea

ARTICLE INFO

Keywords:
Congenital heart disease
Graph neural networks
Self-explainable model
Sulcal pattern analysis

ABSTRACT

Examining the altered arrangement and patterning of sulcal folds offers insights into the mechanisms of neurodevelopmental differences in psychiatric and neurological disorders. Previous sulcal pattern analysis used spectral graph matching of sulcal pit-based graph structures to assess deviations from normative sulcal patterns. However, challenges exist, including the absence of a standard criterion for defining a typical reference set, time-consuming cost of graph matching, user-defined feature weight sets, and assumptions about uniform node distribution. We developed a deep learning-based sulcal pattern analysis to address these challenges by adapting

* Corresponding author at: Department of Electronic Engineering, Hanyang University, Seoul 04763, South Korea.
** Corresponding author at: Fetal Neonatal Neuroimaging and Developmental Science Center, Boston Children's Hospital and Harvard Medical School, 300 Longwood Ave, Boston, MA 02115, USA.
E-mail addresses: ljm@hanyang.ac.kr (J.-M. Lee), Kiho.Im@childrens.harvard.edu (K. Im).
[1] These authors are joint supervising authors.

https://doi.org/10.1016/j.media.2025.103538
Received 26 July 2024; Received in revised form 22 February 2025; Accepted 27 February 2025
Available online 28 February 2025

prototype-based graph neural networks to sulcal pattern graphs. Additionally, we proposed a prototype inverse-projection for better interpretability. Unlike other prototype-based models, our approach inversely projects prototypes onto individual node representations to calculate the inverse-projection weights, enabling efficient visualization of prototypes and focusing the model on selective regions. We evaluated our method through a classification task between healthy controls ($n = 174$, age $= 15.4 \pm 1.9$ [mean $\pm$ standard deviation, years]) and patients with congenital heart disease ($n = 345$, age $= 15.8 \pm 4.7$) from four cohort studies and a public dataset. Our approach demonstrated superior classification performance compared to other state-of-the-art models, supported by extensive ablative studies. Furthermore, we visualized and examined the learned prototypes to enhance understanding. We believe our method has the potential to be a sensitive and understandable tool for sulcal pattern analysis.

1. Introduction

Sulcal patterning refers to the global patterning of the position, arrangement, number, and size of sulcal folds and their intersulcal relationships. Sulcal patterning strongly regulates the spatiotemporal dynamics of cortical expansion and folding (Im et al., 2017). Primary sulcal folds exhibit interrelated developmental patterns, reflecting the optimal organization of functional cortical areas under tight genetic control. Disruptions in the global patterning of primary sulci occur due to alterations to neurodevelopmental processes in various brain malformations and psychiatric and neurological disorders (Barkovich et al., 2012; Im et al., 2013b, 2016; Nakamura et al., 2007). Therefore, regional features of sulcal folds and their inter-sulcal relationships should be considered when identifying abnormal sulcal patterns (Im et al., 2017, 2013b; Ortinau et al., 2019). However, visual detection of changes in the interrelated arrangement and patterning of sulcal folds is challenging, necessitating quantitative sulcal pattern analysis (SPA) to analyze neurodevelopmental disabilities in the human cerebral cortex (Im et al., 2013b; Tarui et al., 2023).

We have developed a comprehensive quantitative method to analyze sulcal patterns using a graph structure with primary sulcal pits (the deepest local points of sulci) as nodes derived from magnetic resonance imaging (MRI) data (Im et al., 2017, 2011, 2013b; Morton et al., 2020, 2021; Ortinau et al., 2019; Tarui et al., 2018) (Fig. 1). The mean similarity index between individual sulcal pattern graphs and a set of sulcal graphs from healthy controls was calculated using spectral graph matching to measure deviations from typical sulcal patterns. Previous studies have demonstrated the sensitivity of this technique in detecting abnormal sulcal patterning and altered brain development across various disorders, including congenital heart disease (CHD) (Asschenfeldt et al., 2021; Maleyeff et al., 2024a; Morton et al., 2020, 2021; Ortinau et al., 2019), developmental dyslexia (Im et al., 2016),

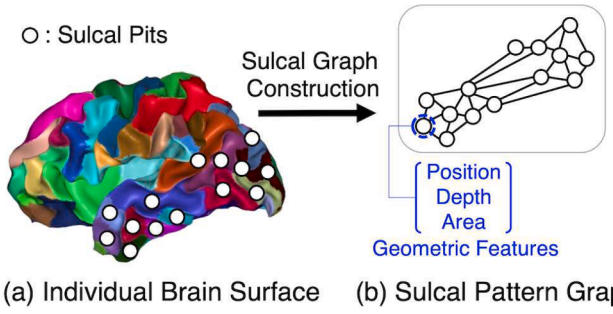

Fig. 1. Illustration of the construction of the sulcal pattern graph. For each cortical surface (a), we constructed a sulcal pattern graph (b), wherein each sulcal pit (a white circle) served as a node in the graph. Sulcal pits (graph nodes) were connected by an edge when their sulcal basins met, and we utilized geometric features such as the position and depth of the sulcal pits and the area of the corresponding sulcal basin to represent the graph nodes. In the figures in this paper, sulcal pits in regions other than the left-temporal lobe were omitted for simplicity. It is noteworthy that we utilized whole brain sulcal pits for our analyses.

ventriculomegaly (Tarui et al., 2023), isolated agenesis of the corpus callosum (Tarui et al., 2018; Vasung et al., 2020), and cerebral malformations (Bae et al., 2014; Im et al., 2013b). Despite recent advances, this method has several limitations. First, this method relies on a typical subject set determined by inconsistent criteria, such as a subset of the dataset being analyzed (Im et al., 2011, 2013b), an external typical control cohort (Morton et al., 2020), and template brains (Im et al., 2017; Tarui et al., 2023). Selecting typical subjects may introduce bias or fail to represent healthy controls adequately. While more extensive typical subject sets can mitigate these issues, the spectral matching algorithm presents a challenge to conventional SPA. To calculate sulcal pattern similarity, large affinity matrices $\mathbb{M}_{ij}$ of dimensions $N_iN_j \times N_iN_j$ for each graph $\mathcal{G}_i$ are constructed using all graphs $\mathcal{G}_j$, $\forall\, j = 1, \ldots, |\mathcal{D}_T|$ from the typical subject set $\mathcal{D}_T$, where N_i and N_j denote the number of nodes of the target graph pair $\mathcal{G}_i$, and $\mathcal{G}_j$, respectively. Constructed affinity matrices $\mathbb{M}_{ij}$ are then used to build a subset of consistent assignments guided by their eigenvectors, demanding high computation time and memory resources. These computation requirements impede similarity calculations with a large number of typical subjects, limiting the generalizability of the method. Second, spectral graph matching usually assumes equal contributions across sulcal pits to calculate the similarity index. However, patients with developmental disease may exhibit marked differences in sulcal patterns in specific brain regions (Im et al., 2013b, 2016; Morton et al., 2020, 2021). Third, geometric features for nodes are weighted to define inter-graph node correspondence and relative importance during target graph matching (Im et al., 2011, 2013b). Previous studies repeated the estimation of similarity indices using different weights and statistically compared the similarity indices. However, this approach cannot explicitly ensure optimal weight set selection.

Recently, graph neural networks (GNNs) have garnered attention for addressing complex relationships between graph nodes in various real-world graph tasks (Defferrard et al., 2016; Gilmer et al., 2017; Kawahara et al., 2017). Among these high-performance GNNs, prototype-based models follow a concept-based learning approach where class representative patterns (prototypes) are learned and used during the reasoning process, akin to human decision-making processes (Alvarez Melis and Jaakkola, 2018; Keswani et al., 2022; Rymarczyk et al., 2022; Zheng et al., 2019). Specifically, prototype-specific similarities are computed by comparing the most important part of an individual input graph with class-specific prototypes. These similarity values are then used to inform the probability of the input graph belonging to a specific class for classification. Extensive studies have demonstrated the ability to learn trustworthy explanations without sacrificing predictive performance (Bécigneul et al., 2020; Dai and Wang, 2022; Ragno et al., 2022; Vincent-Cuaz et al., 2022; Zhang et al., 2022).

Interestingly, prototype-based GNNs and our previous graph-based SPA utilize class-representative patterns (prototypes and subject sets) to quantitatively assess the similarity between individual samples and specific class-representative patterns. However, unlike SPA, GNN prototypes are learned from training samples in a data-driven manner, enhancing model generalizability. Prototype-based models enforce

diverse (Rymarczyk et al., 2021; Zhang et al., 2022) and class-representative (Kang et al., 2022; Wang et al., 2021) prototype patterns via various constraints, overcoming issues with existing typical subject sets in SPA. Consequently, prototypes enable larger training graphs without necessitating affinity matrices for all training graphs. Furthermore, input geometric features can be considered using the optimal weight sets through trainable parameter optimization in deep-learning-based models. However, the concerns regarding the unequal contribution of nodes remain. This challenge might be overcome by including additional modules, such as subgraph extraction networks, into GNNs. These modules can extract and consider the most crucial parts (e.g., specific brain regions) from entire graphs, albeit at the cost of computational complexity and time (Zhang et al., 2022). Furthermore, mapping learned prototypes in the latent space using input data enables visualization (Gautam et al., 2022). To achieve this, recent prototype-based models have been applied to various mechanisms, including a prototype decoder (Gautam et al., 2022; Ragno et al., 2022), and prototypes have replaced the nearest part of the training sample during training (Rymarczyk et al., 2022; Zhang et al., 2022). However, these approaches introduce challenges to the training process and increase model complexity.

This study proposes a solution involving a prototype-based GNN mechanism called the prototype inverse-projection (PIP). Specifically, prototypes are inversely projected onto all nodes in individual graphs rather than directly projecting individual graph representations onto learnable prototypes (Fig. 2). The inverse-projection weights linking prototypes and nodes are used to calculate the prototype-specific similarity scores for subsequent classification tasks. This mechanism enables the model to prioritize the most important subgraph while the learned prototypes can be spatially delineated using inverse-projection weights. We evaluated the proposed model by applying the model to sulcal pattern graphs from patients in CHD cohorts to investigate atypical sulcal patterning. CHD is the most common congenital abnormality and is associated with impaired neurogenesis and neurodevelopmental disabilities (Marelli et al., 2016; Mebius et al., 2017; Wernovsky, 2006). Recent MRI-based studies have shown that abnormal neurodevelopment in CHD is associated with region-specific brain vulnerability, including delayed gyrification and atypical sulcal folding patterns (Clouchoux et al., 2013; Olshaker et al., 2018; Ortinau et al., 2019). Thus, we set a binary classification problem (CHD vs. healthy controls) and compared the performance of the proposed model with those of other state-of-the-art GNNs and existing prototype-based GNNs. Extensive ablation studies were performed to explore various configurations of the proposed model. Furthermore, we performed group-level visualization of the learned prototypes by mapping individual prototype-specific inverse-projection weights onto a brain surface model in a common space.

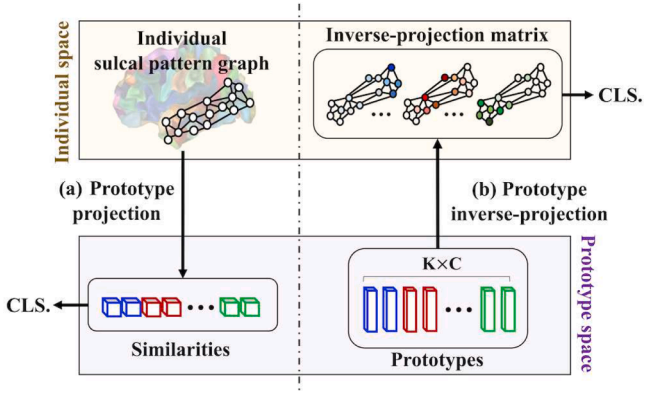

Fig. 2. Graphical representation of the (a) prototype projection and (b) prototype inverse projection (PIP). (a) Prototype projection was used to estimate prototype similarity for each prototype in existing prototype-based models. (b) In contrast, PIP inversely projects the prototypes onto the representations of nodes in individual graphs to identify the most relevant regions (nodes) with the highest inverse projection weights for each prototype. CLS. denotes downstream classification task.

2. Related works

2.1. Quantitative cortical folding pattern analysis

Although the precise mechanisms underlying human-specific sulcal folding patterning have not been fully investigated, it is widely accepted that cortical sulcal patterning is a pattern-specific folding process rather than a random one (Rakic, 1988). To quantify abnormalities of sulcal patterns, pioneer MRI studies have characterized cortical growth and folding development by capturing various structural features such as the gyrification index, mean cortical surface curvature, sulcal depth, sulcal length, and sulcal area (Clouchoux et al., 2013; Lefèvre et al., 2015; Wright et al., 2014; Zilles et al., 1988). For example, Duan et al. constructed multi-view curvature features on inner cortical surfaces to explore representative folding patterns of infant brains (Duan et al., 2019). Guillon et al. utilized 3D skeleton maps of grey matter and the cerebrospinal fluid union to identify rare cortical folding patterns (Guillon et al., 2024). Chamfer distance map based on the skeleton was cropped and then fed into a beta variational auto-encoder model to encode the relative folding characteristics of the left central sulcus region in the latent space.

2.2. Primary sulcal folding pattern

To better understand the mechanisms of typical and atypical cortical sulcal folding patterns, it could be crucial to observe early sulcal folds that occur in developing brain before third trimester (Lohmann et al., 2008). Since the first sulcal folds may develop into the deepest local points of sulcal fundus regions with spatial stability, *in vivo* MRI has been used for identifying sulcal pits to reflect putative early sulcal folds in mature brains (Im et al., 2010; Lohmann et al., 2008). Im et al. first proposed a comprehensive and quantitative SPA to quantify abnormalities in primary sulcal patterns by applying a spectral graph-matching algorithm to pairs of primary sulcal pattern graphs between an individual and a pre-defined set of typical subjects (Im et al., 2011). This approach has since been widely adopted to detect altered sulcal patterns in various diseases (Im et al., 2016; Maleyeff et al., 2024b; Morton et al., 2020, 2021; Ortinau et al., 2019; Tarui et al., 2018, 2023). Meng et al., also developed a primary sulcal pattern graph-based method to mine the common patterns of cortical folding (Meng et al., 2018). They constructed a comprehensive similarity matrix for the entire dataset by adaptively fusing the matrices from six distinct metrics, such as sulcal pit depth and area, to capture the complementary information. Yadav et al. proposed a population-based graph-matching approach that accounts for the significant inter-individual variability in the topology of sulcal graphs (Yadav et al., 2023). Moreover, these authors explored the potential of graph representation learning by applying GNN for gender classification task (Yadav et al., 2024).

2.3. Prototype-based networks for the image domain

A growing body of literature in computer vision explores accurate yet interpretable image classification models by integrating case-based reasoning into networks. Prototypical part networks (ProtoPNet) (Chen et al., 2019) and TesNet (Wang et al., 2021) stand as the most representative works, wherein a prototypical layer captures activation patterns through the comparison of important image segments with class-specific and learnable prototypes, preceding the fully connected classification layer. This concept was refined by adopting a data-driven merging-pruning process (Rymarczyk et al., 2021) and a neural decision tree (Nauta et al., 2021) to reduce the number of prototypes for better

effectiveness. Rymarczyk et al. (2022) introduced a differentiable prototype assignment by adopting a prototype pool for the soft assignment of prototypes. Gautam et al. (2022) defined three properties of prototype-based models and proposed a variational autoencoder-based prototypical network (ProtVAE) designed to enforce transparent, diverse, and trustworthy models.

2.4. Graph representation learning

GNNs are promising methods that are used for major graph representation tasks, including graph classification, edge prediction, and node classification (Corso et al., 2020; Gasteiger et al., 2018). In recent years, the most widely used methods for GNNs have learned node representations by gradually aggregating local neighbors, which is known as message-passing (Dwivedi et al., 2023; Hamilton et al., 2017; Kipf and Welling, 2016). Among these, the graph convolutional networks (GCN) provide a localized first-order approximation of Chebyshev filters on graphs for a simplified graph propagation rule that bridges the gap between spectral and spatial graph convolutions (Kipf and Welling, 2016). GraphSage is one of the most representative spatial GNNs, in which the aggregation function adopts a sampling strategy for each node (Hamilton et al., 2017). Graph attention networks (GATs) employ attention mechanisms to learn relative weights between the target node and its neighbors (Bresson and Laurent, 2017; Velickovic et al., 2017). Furthermore, Xu et al. explored the expressive power of the GNNs based on the Weisfeiler-Lehman Isomorphism Test (Leman and Weisfeiler, 1968; Xu et al., 2018). Although this model was proposed to improve theoretical limitations of message-passing GNNs, it can be interpreted as a GCN (Dwivedi et al., 2023).

2.5. Prototype-based GNNs

Despite advances in prototype-based image models, little effort has been made for GNNs. As the closest analogy to prototype-based models in the image domain, ProtGNN was introduced to focus on graph classification tasks (Zhang et al., 2022). The closest subgraphs were identified from training graphs for each prototype by employing a Monte Carlo tree search algorithm to interpret learned prototype vectors in latent space. However, the learned prototypes are inherently limited to parts of the training dataset, and the tree-searching algorithm is time-consuming. Ragno et al. (2022) adapted ProtoPNet and TesNet to focus on graph and node classification tasks in the graph domain. They assumed that individual node embedding encapsulates latent information of k-hop subgraphs surrounding the node and compared this node embedding with prototypes. In contrast, ProtoPNet and TesNet used patches in the image for comparison. A prototype-based self-explainable GNN (PxGNN) uses a graph generator to construct prototype graphs based on learned prototype embeddings (Dai and Wang, 2022). Specifically, a graph-based autoencoder is pretrained using a training dataset with reconstruction loss concerning connectivity and node attributes, which is subsequently used during training for the primary task. However, the authors imposed a stringent regularization constraint to produce realistic prototype graphs using the graph generator, ensuring that the initialized prototypes from the pretrained generator remain unchanged.

3. Materials and method

3.1. Dataset

In this study, we included MRI datasets of patients with CHD and healthy controls from previously published studies: single ventricle (SV)-CHD (Morton et al., 2020), transposition of the great arteries (TGA)/tetralogy of Fallot (ToF)-CHD (Morton et al., 2021), and Pediatric Cardiac Genomics Consortium (PCGC) (Maleyeff et al., 2024a; Morton et al., 2023; Richter et al., 2020) cohort studies. These studies were approved by the Boston Children's Hospital (BCH) institutional review board (SV-CHD and TGA/ToF-CHD cohort studies) or the central Institutional Review Board (PCGC cohort study). Reliance agreements were approved at each study center for the multicenter cohort. Written informed consent was obtained from the participants (aged $\geq$18 years) or their parents or guardians (aged $<$18 years). We enrolled participants in the SV-CHD cohort between 2010 and 2012. Participants with SV-CHD were recruited using the inclusion and exclusion criteria described in our previous study (Morton et al., 2020). Healthy control participants of a similar age were recruited from local pediatric practices, our institutional adolescent clinic, and through posted notices. The exclusion criteria from the National Institutes of Health (NIH) MRI study of normal brain development were applied to the participants (Evans and Group, 2006). Patients with ToF-CHD were enrolled from 2004 to 2008 at the BCH with the inclusion and exclusion criteria as previously described (Morton et al., 2021). The inclusion criteria for TGA-CHD were (1) diagnosis of TGA with or without a ventricular septal defect and (2) planned arterial switch surgery at 3 months of age. Patients with TGA-CHD were recruited between 1988 and 1992 with the same inclusion and exclusion criteria described in the previous study (Maleyeff et al., 2024b). A group of healthy control adolescents was recruited using the exclusion criteria of the NIH MRI study of healthy participants. For the PCGC cohort, participants from our previous study were enrolled using the same inclusion and exclusion criteria (Maleyeff et al., 2024a). Finally, we included a publicly available dataset of healthy controls from the Human Connectome Project (HCP) (Van Essen et al., 2012) (22–25 years, 53 % male sex, released in November 2014), as in our previous study (Morton et al., 2020). Our study included 345 patients with CHD (115 SV-CHDs, 92 TGA-CHD, 41 ToF-CHD, and 97 patients with CHDs from the PCGC cohort) and 174 healthy controls without CHD (45 controls from the SV-CHD cohort, 49 controls from the TGA/TOF-CHD cohort, and 80 controls from the HCP) (Table 1).

3.2. MRI acquisition

In the SV-CHD cohort, MRI was acquired using a 1.5-T General Electric Twin-speed magnetic resonance scanner for subjects with implanted cardiovascular devices or coils or a 3-T General Electric system (General Electric Medical Systems). For the TGA/ToF-CHD cohort, an MRI was performed using a 1.5 Tesla GE Twin Speed magnetic

Table 1
Participant characteristics for each cohort.

Variable (CHD/non-CHD)	SV-CHD	TGA/ToF-CHD	PCGC—CHD	HCP	All
n	115/45	133/49	97/0	0/80	345/174
Age at MRI	14.8±3.0/ 15.5±2.5	15.8±1.0/ 15.4±1.2	17.2±8.0/ -	-/ 23.6±1.1	15.8±4.7/ 15.4±1.9
Male sex	67 (59 %)/ 26 (58 %)	97 (73 %)/ 20 (41 %)	56 (59 %)/ -	-/ 42 (53 %)	220 (64 %)/ 88 (51 %)

Mean ± standard deviations are reported for *age at MRI* (years). For *Male sex*, data are expressed as *n* (%) for male participants.

resonance scanner at the BCH. By using a T1-weighted 3D spoiled gradient recalled steady-state sequence, patients were scanned without sedation using the following parameters: repetition time (TR)/echo time (TE) = 7 ms/2.8 ms, flip angle = 8°, acquisition matrix = 256 × 256, field of view (FOV) = 256 mm, slice thickness = 1 mm (3T), and TR/TE = 40 s/4 s, flip angle = 20°, acquisition matrix = 256 × 192, FOV = 240 mm, and slice thickness = 1.5 mm (1.5T). MRI acquisition in the PCGC cohort closely followed the Adolescent Brain Cognitive Development Study, as previously described (Maleyeff et al., 2024a). A neuroradiologist inspected all images to ensure data quality and exclude structural abnormalities, including tumors, stroke, and major injuries.

3.3. Image preprocessing and cortical sulcal pattern graph construction

We used the FreeSurfer pipeline to process the images and extract cortical surfaces (Dale et al., 1999; Fischl, 2012; Fischl et al., 1999). This study used left and right white matter surfaces (gray/white matter boundaries). To ensure accuracy, we visually inspected the cortical surface of each participant. The global sulcal pattern was represented using a graph structure with sulcal pits and surrounding catchment basins decomposed from primary sulcal segments as nodes (Im et al., 2010, 2011). We generated a sulcal depth map on the inner cortical surface using FreeSurfer and smoothed the map using a surface-based heat kernel smoothing with a full-width half-maximum of 10 mm to prevent over-extraction of the sulcal pits (Chung et al., 2005; Im et al., 2010). We used this smoothed sulcal depth map to identify sulcal pits and their sulcal basins based on a watershed segmentation algorithm (Im et al., 2013b) (Fig. 1a). The sulcal pits were connected to an edge in the sulcal graph representation if their sulcal basins coincided. Each node had geometric features, comprising the sulcal area, sulcal depth, and 3D position of the sulcal pit (Fig. 1b). Moreover, we compared basic statistics, such as the number of nodes, of the sulcal pattern graphs between CHD patients and healthy controls using linear regression model, adjusting for age and sex ($Y = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 group$, where Y denotes one of the statistics). As shown in Table 2, the number of nodes, and edges, mean sulcal area, and mean sulcal depth were not significantly different between two groups.

3.4. Proposed model: PIP

In this section, we describe the architecture of the proposed PIP methodology. As shown in Fig. 3, the PIP consists of four key components: (1) a graph encoder, (2) a prototype inverse-projection layer, (3) a prototype similarity layer, and (4) a prediction layer.

3.4.1. Preliminaries

We set $\{\mathcal{G}_i, y_i\}_{i=1}^{N}$ as the training dataset, where $\mathcal{G}_i$ is an input individual sulcal pattern graph for subject i and $y_i \in \{1, \ldots, C\}$ is the corresponding class label for the graph (C is the number of classes). The graph $\mathcal{G}_i$ can be denoted as (V_i, E_i) with a set of nodes V_i and edges $E_i \in V_i \times V_i$. We let $A_i \in \{0,1\}^{|V_i|\times|V_i|}$ denote an adjacency matrix describing the topology of the graph $\mathcal{G}_i$. The set of nodes V_i is associated with the attribute matrix $X_i = \left[\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \ldots, \boldsymbol{x}_i^{|V_i|}\right]^T$, where $\boldsymbol{x}_i^v \in \mathbb{R}^{d_{input}\times 1}$ denotes the input node feature vector for node $v \in \{1, \ldots, |V_i|\}$ in subject i. d_{input} represents the number of input feature dimensions.

Table 2
Basic statistics of sulcal pattern graphs for each group.

	Group		β±SE	P value
	CHD	Control		
# of nodes	168.50±16.05	170.06±14.44	3.34±3.94	0.40
# of edges	934.16±93.47	944.32±85.12	18.12±23.17	0.44
Sulcal area	1120.00±76.36	1092.04±65.53	6.52±18.49	0.73
Sulcal depth	0.70±0.05	0.68±0.04	−0.01±0.01	0.57

Mean ± standard deviation for each group (CHD, and Control) are reported. β: regression coefficient; *SE*: standard error.

We used sulcal pattern graphs defined on the whole brain for each individual. Based on the surface model constructed using the FreeSurfer algorithm, we initially have pairs of mid-surfaces for the left and right hemispheres, resulting in two independent sulcal pattern graphs. Thus, we combined them by concatenating the node feature matrices and constructing a block diagonal adjacency matrix. Specifically, the node feature matrix X_i and an adjacency matrix A_i can be obtained as follows:

$$X_i = concat\left(X_i^{left}, X_i^{right}\right) \tag{1}$$

$$A_i = \begin{bmatrix} A_i^{left} & 0 \\ 0 & A_i^{right} \end{bmatrix} \tag{2}$$

where $concat(\cdot)$ denotes the matrix concatenation operation, and $X_i^{left} \in \mathbb{R}^{|V_i^{left}|\times d_{input}}$, and $X_i^{right} \in \mathbb{R}^{|V_i^{right}|\times d_{input}}$ are node feature matrix for left, and right hemisphere. $|V_i^{left}|$, and $|V_i^{right}|$ denote the number of nodes for each hemisphere. Similarly, $A_i^{left} \in \{0,1\}^{|V_i^{left}|\times|V_i^{left}|}$, and $A_i^{right} \in \{0,1\}^{|V_i^{right}|\times|V_i^{right}|}$ denote adjacency matrices for each hemisphere.

3.4.2. Node embedding

We first calculated node hidden embeddings $H = \left[\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{|V|}\right]^T$, where $\boldsymbol{h}_v \in \mathbb{R}^{d\times 1}$ using multilayer perceptron (MLP) with a hidden layer for each node $v \in \{1, \ldots, |V|\}$:

$$\boldsymbol{h}_v = U_2(ReLU(U_1\boldsymbol{x}_v)). \tag{3}$$

$ReLU(\cdot)$ is a rectified linear unit activation function, and $U_1 \in \mathbb{R}^{d\times d_{input}}$, and $U_2 \in \mathbb{R}^{d\times d}$ are trainable parameters. We present the equations without the bias terms and subject indices for notational convenience. We then applied a message-passing-based GNN (GraphSage) to the node embeddings to learn node representations (Hamilton et al., 2017). With the input node embedding $\boldsymbol{z}_v^0 = \boldsymbol{h}_v$, the GNN comprises L layers, and the updating rule for layer l is as follows:

$$\boldsymbol{z}_v^{l+1} = ReLU\left(W_l concat\left(\boldsymbol{z}_v^l, (1/\mathcal{N}_v)\sum_{u\in\mathcal{N}_v}\boldsymbol{z}_u^l\right)\right) \tag{4}$$

where $\mathcal{N}_v$ is the number of neighboring nodes for node v. $W_l \in \mathbb{R}^{d\times 2d}$ is a trainable parameter for layer l. For each layer, node representation $\boldsymbol{z}_v^{l+1}$ was $l2-$ normalized before passing to the following layer.

3.4.3. PIP

Following the concept reasoning paradigm, we had trainable prototype patterns $P_c = \left\{p_{c,k} \in \mathbb{R}^{d\times 1}\right\}_{k=1}^{K}$ for all classes $c \in \{1, \ldots, C\}$. Before calculating the similarity scores, the prototype patterns were inversely projected onto the subspace spanned by the learned node representations $Z^L = \left[\boldsymbol{z}_1^L, \boldsymbol{z}_2^L, \ldots, \boldsymbol{z}_{|V|}^L\right]^T$. Based on previous work on natural language processing (Bahdanau et al., 2014), an attention-based network was used to estimate inverse-projection weights. Formally, an inverse-projection weight $\omega_v^{p_{c,k}}$ between the $k-$ th prototype $p_{c,k}$ from class c and the embedding vector $\boldsymbol{z}_v^L$ of $v-$ th node was calculated as:

$$\omega_v^{p_{c,k}} = \theta_2^T\left(tanh\left(\Theta_1 concat\left(p_{c,k}, \boldsymbol{z}_v^L\right)\right)\right) \tag{5}$$

where $\Theta_1 \in \mathbb{R}^{d\times 2d}$, and $\theta_2 \in \mathbb{R}^{d\times 1}$ are trainable parameters and $tanh(\cdot)$ denotes the hyperbolic tangent activation function.

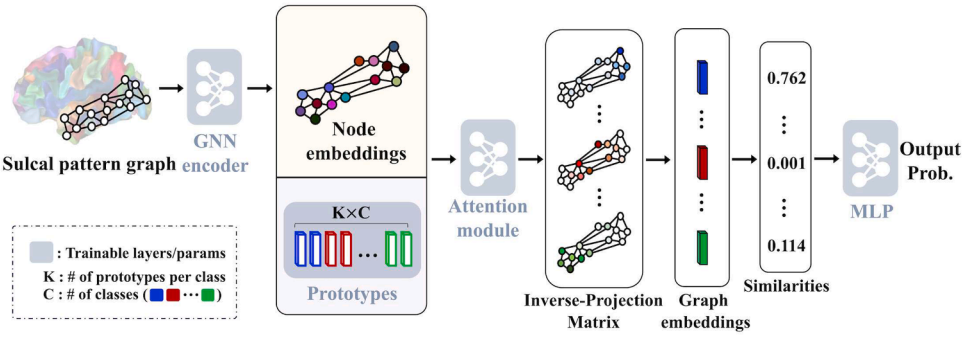

Fig. 3. Architecture of the proposed framework. After passing through the graph neural networks-based node embedding layer (GNN encoder), the learnable prototype vectors are inversely projected onto the node embeddings to calculate the inverse projection matrix using the attention-based module (Attention module). Then the inverse projection matrix is used to obtain prototype-specific graph-level representations for each prototype, and prototype-specific similarities are calculated by using them. Finally, a multi-layer perceptron (MLP) is utilized using these similarity values to estimate the class probability (Output Prob.) for classification.

3.4.4. Prototype focal similarity

Once an inverse-projection matrix $\Omega_{c,k} = \left\{\omega_v^{p_{c,k}}\right\}_{v=1}^{|V|} \in \mathbb{R}^{|V| \times 1}\ \forall \{c,k\}$ was obtained, a prototype-specific graph representation $z_{\mathcal{G}}^{p_{c,k}}$ was defined as:

$$z_{\mathcal{G}}^{p_{c,k}} = \left(\Omega_{c,k}^T Z^L\right)^T \tag{6}$$

This can be considered an attention-based readout function involving prototype-specific pooling vector $\Omega_{c,k}^T$. Instead of applying a conventional readout function (e.g., readout using mean pooling vector $\Omega_{mean}^T = [1/|V|, \ldots, 1/|V|]^T$ and sum pooling vector $\Omega_{sum}^T = [1, \ldots, 1]^T$), we used this prototype-specific graph representation to calculate similarity scores $sim\left(z_{\mathcal{G}}^{p_{c,k}}, p_{c,k}\right)$ for each prototype as follows:

$$sim\left(z_{\mathcal{G}}^{p_{c,k}}, p_{c,k}\right) = \log\left(\left(\| z_{\mathcal{G}}^{p_{c,k}} - p_{c,k}\|_2^2 + 1\right) / \left(\| z_{\mathcal{G}}^{p_{c,k}} - p_{c,k}\|_2^2 + \epsilon\right)\right) \tag{7}$$

As in Rymarczyk et al. (2022), we employed a focal similarity strategy that enabled the proposed model to focus on crucial parts of the input graph. The focal similarity maximizes the gap between the maximal and average similarity activations in the image domain. To calculate the focal similarity $g_{p_{c,k}, z_{\mathcal{G}}^{p_{c,k}}}$ between the prototype-specific graph representation and the prototype, we adapted the existing method based on the image domain to our graph-structured dataset as follows:

$$g_{p_{c,k}, z_{\mathcal{G}}^{p_{c,k}}} = sim\left(z_{\mathcal{G}}^{p_{c,k}}, p_{c,k}\right) - sim\left(z_{\mathcal{G}}^{mean}, p_{c,k}\right) \tag{8}$$

where $z_{\mathcal{G}}^{mean} = \Omega_{mean}^T Z^L$ denotes the graph representation defined by the readout function with the mean pooling vector Ω_{mean}^T.

3.4.5. Learning objective

To solve the problems associated with existing SPA, we considered an objective function, including cross-entropy loss and several constraints in training our model. First, two cluster costs are used to make the learned prototypes class-representative and meaningful as follows:

$$\mathcal{L}_{clst,1} = 1/N \sum_{i=1}^{N} \min_k \| z_{\mathcal{G}_i}^{p_{y_i,k}} - p_{y_i,k} \|_2^2 \tag{9}$$

$$\mathcal{L}_{clst,2} = 1/C{\cdot}K \sum_{c=1}^{C} \sum_{k=1}^{K} \min_i \| p_{c,k} - z_{\mathcal{G}_i}^{p_{c,k}} \|_2^2 \tag{10}$$

The minimization of the first term ($\mathcal{L}_{clst,1}$) requires that the training samples form clusters around the prototypes in the latent space. The second term ($\mathcal{L}_{clst,2}$) causes the learned prototypes to be close to at least one training sample, resulting in meaningful prototypes. Furthermore, the orthogonal cost was defined to encourage a diversity of prototypes as follows:

$$\mathcal{L}_{orth} = 1/C \sum_{c=1}^{C} \| P_c P_c^T - I_K \|_F \tag{11}$$

where $I_K \in \mathbb{R}^{K \times K}$ is the identity matrix and $\| \cdot \|_F$ denotes the matrix Frobenius norm operator. In summary, with the final classification layer consisting of fully connected layers $f(\cdot) : \mathbb{R}^{C \cdot K} \rightarrow [0,1]^C$ and $C{\cdot}K$ similarity scores $g_i = \left\{g_{p_{c,k}, z_{\mathcal{G}_i}^{p_{c,k}}}\right\}_{c=1..C,\ k=1..K} \in \mathbb{R}^{C \cdot K \times 1}$ for each subject i, our optimization objective $\mathcal{L}$ was defined as:

$$\mathcal{L} = 1/N \sum_{i=1}^{N} CE(f(g_i), y_i) + \lambda_{clst,1} \cdot \mathcal{L}_{clst,1} + \lambda_{clst,2} \cdot \mathcal{L}_{clst,2} + \lambda_{orth} \cdot \mathcal{L}_{orth} \tag{12}$$

where $CE(\cdot)$ denotes the cross-entropy loss and $\lambda_{clst,1}$, $\lambda_{clst,2}$, and λ_{orth} are hyperparameters for weighting each constraint.

4. Experiments

To evaluate the proposed model, we compared the classification performance of our PIP with those of black-box models and prototype-based GNNs. Ablation studies were conducted to investigate the effects of the various configurations of the proposed model. We also visualized the prototypes learned from our model by aggregating the similarity scores from the inverse-projection matrix. Furthermore, the relationships between the prototypes were investigated for a better understanding.

4.1. Details of implementation

We split the data using 10-fold stratified cross-validation with 80 % training, 10 % validation, and 10 % test samples. We then reported the mean and standard deviation of the area under the ROC curve (AUC) for the test samples across 10 folds. A stochastic gradient descent optimizer adopting an early stopping strategy with an initial learning rate of 1e-3 reduction for every 25 patients was employed. Training was performed for a maximum of 1000 epochs for each fold with a batch size of 48. The minimum learning rate was set to 1e-5. We set the momentum and weight decay to 0.9 and 5e-4, respectively. For classification, we set the number of prototypes per class to two ($K = 2$), resulting in four prototypes for the two classes (CHD vs. healthy control; $C = 2$). The node-embedding networks consisted of $L = 2$ layers with a hidden dimension d of 64. In the objective function, we set $\lambda_{clst,1}$, $\lambda_{clst,2}$, and λ_{orth} to 1, 0.1, and 0.5, respectively. All the hyperparameters were selected based on a grid search algorithm or related studies. We conducted all the experi-

ments with an NVIDIA Titan V 12 GB GPU using the PyTorch framework.

4.2. Comparison with state-of-the-art models

We compared the classification performance of the proposed PIP with those of state-of-the-art black-box networks and prototype-based GNNs. An MLP consisting of two hidden layers and GNNs, including GCNs, GraphSage, and a GAT, was employed for black box benchmarking. The prototype-based GNNs were also compared with our proposed model. For ProtoPNet and TesNet, Ragno et al. (2022) used a global max-pooling readout layer using only the single most similar node for each prototype to adapt the original model in the image domain to graph-structured data. However, we hypothesized that a single node encoding only a subgraph of L-hop neighbors is insufficient for identifying abnormal sulcal patterning because of the interdependence of cortical areas. Thus, we modified these models slightly by utilizing the mean-pooling readout layer in addition to the max-pooling operator to enable the model to capture more global effects on abnormal cortical folding. We adopted an unofficial implementation because some benchmarking prototype-based GNNs are not publicly available. We also performed a non-parametric Wilcoxon signed-rank test (Wilcoxon, 1992) to investigate whether the performance of the proposed model achieved significantly higher test AUC scores compared to the other models across training folds. Moreover, to verify the consistency of the proposed model, we trained it using five different random seeds. Except for the result from one seed reported here, the results from the other four random seeds were provided in the supplementary materials (Table S1). All codes, hyperparameters, and models used in this study are available at https://github.com/hookhy/surfacepip.

4.3. Ablative study

To investigate the effectiveness of the proposed PIP mechanism, we conducted an ablation study for various model configurations. Specifically, we replicated the classification analysis by modifying the model configuration using several approaches. First, we excluded the inverse-projection from the loss terms by modifying the cluster constraints in Eqs. (9) and (10) as follows:

$$\widehat{\mathscr{L}}_{clst,1} = 1/N \sum_{i=1}^{N} \min_{k} \| \boldsymbol{z}^{mean}_{\mathscr{G}_i} - \boldsymbol{p}_{y_i,k} \|_2^2 \tag{13}$$

$$\widehat{\mathscr{L}}_{clst,2} = 1/C{\cdot}K \sum_{c=1}^{C} \sum_{k=1}^{K} \min_{i} \| \boldsymbol{p}_{c,k} - \boldsymbol{z}^{mean}_{\mathscr{G}_i} \|_2^2 \tag{14}$$

where $\boldsymbol{z}^{mean}_{\mathscr{G}_i} = \boldsymbol{\Omega}^T_{i,mean} Z_i^L$ denotes the graph representation for subject i defined by the readout function $\boldsymbol{\Omega}^T_{i,mean} \in \mathbb{R}^{|V_i| \times 1}$. This configuration mimicked the loss terms used in existing prototype-based models. Second, focal similarity was excluded in Eq. (8), resulting in a regular prototype similarity score between the prototype-specific graph representation and the prototype pattern, as follows:

$$\widehat{g}_{\boldsymbol{p}_{c,k}, \boldsymbol{z}^{\boldsymbol{p}_{c,k}}_{\mathscr{G}}} = sim\left(\boldsymbol{z}^{\boldsymbol{p}_{c,k}}_{\mathscr{G}}, \boldsymbol{p}_{c,k}\right) \tag{15}$$

The only difference was that this configuration did not suppress the average similarity activations $sim\left(\boldsymbol{z}^{mean}_{\mathscr{G}}, \boldsymbol{p}_{c,k}\right)$. Finally, the inverse-projection layers of the prototype were excluded from the architecture. In this setting, conventional prototype projection was performed for classification. We also performed an ablation study on loss constraints by excluding the cluster term ($\lambda_{clst,2} = 0$) or an orthogonal term ($\lambda_{orth} = 0$) from the final objective and calculating the cosine similarity matrix between the learned prototype patterns.

Finally, we conducted an ablation study on the number of prototypes per class K by replicating the classification experiment using different values of K.

Similarly to the classification experiments in Section 4.2, we also performed the Wilcoxon signed-rank test to investigate whether the performance variations of the proposed model are significantly higher than other configurations in these ablation studies.

4.4. Visualization of the PIP matrix

We visualized the inverse-projection matrix on the template brain surface model to interpret the learned prototypes better. The individual surfaces were aligned to the MNI ICBM152 surface template using surface registration (Boucher et al., 2009; Lyttelton et al., 2007). The sulcal pits with the top 10 % similarity values were identified for each subject by thresholding the inverse-projection weights. To consider the individual variance in the spatial distribution of sulcal pits, binary spheres with a diameter of 5 mm centered at the surviving sulcal pits were created by performing a distance transform on the surface. Frequency maps were constructed by overlaying the spherical regions of all the participants for each prototype on the template surface. Finally, the merged frequency maps were averaged by the number of participants to visualize the heatmap of the inverse-projection matrix for each prototype.

5. Results

5.1. Classification performance

We compared the classification performance of the proposed PIP with that of the other methods and reported the results in Table 3. Our model with GraphSage encoder achieved the highest test AUC with a small increase in the model parameters compared to other models. Specifically, the PIP corroborated trustworthy interpretability for sulcal pattern analysis by surpassing the black-box counterpart models, such as other GNNs and MLPs, unlike other prototype-based benchmarks. TesNet and ProtoPNet's global counterparts achieved superior classification performances compared to the original models; however, they still showed less satisfactory results than our proposed model. Fig. 4 shows that the proposed model can learn class-representative prototypes. Visualization of the training graph embeddings and learned prototypes using t-SNE in the latent spaces revealed that the learned prototypes were well represented by a compact cluster centered around the training data points of the corresponding class. Furthermore, since we utilized

Table 3
Classification performances.

Model	Test AUC	# Params	# Epochs	Epoch/Total
MLP†	61.7 ± 5.3	1793	206.6	0.80 s
GCN	70.5 ± 7.6	2977	295.8	0.90 s
GraphSage	71.7 ± 5.8	5025	214.0	1.05 s
GAT	71.8 ± 7.0	3105	228.4	1.17 s
ProtGNN	71.0 ± 7.4	4612	252.9	1.77 s
ProtoPNet†	59.8 ± 9.5	4612	267.6	5.17 s
TesNet	69.8 ± 8.9	4612	232.5	1.44 s
ProtoPNet*,†	61.3 ± 8.4	4612	294.7	4.51 s
TesNet*	68.4 ± 8.9	4612	253.5	2.61 s
PxGNN†	58.3 ± 11.5	5108	202.9	8.95 s
PIP (GCN)	70.1 ± 8.7	5832	647.8	4.10 s
PIP (GraphSage)	73.5 ± 5.6	6692	461.6	2.43 s
PIP (*GAT*)	70.5 ± 4.7	5960	558.7	4.70 s

Test mean ± standard AUC scores of all training folds are reported. Params, epochs, and epoch/total denote the number of trainable parameters, training epochs, and the average time in seconds per epoch, respectively. * denotes that a modified version of the model was applied. †: FDR-corrected P-value < 0.05 significance in the Wilcoxon signed-rank test compared to the PIP (GraphSage). The classification performances of our models with different GNN-based encoder (PIP (GCN), and PIP (GAT)) are also reported.

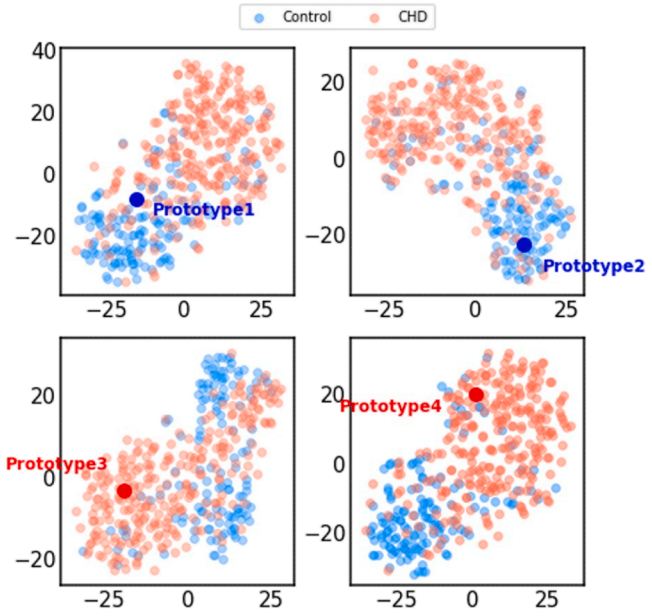

Fig. 4. Visualization of training samples and learned prototypes. Because the different inverse projection weights for each prototype are applied to obtain the final graph-level representation, we show four different subfigures for each latent space corresponding to four different prototypes. The small blue and red circles denote training samples of the healthy control and CHD classes, respectively, and the four large circles denote the learned prototypes. We used t-distributed stochastic neighbor embedding (T-SNE) to obtain the latent spaces.

Table 4
Classification performances for various configurations of the proposed model.

Model	Test AUC	# Params	# Epochs	Epoch/Total
w/o AC	73.6 ± 4.9	6692	376.9	2.76 s
w/o FS	72.1 ± 5.7	6692	454.2	2.59 s
PP	69.0 ± 5.4	6692	269.9	1.51 s
PIP (*ours*)	73.5 ± 5.6	6692	461.6	2.40 s

w/o AC: model trained without attentive constraints (Eqs. (13) and (14)); w/o FS: model trained without focal similarity (Eq. (15)); PP: model with prototype projection and without a prototype inverse-projection.

multicenter cohorts collected from different sites and scanning protocols, potential bias due to site effects may exist. Therefore, we additionally visualized the t-SNE latent space with cohort labels for each training embedding to investigate whether the learned prototypes and training embeddings have inadvertently captured biases due to site effects (Fig. S1).

5.2. Ablative study

Table 4 shows the performances of various configurations of the proposed model. The model employing the modified cluster constraints achieved no loss in classification performance and performed slightly better than the proposed model. To explain this result, we examined the trained inverse-projection matrices of the training subjects, as shown in Fig. 5. The trained inverse-projection matrices exhibited a uniform distribution, as shown in Fig. 5a, indicating that this configuration did not provide any interpretation. However, our proposed model focuses on specific regions and provides better interpretability by showing diverse patterns for the inverse-projection matrices (Fig. 5c). When we removed the focal similarity and prototype inverse-projection layers, performance degradation was observed.

Table 5 shows that the proposed model achieved the best classification performance, supporting the effects of the cluster and orthogonal constraint terms in the final objective function. Moreover, we examined the inverse-projection matrices shown in Fig. 6 to better understand the effects of these constraint terms. Interestingly, we found that the elimination of the second cluster term ($\lambda_{\text{clst},2} = 0$) caused some of the prototypes to have inverse-projection weights that were too small (Fig. 6a). This demonstrates the limited interpretability of the model due to the absence of the second cluster term. Unlike the first cluster term ($\mathscr{L}_{\text{clst},1}$) enforces the training samples to form clusters around the prototypes, the second cluster term ($\mathscr{L}_{\text{clst},2}$) requires every prototypes to be meaningful by pushing them to be close to at least one training sample. Thus, without the second cluster term, certain prototypes could be excluded from model training due to a low gradient.

Moreover, when we removed the orthogonal term ($\lambda_{orth} = 0$), the trained inverse-projection matrices showed non-discriminative patterns between prototypes (Fig. 6b). This result indicates that the trained prototypes collapse into a single point, which is undesirable for prototype-based learning (Gautam et al., 2022). Quantitatively, the cosine similarities among the trained prototype vectors were close to 1 when the orthogonal term was removed ($\lambda_{orth} = 0$). In contrast, the proposed model generated diverse prototypes exhibiting almost zero inter-prototype correlations (Fig. 7).

Next, we investigated the effect of the number of prototypes per class (K) among {1, 2, …, 20} and reported the results in Fig. 8. The upper bound for this grid search was defined based on the observation that the classification performance decreased after K=4. We suspect that either the classification task on CHD cohorts is too simple to learn the inherent and iconic sulcal patterns or that the lack of training data limits the full potential of the proposed model. As shown in the figure, the PIP with $K = 2$ and $K = 4$ achieved the best classification performance. Thus, we set K to 2 because a larger model with a larger K requires more memory and time. Moreover, the number of prototypes per class is expected to have a major influence on both classification performance and learnt prototype(s). However, the experiment reported on Fig. 8 suggests that the influence of K on the classification performance is relatively subtle. Thus, we explored the latent spaces by visualizing them in the same way as shown in Fig. 4 for each value of K, to assess whether the learned prototypes are sufficiently representative of the corresponding classes. As shown in the supplementary materials (Figs. S2–S4), the prototypes learned by the PIP appeared to be sufficiently representative, even when the number of prototypes set to be large, supporting our claim of generative ability in the proposed network. We concluded that the lack of significant performance improvements, despite a large number of prototypes per class, might be due to overfitting resulting from the limited size of the dataset.

5.3. Prototype analysis

The inverse-projection weights for each prototype are visualized in Fig. 9. The prototypes of the healthy control group were most strongly activated in similar regions, including the left transverse temporal gyrus, left precuneus, paracentral lobule, and right postcentral gyrus (Fig. 9a). They also had different activation patterns in the right superior temporal and right transverse temporal gyri and right paracentral lobule. In the case of prototypes of the CHD class, we observed that the first prototype (prototype 3 in Fig. 9b) had distinct patterns in specific regions in the left hemisphere, including the insula cortex, inferior parietal lobule, superior temporal gyrus, and pericalcarine cortex, compared to prototypes of the healthy control class. However, the left paracentral lobule and left precuneus were not identified in the first prototype of the CHD class (Prototype 3), whereas the other prototypes showed high activation in these regions. The right superior temporal and right transverse temporal gyri had patterns similar to those of the first prototype of the healthy control class (prototype 1 in Fig. 9a). Finally, the second prototype of the CHD class (prototype 4 in Fig. 9b) showed high activation patterns in the

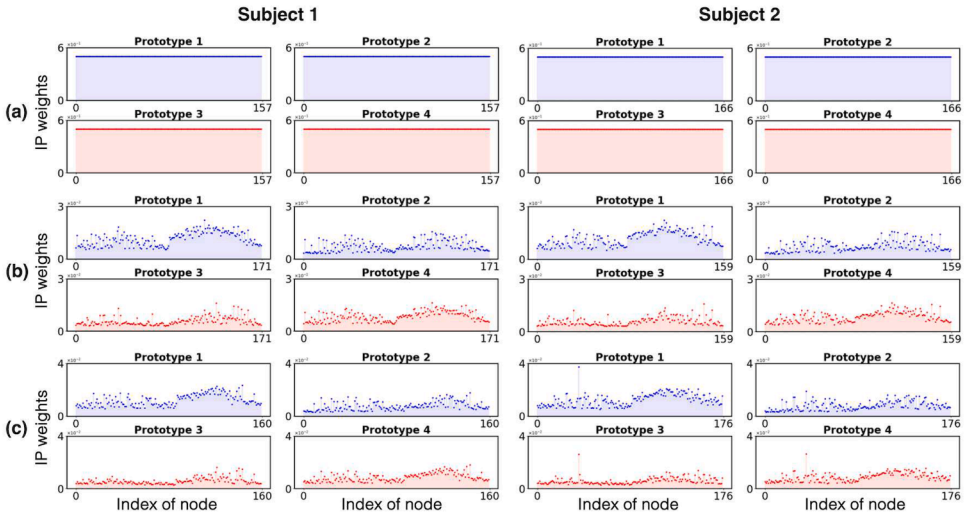

Fig. 5. Inverse projection weights (y-axis) across nodes (x-axis) on two example subjects (Left: CHD patient. Right: healthy control) for various configurations of the proposed model. Example cases include (a) a model without attentive constraints by using the mean pooling strategy in two cluster loss terms instead of using the inverse projection matrix (Eq. (13), and Eq. (14)), (b) a model without focal similarity, and (c) the proposed method. The blue and red colors indicate inverse projection weights for the healthy control and CHD classes, respectively.

Table 5
Classification performances for various loss configurations of the proposed model.

Model	Test AUC	# Params	# Epochs	Epoch/Total
w/o clst2	70.9 ± 4.0	6692	405.8	2.65 s
w/o orth	69.6 ± 10.4	6692	337.3	2.51 s
PIP (*ours*)	73.5 ± 5.6	6692	461.6	2.43 s

w/o clst2: the model was trained without the second cluster constraint in Eq. (10); w/o orth: the model was trained without an orthogonal constraint in Eq. (11).

regions identified by the prototypes of the healthy control class.

To provide deeper insight into the learned prototype patterns, we analyzed the inter-prototype relationships, as shown in Fig. 9c. The joint similarity distributions for each pair of prototypes were also investigated. The similarities for prototypes 1 and 4 were highly correlated and interestingly, patients with CHD had a higher similarity for prototype 4 (and prototype 1), a relatively higher similarity for prototype 2, and lower similarity for prototype 3. Among healthy control subjects, individuals showing higher similarity for prototype 1 (and prototype 4) exhibited the opposite trend (low similarity for prototype 2 and high similarity for prototype 3). These examples demonstrate that prototypes 1 and 4 may have intermediate-level atypical patterns between patients with CHD and healthy controls. Conversely, the similarities for prototypes 2 and 3 showed a strong anti-correlation, and we found that these two prototypes could provide more discriminative patterns between the two classes.

The fundamental concept of the quantitative sulcal pattern analysis is to detect subtle and global abnormalities in sulcal patterns that are hard to inspect visually. To capture the alterations in global sulcal patterns, we need to consider not only the geometric features of the sulcal folds themselves but also the interrelated arrangement and topology of these folds. Fig. 10 shows the example of individual cases who exhibit high and low output probabilities of healthy controls. As shown in Fig. 10, the inter-subject variability of complex sulcal patterns makes it

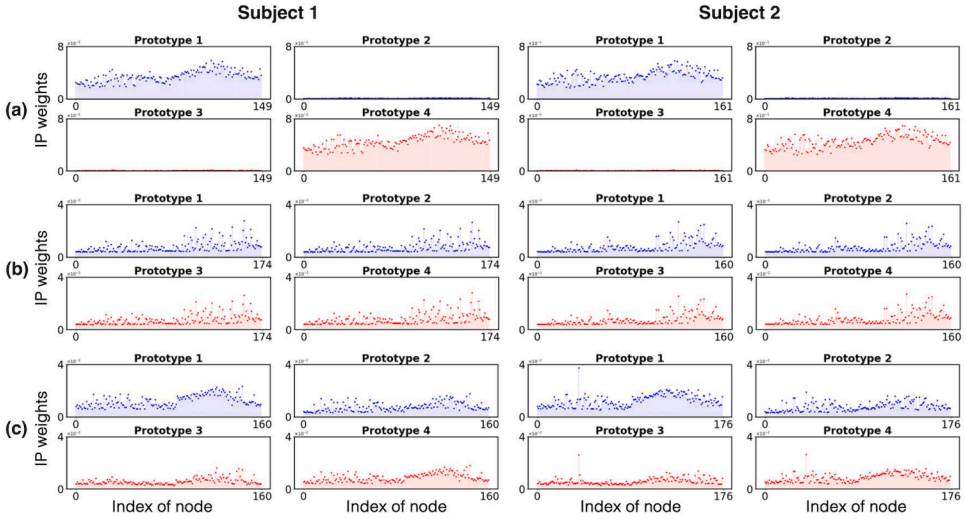

Fig. 6. Inverse projection weights (y-axis) across nodes (x-axis) on two example subjects (Left: CHD patient. Right: healthy control) for various configurations of the proposed model. Example cases include (a) the model without the second cluster term ($\lambda_{clst2} = 0$), (b) the model without the orthogonal term ($\lambda_{orth} = 0$), and (c) the proposed method. The blue and red colors indicate inverse projection weights for the healthy control and CHD classes, respectively.

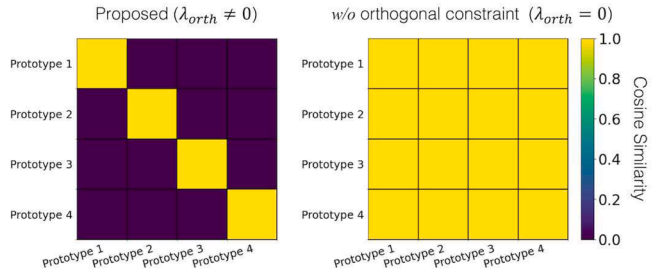

Fig. 7. Inter-prototypical cosine similarity matrices among the learned prototypes from (left) the proposed method and (b) the model without an orthogonal term. The yellow colors in the nondiagonal elements denote that the corresponding prototype pairs show higher intercorrelations, indicating prototype collapse.

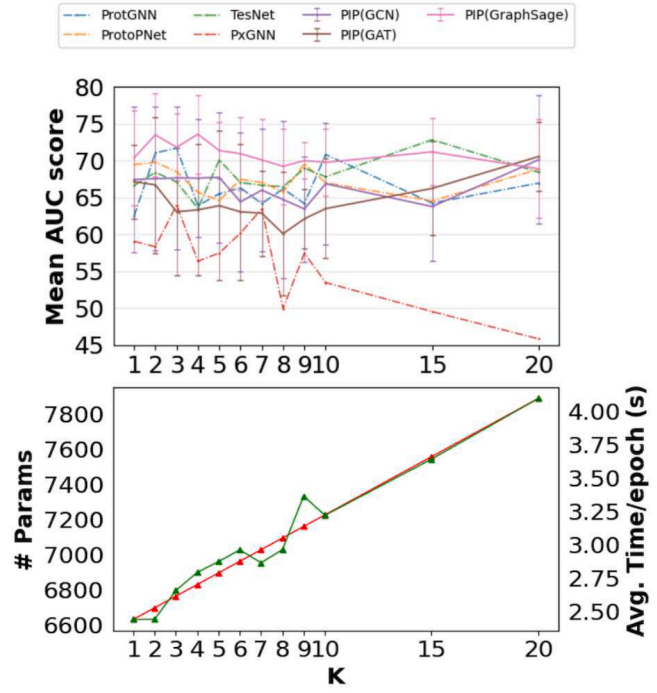

Fig. 8. Experimental results for the number of prototypes (*K*) for each class. (top) The mean and standard deviation (error bars) of AUC scores across 10 training folds for congenital heart disease classification by varying *K*. (bottom) We also examine the number of learnable parameters (# Params, red) for the proposed model (PIP(GraphSage)) and average time (s) per epoch (Avg. Time/epoch (s), green) for each *K*.

difficult to identify alterations in CHD patients by visual inspection, even though they have relatively high similarities for prototypes of CHD class (prototype 3 and 4) according to the proposed model.

Finally, we used t-SNE showing the latent space for all the prototypes along with ten training folds to highlight the consistency along with all the training folds (Fig. 11). Those learned prototypes were remarkably consistent across all training folds, supporting the generalizability of our proposed model.

6. Discussion

6.1. Effect of prototype-based GNNs on the SPA

In this study, we developed a deep learning-based SPA framework by adopting prototype-based GNNs and performed a CHD classification task on sulcal pattern graphs. The experimental results demonstrated the effectiveness of our approach, highlighting its ability to achieve the highest AUC score among state-of-the-art models. Moreover, the learned prototypes are consistent and can be regarded as superior alternatives to the typical subject set of conventional SPA in terms of diversity and capability to learn class-representative patterns (Im et al., 2011, 2013b). In the Wilcoxon signed-rank test, the classification performance of the proposed model (PIP(GraphSage)) was significantly higher than that of MLP, ProtoPNet(*), and PxGNN. Unfortunately, no significant differences were observed between the proposed model and the other models. Similarly, the differences in classification performance between the proposed model and other implementations in ablative studies, such as the number of prototypes (*K*) in Fig. 8 and different (loss) configurations in Tables 4 and 5, were also not statistically significant. However, we observed that the proposed model consistently achieved a higher mean AUC with a lower standard deviation compared to the others, both in the classification task and in the ablative studies.

The final classification layer was applied to the prototype-specific similarities to calculate the output probability for a healthy control class, which can be considered an analogy of the sulcal pattern similarity index of our previous approach. The proposed model also considers the disease class, whereas the conventional SPA only considers the healthy control class. In the conventional approach, the typical subject set consists of typically developed subjects only, which are compared to individual brains to quantify how similar they are to normal sulcal folding patterns. However, our framework calculated the prototype-specific similarities using the prototypes for all classes (healthy controls and CHD), indicating that we considered normal and abnormal sulcal patterns. The use of the prototype and corresponding similarity index for an abnormal case (CHD in this study) promoted the interpretability of the analysis. Specifically, the conventional method could not generate representative cases of abnormal sulcal folding patterns, which could be learned using the proposed framework. Furthermore, node attributes such as the sulcal pattern area, sulcal depth, and coordinates of the sulcal pit were considered with optimal weights via trainable parameters (e.g., U_1 and U_2 in Eq. (3)) in the node embedding networks instead of the conventional approach, which uses predefined weights based on prior knowledge. Beyond these features, future work exploring the potential of additional descriptive features will help improve expressivity of the proposed model.

6.2. Understanding the learned prototypes

We revealed that the similarities for prototypes 1 and 4 were highly correlated. Moreover, healthy control subjects with higher similarities to these prototypes were found to exhibit relatively high similarity to prototype 3, while showing low similarity to prototype 2. In contrast, CHD patients with higher similarities to these prototypes exhibited the opposite pattern (higher similarities to prototype 2, and lower similarities to prototype 3). Based on these results, we come to the hypothesis that these prototypes might represent intermediate-level atypical sulcal patterns between healthy controls and CHD patients. For example, CHD patients with high similarity to prototype 4 may not exhibit severe abnormalities in their sulcal patterns, and thus, may also show high similarity to prototype 1 (which represents the healthy control group). In contrast, they show relatively low similarity to prototype 3, which is highly representative of atypical sulcal patterns in CHD patients. Thus, even though these prototypes were observed to have relatively low similarities across both healthy control and CHD groups, they could contribute to more stable learning of the prototypes, allowing the model to account for residual patterns that are not covered by other representative prototypes (such as prototypes 2 and 3).

On the other hand, prototypes 2 and 3 showed a strong negative correlation, which seems to imply that two contrasting cortical folding morphologies might exist. In this case, an individual may tend to resemble one morphology while being dissimilar to the other, despite these patterns being located in distinct or distal regions of the cortex.

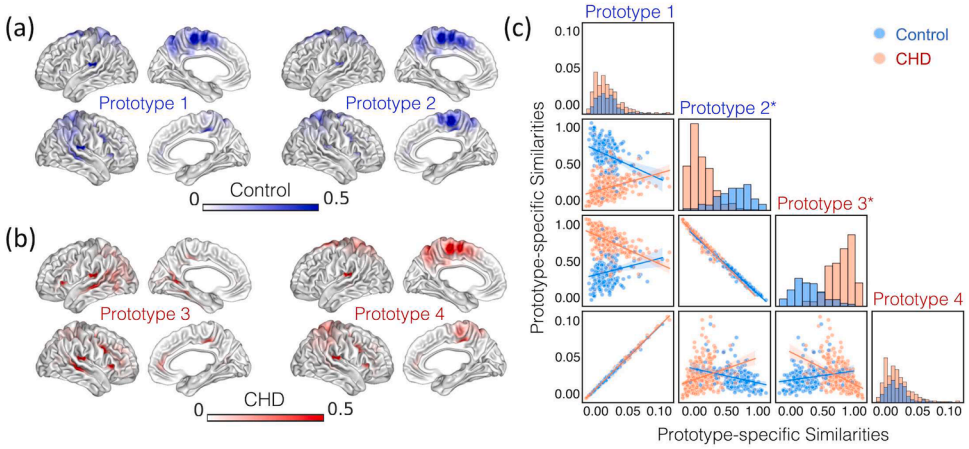

Fig. 9. Visualization of the learned prototypes (a-b) and the inter-prototypic relationships (c). The blue and red colors show the heatmaps of the inverse projection matrix for each prototype in healthy controls (a) and the congenital heart disease (CHD) class (b), respectively. (c) Diagonal elements showing the distribution of the prototype-specific similarities ($g_{p,\mathbf{z}_{\mathscr{G}}^{p}}$ for input graph $\mathscr{G}$ in Eq. (8)) for each prototype p. The blue and red dots denote the individuals in healthy controls and the CHD class, respectively, and the asterisk (*) indicates significantly different similarity distributions for the prototypes (Bonferroni corrected P value <0.05 after multiple comparison correction) between the classes according to the two-sample T test. Nondiagonal elements demonstrate the joint distribution for each prototype pair.

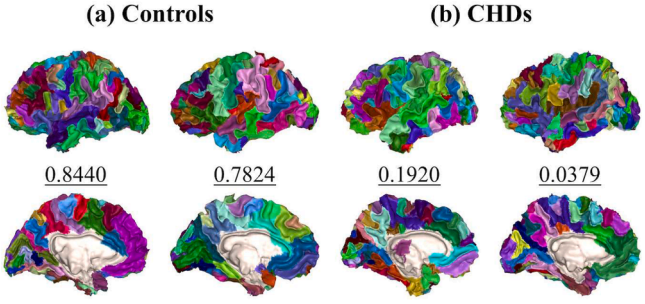

Fig. 10. Visualization of left-hemispheric cortical surfaces in sample individuals for (a) the healthy control group and (b) the CHD group. The sulcal catchment basins, derived from the sulcal depth map, are identified by different colors. The underlined numbers indicate the output probabilities of the healthy control group from the PIP, for each individual (each column). A lower probability value suggests that the corresponding sulcal pattern graph is more closely aligned with the CHD prototype, while a higher probability indicates stronger similarity to the healthy control prototype.

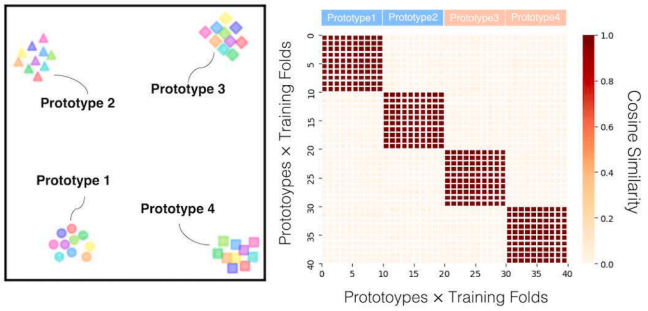

Fig. 11. Visualization of the learned prototypes for each training fold and their inter-prototypic similarity matrix. (left) Different colors denote different folds. (right) The inter-prototypic cosine similarity matrix for all prototypes and 10 training folds is illustrated. A darker color indicates greater similarity between corresponding prototype pairs.

However, the fundamental concept of the quantitative sulcal pattern analysis, including this study, is to detect subtle and global abnormalities in sulcal patterns that are hard to inspect visually. The biggest motivation behind this is that cortical areas develop interdependently, with their growth related to other functional areas through optimized white matter connections (O'Leary et al., 2007). This aspect of interdependent cortical regionalization during development provides the rationale for using a sulcal pattern graph, which is designed to characterize the global patterning of primary sulcal folds (Morton et al., 2020), rather than focusing on the well-defined and localized patterns such as the power button symbol in the precentral sulcus in type 2 focal cortical dysplasia (Mellerio et al., 2015). However, it is still worth investigating these patterns. Therefore, we visually inspected the sulcal patterns using surface models labeled with the corresponding sulcal catchment basins to evaluate if there were any visible symbols or landmarks. Unfortunately, no visually perceptible landmarks were found in our data, as shown in Fig. 10. This suggests that the proposed model takes into account not localized alterations, but rather subtle and complex global patterning, such as the interrelated arrangement of sulcal folds, to characterize the abnormal sulcal patterns in both disease and healthy control groups. Thus, the proposed approach has the potential to detect subtle and complex abnormalities in other developmental diseases that have not yet been clearly explored, even those diseases known to exhibit stochastic deviations from typical sulcal folding patterns without well-defined, class-specific atypical patterns.

6.3. *Effectiveness of the PIP approach*

Inverse-projection differs from the conventional prototype projection mechanism in two ways: 1) projection with opposite directions is involved, and 2) instead of graph pooling to obtain a graph-level representation, inverse-projection directly used node-level embeddings. Meanwhile, the prototype vectors were trained to include information about specific subgraph structures, representing the most frequent patterns for each class, similar to Ragno et al. (2022). Training is achieved by comparing the node representations in the inverse-projection layer, which encodes information from the L-hop neighboring subgraph centered at the node after passing through an L-layered GNN. Therefore, we illustrated four distinct latent spaces for the graph-level representation of the training samples for each prototype in Fig. 4. The training samples shown in Fig. 4 exhibit diverse patterns across the four prototype-specific subspaces, indicating that the prototypes learned to

encode information from different class-representative subgraph structures.

Existing prototype-based GNNs pose significant challenges for learning inherently interpretable prototypes because graphs are non-Euclidean data with irregular and arbitrary structures and various permutation-invariant nodes. For example, defining graph-structured prototypes is challenging, given the ambiguity in determining the number of nodes to be utilized. However, by utilizing inverse-projection weights, which indirectly visualize the corresponding prototypes, the proposed method addresses this problem without using laborious or costly processes, such as the Monte Carlo tree search algorithm (Zhang et al., 2022) or a graph autoencoder with strict regularization (Dai and Wang, 2022). Thus, PIP can achieve better interpretability by incorporating transparent graph-based prototypes.

Another benefit of the inverse-projection approach is that it focuses on more important regions. The spectral graph matching algorithm utilized in conventional SPA assumes equal contributions from all nodes (Im et al., 2011). Existing studies have considered this problem by replicating the SPA on parcellated sulcal graphs such as lobar areas (Morton et al., 2020, 2021). This approach does not directly handle the equal-contribution problem and can even deteriorate model sensitivity by decreasing the statistical power with more statistical tests. However, this study aggregates node representations using PIP weights, which can be considered an attention-based graph pooling mechanism. As the resulting graph representation containing information from the nodes exhibits high inverse-projection weights, the model can selectively consider the crucial nodes for each prototype. Moreover, the focal similarity strategy maximizes this advantage by emphasizing features from the selected regions while suppressing the background regions.

6.4. Differences with previous studies

Previous studies have utilized various geometric features to quantitatively analyze the sulcal pattern development. Multi-scale curvature maps were used to construct pairwise similarity matrices among healthy adults (Duan et al., 2019), and various geometrical features such as gyrification index, and sulcal area were compared between healthy fetuses and preterm newborns (Lefèvre et al., 2015). However, these approaches are inherently limited for analyzing the global patterning of primary sulci, since they do not take the interrelated arrangement and the topological relationships of the sulcal folds into consideration. Although the beta variational auto-encoder proposed by Guillon et al. (2024) can encode relevant characteristics of folding patterns in latent space, the authors were interested in focusing on the local folding patterns of right central sulcus area, rather than the global sulcal arrangement. Cortical areas develop interdependently, with their growth related to other functional areas through optimized white matter connections (O'Leary et al., 2007). This aspect of interdependent cortical regionalization during development provides the rationale for using sulcal pattern graph, which is designed to characterize the global patterning of primary sulcal folds (Morton et al., 2020).

Although the global sulcal pattern is important for analyzing brain development, defining precise anatomical correspondences and analyzing the sulcal shapes across different brains remains a significant challenge due to the complexity and high inter-individual variability of sulcal patterns. Ono's atlas was used to visually describe the geometric and topological relationships of sulcal folds by utilizing the connection and interruption patterns across local neighboring sulci (Kubik and Abernathey, 1990). More recently, we and other researchers have developed a comprehensive and quantitative analytic tool by utilizing the sulcal pattern graph with the sulcal pits and its catchment basins as graph nodes to consider not only the geometric features of sulcal folds themselves but also their intersulcal topological inter-related relationships (Im et al., 2017, 2011, 2013b; Meng et al., 2018; Morton et al., 2020; Ortinau et al., 2019; Tarui et al., 2018, 2023; Yadav et al., 2023). However, we mentioned the inherent limitations of this approach, including the absence of a standard criterion for defining a typical reference set, time-consuming cost of graph matching, user-defined feature weight sets, and assumptions about uniform node distribution, which make the analysis less generalizable. This is why we developed a new deep learning framework based on prototype-based GNNs in this study. Through prototype learning, the calculation of the affinity matrix, as well as the procedures for defining a typical dataset and feature weight set, are no longer required. Inverse-projection of the prototypes also enables the model to selectively consider the crucial part. Although Yadav et al. previously employed a GNN-based gender classifier on sulcal pattern graphs, PIP is the first to develop a self-explainable GNN that does not require any post-hoc explanatory model, to interpret how sulcal patterns influence the result conditions (e.g., disease or gender).

6.5. Multicenter cohorts

We included the multicenter cohorts to provide more extensive analyses in this study. The multicenter cohorts used in this study have been examined to detect sulcal pattern abnormalities in CHD, thereby supporting the relevance and quality of the dataset (Maleyeff et al., 2024b). It was proposed that association between the sulcal pattern similarities and magnetic field strength did not significant (Morton et al., 2020). Furthermore, presence and spatial distribution of the sulcal pits have revealed robustness to variability in scanners and scan sessions, supporting reliability of sulcal pit extraction (Im et al., 2013a). Based on these findings, our recent study on multicenter CHD cohorts did not adjust field strength for their sulcal pattern analysis (Maleyeff et al., 2024b). Our previous works on the different subset of the CHD cohorts used in this study showed that alterations in sulcal patterns were observed in whole brain regions as well (Morton et al., 2020, 2021). To dates, no distinct regions of abnormal sulcal patterns have been identified across the different CHD subtypes. Thus, we hypothesize that the site effect with distinct subset of CHD do not influence the analysis in this study. Indeed, no remarkable bias across different sites and scanners was observed in the training prototypes or the corresponding sample embeddings, further supporting the site reliability of our approach (Fig. S1).

In our previous study, nominally significant associations were found between sulcal pattern similarity and white matter volume in the SV-CHD cohort, which is also used in this study (Morton et al., 2020). Thus, we conducted a linear regression analysis to investigate whether total brain volume and white matter volume are associated with the output of the proposed approach. The probabilities for the healthy control class derived from the proposed model were calculated for each individual and used as a dependent variable in the linear regression analysis. Age, sex, group (CHD vs. healthy control), and volumetric measurements (total brain volume or white matter volume) were included as independent variables. In this analysis, no statistically significant correlations were found between the output probabilities and total brain volume ([regression coefficient $\beta\pm$ standard error (SE), $\times$ E-07]:-1.93 ± 0.50, P=0.70), and white matter volume ([$\beta\pm$SE, $\times$ E-07]:2.25 ± 1.49, P=0.13).

6.6. Limitations and conclusion

The proposed method has several limitations. First, the learned prototypes are visualized by showing a local region that exhibits a high inverse-projection weight. Thus, this approach enables us to identify

where the class-representation patterns appear in the sulcal pattern graph. However, it is still unclear what type of class-representative patterns (e.g., abnormal patterning, such as the positioning and arrangement of cortical folds in patients with CHD) were embedded in each prototype. Future studies should develop a method to provide further information on the prototypes for better interpretability. Second, because the proposed method is based on a fully supervised framework, adapting our method to other disorders requires additional training and hyperparameter tuning, which may be laborious and time-consuming. In future studies, we will explore a more generalized method by incorporating an unsupervised anomaly detection model to develop a pathology-agnostic sulcal pattern analysis framework. This approach may also help address the first limitation by providing a finer interpretation. For example, a graph reconstruction model (such as a graph autoencoder) trained on typical sulcal pattern graphs will learn a normative distribution of sulcal patterns, which will aid in providing a more refined interpretation of how and to what extent the node features align with the learned normative pattern. Thirdly, we utilized only the node attributes, treating their connections as binary edges. Thus the message-passing network used here is less capable of considering such complex information regarding edges. Nevertheless, we would like to highlight that the proposed framework is valuable for several reasons: (1) Firstly, there is a trade-off between the expressivity and model complexity. More features, such as the non-binary edge information, could encourage learning more expressive features in the presence of sufficient training data. However, the lack of training data might lead to overfitting and performance degradation. (2) Despite a growing number of efforts to leverage the edge features for graph representation learning (Bresson and Laurent, 2017; Gong and Cheng, 2019), message-passing strategies focused solely on node features are widely adopted in state-of-the-art GNNs, such as the benchmarkign GNNs used in our performance comparisons, offering advantages in terms of generalizability and versatility. Finally, training on larger and more balanced datasets may marginally improve the classification performance of the proposed model.

In conclusion, we proposed a prototype-based GNN for improved SPA utilizing PIP for better interpretability and more sensitive analysis. The proposed model surpasses existing state-of-the-art models in classification performance, and extensive experiments show that the model achieves diversity, transparency, and trustworthiness. We hope the proposed model will further boost research on SPA in the human cerebral cortex.

CRediT authorship contribution statement

Hyeokjin Kwon: Writing – original draft, Visualization, Software, Formal analysis, Conceptualization. **Seungyeon Son:** Methodology. **Sarah U. Morton:** Validation, Resources, Methodology, Data curation. **David Wypij:** Validation, Resources, Methodology, Data curation. **John Cleveland:** Validation, Resources, Methodology, Data curation. **Caitlin K Rollins:** Validation, Resources, Methodology, Data curation. **Hao Huang:** Validation, Resources, Methodology, Data curation. **Elizabeth Goldmuntz:** Validation, Resources, Methodology, Data curation. **Ashok Panigrahy:** Validation, Resources, Methodology, Data curation. **Nina H. Thomas:** Validation, Resources, Methodology, Data curation. **Wendy K. Chung:** Validation, Resources, Methodology, Data curation. **Evdokia Anagnostou:** Validation, Resources, Methodology, Data curation. **Ami Norris-Brilliant:** Validation, Resources, Methodology, Data curation. **Bruce D. Gelb:** Validation, Resources, Methodology, Data curation. **Patrick McQuillen:** Validation, Resources, Methodology, Data curation. **George A. Porter:** Validation, Resources, Methodology, Data curation. **Martin Tristani-Firouzi:** Validation, Resources, Methodology, Data curation. **Mark W. Russell:** Validation, Resources, Methodology, Data curation. **Amy E. Roberts:** Validation, Resources, Methodology, Data curation. **Jane W. Newburger:** Validation, Resources, Methodology, Data curation. **P. Ellen Grant:** Validation, Funding acquisition. **Jong-Min Lee:** Validation, Supervision, Methodology. **Kiho Im:** Writing – original draft, Validation, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Institute of Neurological Disorders and Stroke (R01NS114087), National Institute of Biomedical Imaging and Bioengineering (R01EB031170), National Heart, Lung, and Blood Institute (R01HL096825, R01HL135061-01, and P50HL74734), and Pediatric Cardiac Genomics Consortium (U01HL098188, U01HL098147, U01HL098153, U01HL098163, U01HL098123, and U01HL098162) of the National Institutes of Health and a grant of the Korea Dementia Research Project through the Korea Dementia Research Center (KDRC), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (RS-2020-KH106773). The acknowledgement for Dr. Duan Xu has not yet been included

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2025.103538.

Data availability

Datasets are available on request. The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

Alvarez Melis, D., Jaakkola, T., 2018. Towards robust interpretability with self-explaining neural networks. Adv. Neural Inf. Process. Syst. 31.
Asschenfeldt, B., Evald, L., Yun, H.J., Heiberg, J., Ostergaard, L., Grant, P.E., Hjortdal, V. E., Im, K., Eskildsen, S.F., 2021. Abnormal left-hemispheric sulcal patterns in adults with simple congenital heart defects repaired in childhood. J. Am. Heart Assoc. 10, e018580.
Bae, B.I., Tietjen, I., Atabay, K.D., Evrony, G.D., Johnson, M.B., Asare, E., Wang, P.P., Murayama, A.Y., Im, K., Lisgo, S.N., 2014. Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. Science 343, 764–768.
Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
Barkovich, A.J., Guerrini, R., Kuzniecky, R.I., Jackson, G.D., Dobyns, W.B., 2012. A developmental and genetic classification for malformations of cortical development: update 2012. Brain 135, 1348–1369.
Bécigneul, G., Ganea, O.E., Chen, B., Barzilay, R., Jaakkola, T.S., 2020. Optimal transport graph neural networks.
Boucher, M., Whitesides, S., Evans, A., 2009. Depth potential function for folding pattern representation, registration and analysis. Med. Image Anal. 13, 203–214.
Bresson, X., Laurent, T., 2017. Residual gated graph convnets. arXiv preprint arXiv: 1711.07553.
Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K., 2019. This looks like that: deep learning for interpretable image recognition. Adv. Neural Inf. Process. Syst. 32.
Chung, M.K., Robbins, S.M., Dalton, K.M., Davidson, R.J., Alexander, A.L., Evans, A.C., 2005. Cortical thickness analysis in autism with heat kernel smoothing. Neuroimage 25, 1256–1265.
Clouchoux, C., Du Plessis, A., Bouyssi-Kobar, M., Tworetzky, W., McElhinney, D., Brown, D., Gholipour, A., Kudelski, D., Warfield, S., McCarter, R., 2013. Delayed cortical development in fetuses with complex congenital heart disease. Cereb. Cortex 23, 2932–2943.
Corso, G., Cavalleri, L., Beaini, D., Liò, P., Veličković, P., 2020. Principal neighbourhood aggregation for graph nets. Adv. Neural Inf. Process. Syst. 33, 13260–13271.
Dai, E., Wang, S., 2022. Towards prototype-based self-explainable graph neural network. arXiv preprint arXiv:2210.01974.
Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage 9, 179–194.
Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. Adv. Neural Inf. Process. Syst. 29.

Duan, D., Xia, S., Rekik, I., Meng, Y., Wu, Z., Wang, L., Lin, W., Gilmore, J.H., Shen, D., Li, G., 2019. Exploring folding patterns of infant cerebral cortex based on multi-view curvature features: methods and applications. Neuroimage 185, 575–592.
Dwivedi, V.P., Joshi, C.K., Luu, A.T., Laurent, T., Bengio, Y., Bresson, X., 2023. Benchmarking graph neural networks. J. Mach. Learn. Res. 24, 1–48.
Evans, A.C., Group, B.D.C., 2006. The NIH MRI study of normal brain development. Neuroimage 30, 184–202.
Fischl, B., 2012. FreeSurfer. Neuroimage 62, 774–781.
Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. Neuroimage 9, 195–207.
Gasteiger, J., Bojchevski, A., Günnemann, S., 2018. Predict then propagate: graph neural networks meet personalized pagerank. arXiv preprint arXiv:1810.05997.
Gautam, S., Boubekki, A., Hansen, S., Salahuddin, S., Jenssen, R., Höhne, M., Kampffmeyer, M., 2022. Protovae: a trustworthy self-explainable prototypical variational model. Adv. Neural Inf. Process. Syst. 35, 17940–17952.
Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. In: Proceedings of the International Conference on Machine Learning. PMLR, pp. 1263–1272.
Gong, L., Cheng, Q., 2019. Exploiting edge features for graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9211–9219.
Guillon, L., Chavas, J., Bénézit, A., Moutard, M.L., Roca, P., Mellerio, C., Oppenheim, C., Rivière, D., Mangin, J.F., 2024. Identification of rare cortical folding patterns using unsupervised deep learning. Imaging Neurosci. 2, 1–27.
Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. 30.
Im, K., Guimaraes, A., Kim, Y., Cottrill, E., Gagoski, B., Rollins, C., Ortinau, C., Yang, E., Grant, P.E., 2017. Quantitative folding pattern analysis of early primary sulci in human fetuses with brain abnormalities. Am. J. Neuroradiol. 38, 1449–1455.
Im, K., Jo, H.J., Mangin, J.F., Evans, A.C., Kim, S.I., Lee, J.M., 2010. Spatial distribution of deep sulcal landmarks and hemispherical asymmetry on the cortical surface. Cereb. Cortex 20, 602–611.
Im, K., Lee, J.M., Jeon, S., Kim, J.H., Seo, S.W., Na, D.L., Grant, P.E., 2013a. Reliable identification of deep sulcal pits: the effects of scan session, scanner, and surface extraction tool. PLoS ONE 8, e53678.
Im, K., Pienaar, R., Lee, J.M., Seong, J.K., Choi, Y.Y., Lee, K.H., Grant, P.E., 2011. Quantitative comparison and analysis of sulcal patterns using sulcal graph matching: a twin study. Neuroimage 57, 1077–1086.
Im, K., Pienaar, R., Paldino, M.J., Gaab, N., Galaburda, A.M., Grant, P.E., 2013b. Quantification and discrimination of abnormal sulcal patterns in polymicrogyria. Cereb. Cortex 23, 3007–3015.
Im, K., Raschle, N.M., Smith, S.A., Ellen Grant, P., Gaab, N., 2016. Atypical sulcal pattern in children with developmental dyslexia and at-risk kindergarteners. Cereb. Cortex 26, 1138–1148.
Kang, E., Heo, D.W., Suk, H.I., 2022. Prototype learning of inter-network connectivity for ASD diagnosis and personalized analysis. In: Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2022. Singapore. Springer, pp. 334–343. September 18–22, 2022, Proceedings, Part III.
Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J. G., Hamarneh, G., 2017. BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage 146, 1038–1049.
Keswani, M., Ramakrishnan, S., Reddy, N., Balasubramanian, V.N., 2022. Proto2Proto: can you recognize the car, the way I do?. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10233–10243.
Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
Kubik, S., Abernathey, C.D., 1990. Atlas of the Cerebral Sulci. Thieme Medical Publishers.
Lefèvre, J., Germanaud, D., Dubois, J., Rousseau, F., de Macedo Santos, I., Angleys, H., Mangin, J.F., Hüppi, P.S., Girard, N., De Guio, F., 2015. Are developmental trajectories of cortical folding comparable between cross-sectional datasets of fetuses and preterm newborns? Cereb. Cortex 26, 3023–3035.
Leman, A., Weisfeiler, B., 1968. A reduction of a graph to a canonical form and an algebra arising during this reduction. Nauchno-Tech. Inform. 2, 12–16.
Lohmann, G., Von Cramon, D.Y., Colchester, A.C., 2008. Deep sulcal landmarks provide an organizing framework for human cortical folding. Cereb. Cortex 18, 1415–1420.
Lyttelton, O., Boucher, M., Robbins, S., Evans, A., 2007. An unbiased iterative group registration template for cortical surface analysis. Neuroimage 34, 1535–1544.
Maleyeff, L., Newburger, J.W., Wypij, D., Thomas, N.H., Anagnoustou, E., Brueckner, M., Chung, W.K., Cleveland, J., Cunningham, S., Gelb, B.D., Goldmuntz, E., Hagler Jr., D.J., Huang, H., King, E., McQuillen, P., Miller, T.A., Norris-Brilliant, A., Porter Jr., G.A., Roberts, A.E., Grant, P.E., Im, K., Morton, S.U., 2024a. Association of genetic and sulcal traits with executive function in congenital heart disease. Ann. Clin. Transl. Neurol. 11, 278–290.
Maleyeff, L., Park, H.J., Khazal, Z.S., Wypij, D., Rollins, C.K., Yun, H.J., Bellinger, D.C., Watson, C.G., Roberts, A.E., Newburger, J.W., 2024b. Meta-regression of sulcal patterns, clinical and environmental factors on neurodevelopmental outcomes in participants with multiple CHD types. Cereb. Cortex 34.
Marelli, A., Miller, S.P., Marino, B.S., Jefferson, A.L., Newburger, J.W., 2016. Brain in congenital heart disease across the lifespan: the cumulative burden of injury. Circulation 133, 1951–1962.
Mebius, M.J., Kooi, E.M., Bilardo, C.M., Bos, A.F., 2017. Brain injury and neurodevelopmental outcome in congenital heart disease: a systematic review. Pediatrics 140.
Mellerio, C., Roca, P., Chassoux, F., Danière, F., Cachia, A., Lion, S., Naggara, O., Devaux, B., Meder, J.F., Oppenheim, C., 2015. The power button sign: a newly described central sulcal pattern on surface rendering MR images of type 2 focal cortical dysplasia. Radiology 274, 500–507.
Meng, Y., Li, G., Wang, L., Lin, W., Gilmore, J.H., Shen, D., 2018. Discovering cortical sulcal folding patterns in neonates using large-scale dataset. Hum. Brain Mapp. 39, 3625–3635.
Morton, S.U., Maleyeff, L., Wypij, D., Yun, H.J., Newburger, J.W., Bellinger, D.C., Roberts, A.E., Rivkin, M.J., Seidman, J., Seidman, C.E., 2020. Abnormal left-hemispheric sulcal patterns correlate with neurodevelopmental outcomes in subjects with single ventricular congenital heart disease. Cereb. Cortex 30, 476–487.
Morton, S.U., Maleyeff, L., Wypij, D., Yun, H.J., Rollins, C.K., Watson, C.G., Newburger, J.W., Bellinger, D.C., Roberts, A.E., Rivkin, M.J., 2021. Abnormal right-hemispheric sulcal patterns correlate with executive function in adolescents with tetralogy of Fallot. Cereb. Cortex 31, 4670–4680.
Morton, S.U., Norris-Brilliant, A., Cunningham, S., King, E., Goldmuntz, E., Brueckner, M., Miller, T.A., Thomas, N.H., Liu, C., Adams, H.R., Bellinger, D.C., Cleveland, J., Cnota, J.F., Dale, A.M., Frommelt, M., Gelb, B.D., Grant, P.E., Goldberg, C.S., Huang, H., Kuperman, J.M., Li, J.S., McQuillen, P.S., Panigrahy, A., Porter Jr., G.A., Roberts, A.E., Russell, M.W., Seidman, C.E., Tivarus, M.E., Anagnoustou, E., Hagler Jr., D.J., Chung, W.K., Newburger, J.W., 2023. Association of potentially damaging De Novo gene variants with neurologic outcomes in congenital heart disease. JAMA Netw. Open 6, e2253191.
Nakamura, M., Nestor, P.G., McCarley, R.W., Levitt, J.J., Hsu, L., Kawashima, T., Niznikiewicz, M., Shenton, M.E., 2007. Altered orbitofrontal sulcogyral pattern in schizophrenia. Brain 130, 693–707.
Nauta, M., Van Bree, R., Seifert, C., 2021. Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14933–14943.
O'Leary, D.D., Chou, S.J., Sahara, S., 2007. Area patterning of the mammalian cortex. Neuron 56, 252–269.
Olshaker, H., Ber, R., Hoffman, D., Derazne, E., Achiron, R., Katorza, E., 2018. Volumetric brain MRI study in fetuses with congenital heart disease. Am. J. Neuroradiol. 39, 1164–1169.
Ortinau, C.M., Rollins, C.K., Gholipour, A., Yun, H.J., Marshall, M., Gagoski, B., Afacan, O., Friedman, K., Tworetzky, W., Warfield, S.K., 2019. Early-emerging sulcal patterns are atypical in fetuses with congenital heart disease. Cereb. Cortex 29, 3605–3616.
Ragno, A., La Rosa, B., Capobianco, R., 2022. Prototype-based interpretable graph neural networks. IEEE Trans. Artif. Intell.
Rakic, P., 1988. Specification of cerebral cortical areas. Science 241, 170–176.
Richter, F., Morton, S.U., Kim, S.W., Kitaygorodsky, A., Wasson, L.K., Chen, K.M., Zhou, J., Qi, H., Patel, N., DePalma, S.R., Parfenov, M., Homsy, J., Gorham, J.M., Manheimer, K.B., Velinder, M., Farrell, A., Marth, G., Schadt, E.E., Kaltman, J.R., Newburger, J.W., Giardini, A., Goldmuntz, E., Brueckner, M., Kim, R., Porter Jr., G. A., Bernstein, D., Chung, W.K., Srivastava, D., Tristani-Firouzi, M., Troyanskaya, O. G., Dickel, D.E., Shen, Y., Seidman, J.G., Seidman, C.E., Gelb, B.D., 2020. Genomic analyses implicate noncoding de novo variants in congenital heart disease. Nat. Genet. 52, 769–777.
Rymarczyk, D., Struski, Ł., Górszczak, M., Lewandowska, K., Tabor, J., Zieliński, B., 2022. Interpretable image classification with differentiable prototypes assignment. In: Proceedings of the 17th European Conference on Computer Vision–ECCV 2022. Tel Aviv, Israel. Springer, pp. 351–368. October 23–27, 2022, Proceedings, Part XII.
Rymarczyk, D., Struski, Ł., Tabor, J., Zieliński, B., 2021. Protopshare: prototypical parts sharing for similarity discovery in interpretable image classification. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1420–1430.
Tarui, T., Madan, N., Farhat, N., Kitano, R., Ceren Tanritanir, A., Graham, G., Gagoski, B., Craig, A., Rollins, C.K., Ortinau, C., 2018. Disorganized patterns of sulcal position in fetal brains with agenesis of corpus callosum. Cereb. Cortex 28, 3192–3203.
Tarui, T., Madan, N., Graham, G., Kitano, R., Akiyama, S., Takeoka, E., Reid, S., Yun, H. J., Craig, A., Samura, O., 2023. Comprehensive quantitative analyses of fetal magnetic resonance imaging in isolated cerebral ventriculomegaly. Neuroimage Clin. 37, 103357.
Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., 2012. The human connectome project: a data acquisition perspective. Neuroimage 62, 2222–2231.
Vasung, L., Yun, H.J., Feldman, H.A., Grant, P.E., Im, K., 2020. An atypical sulcal pattern in children with disorders of the corpus callosum and its relation to behavioral outcomes. Cereb. Cortex. 30, 4790–4799.
Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. Statistics 1050, 10.48550.
Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., Courty, N., 2022. Template based graph neural network with optimal transport distances. arXiv preprint arXiv: 2205.15733.
Wang, J., Liu, H., Wang, X., Jing, L., 2021. Interpretable image recognition by constructing transparent embedding space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 895–904.
Wernovsky, G., 2006. Current insights regarding neurological and developmental abnormalities in children and young adults with complex congenital cardiac disease. Cardiol. Young 16, 92–104.
Wilcoxon, F., 1992. Individual comparisons by ranking methods. Breakthroughs In Statistics: Methodology and Distribution. Springer, pp. 196–202.
Wright, R., Kyriakopoulou, V., Ledig, C., Rutherford, M.A., Hajnal, J.V., Rueckert, D., Aljabar, P., 2014. Automatic quantification of normal cortical folding patterns from fetal brain MRI. Neuroimage 91, 21–32.

Xu, K., Hu, W., Leskovec, J., Jegelka, S., 2018. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826.
Yadav, R., Dupé, F.X., Takerkart, S., Auzias, G., 2023. Population-wise labeling of sulcal graphs using multi-graph matching. PLoS ONE 18, e0293886.
Yadav, R., Dupé, F.X., Takerkart, S., Auzias, G., 2024. Geometric deep learning for sulcal graphs. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI 2024).
Zhang, Z., Liu, Q., Wang, H., Lu, C., Lee, C., 2022. Protgnn: towards self-explaining graph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9127–9135.
Zheng, H., Fu, J., Zha, Z.J., Luo, J., 2019. Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5012–5021.
Zilles, K., Armstrong, E., Schleicher, A., Kretschmann, H.J., 1988. The human pattern of gyrification in the cerebral cortex. Anat. Embryol. 179, 173–179.