Received: 17 December 2020 | Revised: 6 March 2021 | Accepted: 26 March 2021

DOI: 10.1002/hbm.25437

RESEARCH ARTICLE

WILEY

Reasoning network to predict language scores

Deep reasoning neural network analysis to predict language deficits from psychometry-driven DWI connectome of young children with persistent language concerns

Jeong-Won Jeong[1,2,3,4] | Soumyanil Banerjee[5] | Min-Hee Lee[1,4] | Nolan O'Hara[3,4] | Michael Behen[1,2,4] | Csaba Juhász[1,2,3,4] | Ming Dong[5]

[1]Departments of Pediatrics, Wayne State University, Detroit, Michigan

[2]Neurology, Wayne State University, Detroit, Michigan

[3]Translational Neuroscience Program, Wayne State University, Detroit, Michigan

[4]Translational Imaging Laboratory, Children's Hospital of Michigan, Detroit, Michigan

[5]Computer Science, Wayne State University, Detroit, Michigan

Correspondence

Jeong-Won Jeong, Pediatrics, Neurology, and Translational Neuroscience Program, Translational Imaging Laboratory, Children's Hospital of Michigan, Wayne State University. 3901 Beaubien St., Detroit, MI, 48201, USA.
Email: jjeong@med.wayne.edu

Funding information

National Institute of Neurological Disorders and Stroke, Grant/Award Number: R01NS089659; U.S. National Science Foundation, Grant/Award Number: CNS-1637312; National Science Foundation; National Institutes of Health; US

Abstract

This study investigated whether current state-of-the-art deep reasoning network analysis on psychometry-driven diffusion tractography connectome can accurately predict expressive and receptive language scores in a cohort of young children with persistent language concerns (n = 31, age: 4.25 ± 2.38 years). A dilated convolutional neural network combined with a relational network (dilated CNN + RN) was trained to reason the nonlinear relationship between “dilated CNN features of language network” and “clinically acquired language score”. Three-fold cross-validation was then used to compare the Pearson correlation and mean absolute error (MAE) between dilated CNN + RN-predicted and actual language scores. The dilated CNN + RN outperformed other methods providing the most significant correlation between predicted and actual scores (i.e., Pearson's R/p-value: 1.00/<.001 and .99/<.001 for expressive and receptive language scores, respectively) and yielding MAE: 0.28 and 0.28 for the same scores. The strength of the relationship suggests elevated probability in the prediction of both expressive and receptive language scores (i.e., 1.00 and 1.00, respectively). Specifically, sparse connectivity not only within the right precentral gyrus but also involving the right caudate had the strongest relationship between deficit in both the expressive and receptive language domains. Subsequent subgroup analyses inferred that the effectiveness of the dilated CNN + RN-based prediction of language score(s) was independent of time interval (between MRI and language assessment) and age of MRI, suggesting that the dilated CNN + RN using psychometry-driven diffusion tractography connectome may be useful for prediction of the presence of language disorder, and possibly provide a better understanding of the neurological mechanisms of language deficits in young children.

KEYWORDS

deep reasoning network, diffusion-weighted imaging connectome, language deficits, prediction of language score

1 | INTRODUCTION

Late language emergence (LLE) (Ellis Weismer, Murray-Branch, & Miller, 1994; Rescorla, 2000; Zubrick, Taylor, Rice, & Slegers, 2007) involves a delay in the acquisition of age-appropriate language skills. It is presumed to be a neurodevelopmental condition in which children evidence difficulties in the acquisition and/or use of language due to deficits in the comprehension or production of vocabulary, sentence structure, and/or discourse. The prevalence of LLE is estimated to be about 10–20% in 2-year-old children (Rescorla, 1989; Rescorla & Alley, 2001; Zubrick et al., 2007). Boys are three times more likely than girls to exhibit LLE (Roulstone, Loader, Northstone, & Beveridge, 2002). Approximately 50 to 70% of LLE children (Dale, Price, Bishop, & Plomin, 2003; Paul, Hernandez, Taylor, & Johnson, 1996) seem to be "late bloomers" (LB) who are reported to "catch up" to peers and ultimately demonstrate normal language development by late preschool age. However, about 30–50% of LLE children have more persistent deficits in language functions (language impairment or disorder, LI) that continue into at least late preschool age (Rice, Taylor, & Zubrick, 2008). The causes of LLE, and of LB and/or LI, in otherwise healthy children are not presently known.

Early discrimination of persisting LI, and perhaps type of language impairments (expressive vs. receptive or both), within the larger LLE group, especially during the early preschool age group (i.e., at age 2–3 years) would be crucial in launching and directing therapeutic interventions that might lead to improved language outcomes by late preschool/early elementary school ages. However, direct language assessments, using psychometrics tools, are often unreliable in children at such young ages due to the motivational and behavioral difficulties (especially those who have language delay), and the demands of these tools, which require engagement, motivation, effort/persistence (Downing & Perino, 1992; Sattler, 2001); the etiologic yield of existing evaluation strategies is low and highly variable, ranging from 17 to 34% (Shevell et al., 2003). In addition, the results from current clinical MRI protocols in this age group are mostly unremarkable and of limited value in discriminating LLE from LI, except to rule out a lesional etiology (Bishop, Snowling, Thompson, Greenhalgh, & the CATALISE-2 consortium, 2017; Nelson, Nygren, Walker, & Panoscha, 2006). In the present study, we are addressing the lack of early, accurate diagnostic tools that allow for the early discrimination of LLE from LI, by developing and testing develop an imaging protocol that may provide such critical discriminations, and, again, with the hopes that such early discriminations might allow for early, targeted intervention programs for children with persisting LI. Early and reliable imaging signatures would catalyze a revolution in the early diagnostic and targeted intervention. We report on a comprehensive analysis of diffusion-weighted imaging connectome (DWIC), which may allow for early and reliable prediction of preschool/school-aged language functions via advanced deep learning methods and boost the prediction of persisting language impairment.

For the last decade, there has been a series of neuroimaging studies (Badcock, Bishop, Hardiman, Barry, & Watkins, 2012; Chai, Mattar, Blank, Fedorenko, & Bassett, 2016; Verly et al., 2019) that have utilized MRI protocols to provide objective insights into persistent LI by elucidating neural changes that impair development and/or execution of age-appropriate levels of language abilities. Our recent MRI studies (Jeong, Sundaram, Behen, & Chugani, 2016a; Jeong, Sundaram, Behen, & Chugani, 2016b; Sundaram, Sivaswamy, Makki, Behen, & Chugani, 2008) have revealed brain network abnormalities in children with LI and further suggested associations between such abnormalities and type, and potentially magnitude, of LI as well. For instance, distinct cortico-subcortical network abnormalities, identified using whole-brain connectome analysis (Jeong, Sundaram, et al., 2016a), and involving a frontotemporal language network, differentiated children with LI from healthy controls (Lee, O'Hara, Behen, & Jeong, 2020), and also were differentially associated with distinct LI phenotypes (Lee et al., 2020). This included the discrimination of children with LLE from those with more persisting impairments, and also the identification of impairment in specific receptive or expressive skills. Also, emerging data including our recent studies (Jeong, Sundaram, et al., 2016a; Lee et al., 2020; Sundaram et al., 2008) suggest that computational network analysis of DWIC may help noninvasively evaluate atypical alteration in the geometry of axonal connectivity, possibly improving the predictability of the therapeutic outcome. These findings have driven us to test the hypothesis that probes two distinct language networks: expressive and receptive networks, which may be accurately delineated by a psychometry-based DWIC (pDWIC) (Lee et al., 2020); such networks seem to coincide generally well with clinical diagnostic conceptions of language impairment as well (e.g., international classification of diseases, tenth revision). In pDWIC, each language network, Ω = (R, A), is defined as a collection of L-nodes, $R_{m = n = 1\text{-}L}$ (representing L-brain regions) interconnected by edges, $A_{m,n = 1\text{-}L}$, representing pair-wise white matter pathways of which counts of axonal tract streamlines are significantly correlated with language function assessed by psychometry-derived language scores (Semel, Wiig, & Secord, 2006). Graphic network topology (Jeong, Asano, Juhasz, Behen, & Chugani, 2016; Rubinov & Sporns, 2010) has been applied to characterize not only the nodes and edges in terms of their embedding into the network but also network organization as a whole to predict imaging-based language scores, depending on the severity of the pruning disorganization and(or) myelination abnormality in the language network(s) that may provide useful diagnostic and prognostic information for individual children with LLE.

The present study investigates a cutting-edge pDWIC approach for young children to identify very early imaging markers of neuronal disorganization, helping identify persisting LI and ultimately directing therapeutic interventions that could facilitate improved language outcomes. Our working hypothesis is that the location and severity of brain abnormality identified, quantified, and confirmed by pDWIC analysis of two language networks (that are suspected to be differentially related to/involved in expressive and receptive language functions), will allow (a) accurate imaging-based prediction of persisting expressive and receptive language impairments for early targeted-focused intervention, and (b) detailed concepts for understanding neural substrates of types of language impairments in young children evidencing persistent language concerns.

To determine whether the pDWIC imaging features of two language networks: $\Omega_{\text{expressive}}$ and $\Omega_{\text{receptive}}$ can accurately predict later expressive and expressive language impairment/scores of in individual children with early LLE, the present study performed the prediction tasks with an end-to-end deep learning network, called "dilated convolution neural network (CNN) with a relational network (RN) (Banerjee et al., 2020; Banerjee et al., 2021; Santoro et al., 2017), that is a special graph network whose computations explicitly focus on relational reasoning.

CNN is one of the most powerful deep learning models and has been widely used to extract multi-scale features from input data. Specifically, we use CNN to map the connectome matrix, $A_{m,n}$, into a set of features, which are modeled as objects in the RN. Then, the combination of CNN + RN could reason about the relationship between remotely located edges in $A_{m,n}$, thus leading to a more accurate prediction of expressive and receptive scores for individual patients. Since $A_{m,n}$ is often sparse, the present study proposes a dilated CNN (Yu & Koltun, 2015) in the CNN + RN architecture. The dilated CNN has a larger receptive field than a standard one. Hence, when combined with RN (i.e., dilated CNN + RN), it is supposed to reduce the prediction error caused by sparse inputs. This study systematically investigates (a) if the dilated CNN + RN outperforms other advanced machine learning methods to predict the presence and type of persisting expressive and receptive language impairment/scores, and (b) identify which features of pDWIC are most predictive of the expressive and receptive language impairment/scores in the same study cohort, thus providing additional support that the dilated CNN + RN can effectively improve our understanding of the neuroanatomical substrates associated with specific, persistent language impairments in young children, including LLE.

2 | MATERIALS AND METHODS

2.1 | Participants

The present study included 31 children with persistent language concerns (mean ± *SD* of MRI age: 4.25 ± 2.38 years, 1.8–13.6 years old, 20 males) reported in our previous work (Lee et al., 2020). Briefly, all 31 children were seen by their neurologists at the Children's Hospital of Michigan Pediatric Neurology Clinics from 2011 to 2017. All had been referred for comprehensive developmental and behavioral evaluations, including assessment of cognitive, motor, language functions, and behavioral concerns—the specific battery of measures is described previously (Lee et al., 2020). The children also underwent clinical MRI before psychological assessment with a time interval of 2.7 ± 1.8 years. Receptive and expressive language scores were obtained using the age-appropriate version Comprehensive Evaluation of Language Fundamentals (CELF-P2 for preschool-age children, and CELF-4 for 5–6-year-old children)—raw scores on the receptive and expressive subscales were standardized to normative t-scores (mean = 50; *SD* = 10); also mean verbal and nonverbal intelligence indices (acquired with the Wechsler Preprimary and Preschool Scales of Intelligence, WPPSI-4) were also standardized to t-scores. Measured nonverbal intellectual scores for the study group were 41.4 ± 5.8 (34–60). External and internal behavior scores for the group were 60.64 ± 9.80 (50–82) and 60.90 ± 8.04 (50–80), respectively. Expressive and receptive language scores for the group were 33.0 ± 10.1 (20–56) and 29.2 ± 7.3 (20–48.7), respectively. These scores were used as ground truth data to define the pDWIC language network, $\Omega_{\text{expressive}}$, and $\Omega_{\text{receptive}}$, and train our predictive models. The present study excluded children with a history of epilepsy or atypical/complex febrile seizures, a diagnosis of an autism spectrum disorder and/or attention deficits, and early suboptimal care/maltreatment, including neglect or exposure to abuse. The present study was approved by the Wayne State University's Institutional Review Board with a waiver of informed consent.

2.2 | MRI acquisition

All MRI scans were acquired at a 3 T GE-Signa scanner (GE Healthcare, Milwaukee, WI) equipped with an eight-channel head coil and ASSET. DWI was acquired with a multi-slice single-shot diffusion-weighted echo-planar imaging sequence at TR = 12,500 ms, TE = 88.7 ms, FOV = 24 cm, 128 × 128 acquisition matrix, contiguous 3 mm slice thickness, 55 isotropic gradient directions with b = 1,000 s/mm^2, one b = 0 acquisition, and NEX = 1. T1-weighted structural images were also acquired using a 3D fast spoiled gradient-echo sequence at TR/TE/TI of 9.12/3.66/400 ms, with a section thickness of 1.2 mm and a planar resolution of 0.94x0.94 mm^2. Before performing the DWIC tractography analysis, NIH TORTOISE (https://tortoise.nibib.nih.gov/tortoise) and FSL top-up packages (Andersson, Skare, & Ashburner, 2003) were used to correct motion, noise, physiological artifacts, susceptibility-induced distortion, and eddy current-induced distortion in the DWI data.

All MRI data of the present study were acquired for clinical diagnosis. A multidisciplinary team (nurse, child life specialist, MRI technicians) worked to improve the image quality and minimize motion artifacts by allowing longer scan time for multiple trials. To minimize the potential confound from head motion artifact, the present study excluded patients with unsuccessful MRI showing head motion ≥2 mm in DWI encoding data (i.e., voxel size of DWI image), which was evaluated by NIH TORTOISE DWI motion artifact correction package.

2.3 | Preliminary assessments of structural morphology

T1-weighted structural images for all 31 patients were read as normal clinically. To ensure that subclinical variations in brain morphology would not significantly influence the present study's investigation of structural connectivity measures, an additional preliminary assessment was completed using FreeSurfer software (Fischl, 2012). 3D reconstructions of each patient's brain, along with equivalent reconstructions from 16 typically-developing controls (mean ± *SD* of MRI age: 7.26 ± 3.42 years, 2–13 years, 9 males) reported in previous work (Lee et al., 2020), were resampled to common template space. Vertex-wise cortical thickness values were spatially smoothed with a 10 mm full width at half maximum Gaussian kernel and fit to a general linear model with age and sex as covariates. Clusters of thickness contrast between patients and controls were corrected for multiple comparisons by Monte Carlo simulation (Hagler, Saygin, & Sereno, 2006), ultimately revealing no significant clusters of difference between the groups. Global volumetric measures of left and right cortical gray matter, left and right cerebral white matter, and subcortical gray matter, were similarly fit to a general linear model, with patient versus control as a response variable and with age and sex as covariates. Standardized coefficients for these five volumetric measures were also not significant, with associated t-score = −0.84/−1.55/1.59/−1.33/−0.90 and p-value = .41/.13/.12/.19/.37. Taken together, these findings suggest that the 31 patient brains included in the present study are morphologically normal and further highlight the need for advanced methods targeting DWI structural connectivity measures in language disorder.

2.4 | Construction of pDWIC-based language network

For each of the 31 patients, the second-order integration over fiber orientation distribution (iFOD2) (Tounier, Calamante, & Connelly, 2010) was applied to help address the intra-voxel crossing fiber problem in a whole-brain DWI tractography. Automated anatomical labeling (AAL) parcellation atlas (Tzourio-Mazoyer et al., 2002) consisting of a set of 116 nodes, $R_{i\,=\,j\,=\,1–116}$, was then used to reconstruct whole-brain network, Ω = ($R_{i\,=\,j\,=\,1–116}$, $A_{i,j\,=\,1–116}$), where the elements of edges $A_{i,j}$ quantify the pair-wise connectivity strengths between R_i and R_j (i.e., the number of fiber streamlines scaled by both streamline length and volume of the nodes to stabilize inter-subject variability by correcting for age Cheng et al., 2012). Finally, Pearson's correlation with Bonferroni correction was performed between a subject-series of the edges, $A_{i,j\,=\,1–116}$, and CELF expressive (receptive) scores. The corrected *p*-value <.05 was applied to identify $A_{m,n\,=\,1\text{-}L\text{expressive(receptive)}}$ of $L_{expressive(receptive)}$-nodes, $R_{m\,=\,n\,=\,1\text{-}L\text{expressive (receptive)}}$, consisting of the pDWIC-based expressive (receptive) language network (Lee et al., 2020), $\Omega_{expressive(receptive)}$ = ($R_{m\,=\,n\,=\,1\text{-}L\text{expressive(receptive)}}$, $A_{m,n\,=\,1\text{-}L\mathit{expressive(receptive)}}$).

2.5 | Construction of the dilated CNN + RN

The present study utilized the dilated CNN + RN (Banerjee et al., 2020; Banerjee et al., 2021; Santoro et al., 2017) to objectively predict language scores, where the dilated CNN maps the sparse input matrix, $A_{m,n}$, into a set of features, which are modeled as objects in the RN to reason about the relationship between objects, leading to a more accurate prediction of an output: expressive or receptive language score for individual patients (Figure 1). Briefly, the dilated CNN is a generalized CNN operator to provide exponential expansion of the receptive field without resolution loss. The dilated convolution (Yu & Koltun, 2015) between input: $A_{m,n}$ and CNN kernel: c with dilation factor: l, is defined as:

$$(c_{m,n} *_l A_{m,n})_{i,j} = \sum_i \sum_j c_{i,j} \cdot A_{m-l*i,n-l*j} \tag{1}$$

where the size of the receptive field, the block of elements that determine the activation of each CNN unit, is exponentially

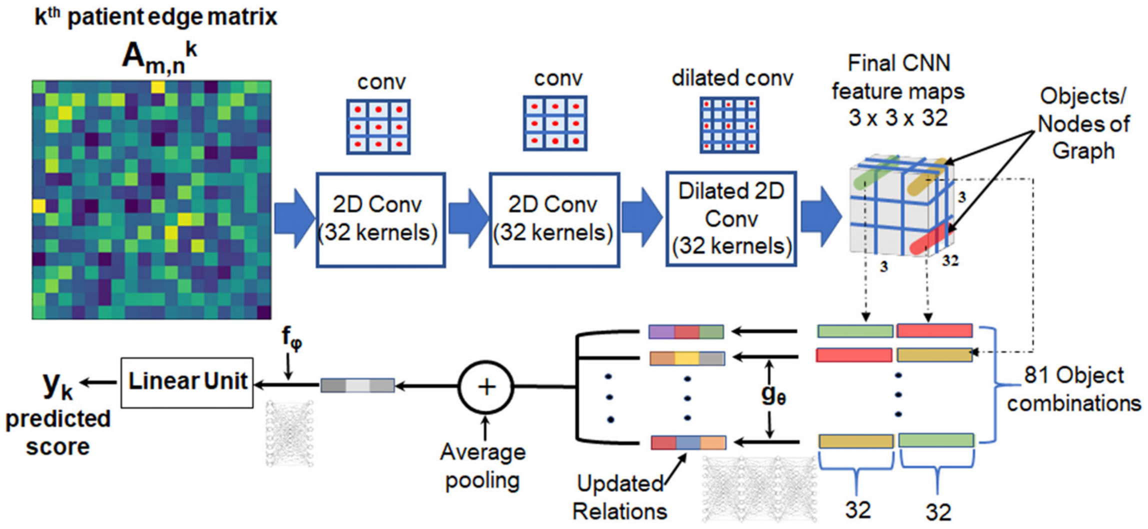

FIGURE 1 An architecture diagram of the proposed dilated CNN + RN which takes a given input: $A_{m,n}^k$ to predict an output: y_k where k is an index of individual subject (i.e., instance). For CELF expressive(receptive) language score, an input: $A_{m,n}^k$ will be a 2D matrix of an expressive (receptive) network features ($L_{expressive(receptive)}$ × $L_{expressive(receptive)}$) while an output: y_k at a linear layer will be a continuous variable ranging from 0 to 50. Two dilated CNN + RNs to learn two sets of $A_{m,n}^k$ from $\Omega_{expressive}$ and $\Omega_{receptive}$ were separately constructed to predict expressive and receptive language scores

increased by the dilation factor, l (i.e., the space between original kernel elements).

It should be noted that the dilation increases the size of the receptive field exponentially at each depth while keeping the number of parameters with a logarithmic growth at each depth, achieving 3.6 times parameter reduction with only a 1% drop of accuracy (Zhou, Wu, Wu, & Zhou, 2016). Thus, compared with conventional convolution (l = 1), the dilated convolution (l > 1) provides more efficient learning at the same computational cost, especially in cases that the input matrix, $A_{m,n}$ has the sparse clusters of edges connecting the remote nodes in the pDIWC-based language network, $\Omega_{expressive(receptive)}$. Furthermore, the spatial arrangement of nodes, $R_{m,n}$ in $\Omega_{expressive\ (receptive)}$, is arbitrarily determined by a fixed cortical atlas (e.g., AAL). This configuration makes local feature patterns of the dilated CNN maps ill-posed to represent specific functional relevance in $\Omega_{expressive\ (receptive)}$. We presume that this limitation can be overcome by combining the dilated CNN with RN, "dilated CNN + RN," which is capable of modeling nonlocal relations among the features extracted by the dilated CNN. In our RN, relational reasoning is performed between every pair of possible objects (i.e., dilated CNN features) regardless of their spatial arrangements in $\Omega_{expressive(receptive)}$. That is, for accurate prediction of language score: y_k, the function: g is applied on each object combination to calculate the relation of every object pair (Banerjee et al., 2020; Banerjee et al., 2021; Santoro et al., 2017).

$$y_k = \mathrm{RN}(O) = f_\varphi\left(1/N\ \Sigma_{i,j}\ g_\theta(o_i, o_j)\right) \tag{2}$$

where o_i and o_j are a possible object pair obtained from the feature maps and O is the set of all objects. The function: g (i.e., four fully connected [FC] layers, 256 units per layer) is an FC network that operates on these object pairs and computes relations between them. The function: f (i.e., two FC layers, 256 units per layer), is another FC network, which operates on the averaged set of relations and predicts the final score. N is the total number of object combinations obtained from the final feature maps. y_k is the final prediction of the network. g_θ corresponds to the update function f_φ for the global attribute.

The function of $g_\theta(o_i,o_j)$ in Equation (2) provides information about the strength of the relationship between each object pair or if the objects are not related. The resulting set of relations is averaged element-wise to ensure combinatorial generalization. Finally, the function: f is applied to this combined vector. The final layer following f is a linear layer with one unit which is required for the language score prediction. An adaptive learning rate approach for stochastic gradient descent using the Adam optimizer (Kingma & Ba, 2014) minimizes the loss function (i.e., the mean square error for language score prediction). Detailed architecture and parameters of the dilated CNN + RN were available in our recent work (Banerjee et al., 2020; Banerjee et al., 2021).

2.6 | Data augmentation

For each of the expressive and receptive language score predictions, threefold cross-validation was performed with two independent cohorts, (a) training cohort (twofolds) and (b) test cohort (onefold), which is independent of the training cohort. Three folds, first fold: patient index 1–10 (age of MRI: 4.14 ± 2.17), second fold: patient index 11–20 (age of MRI: 4.80 ± 3.40), third fold: patient index 21–31 (age of MRI: 3.93 ± 1.24), did not differ at age ($p > .38$). To prevent the overfitting of network layers, the study samples of each cohort were separately augmented by using the synthetic minority oversampling technique (SMOTE) (Brown et al., 2015; Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Hussain, Gimenerz, Yi, & Rubin, 2017; Kawahara et al., 2017). That is, for each cohort, 311 sample instances per patient (i.e., 1 original sample and 310 augmented instances) were generated from a 2-D data matrix in which each row vector constitutes a ($A_{m,n}^k$,t_k) of the kth study subject, sized by a 1× (L^2 + 1) vector consisting of L^2 elements of $A_{m,n}^k$ and a scalar t_k (i.e., actual measurement: CELF score). Then each row vector of this data matrix was augmented 310 times by randomly interpolating its six nearest neighbors of t_k. These augmentations resulted in (a) 6,531 sample instances of the training cohort sized by 6,531 × (L^2 + 1) (i.e., 21 original instances and 6,510 augmented instances) and (b) 3,110 sample instances of the test cohort sized by 3,110 × (L^2 + 1) (i.e., 10 original instances and 3,100 augmented instances). In this manner, we performed a threefold cross-validation over the entire cohort data set (n = 31 patients), first fold was to predict the language scores of patient index: 1–10 using the training cohort of patient index: 11–31, second fold to predict the language sores of patient index: 11–20 using the training cohort of patient index: 1–10 and 21–31, and third fold to predict the language scores of patient index: 21–31 using the training cohort of patient index: 1–20. This validation comprised of 9,641 (31 original +310 × 31 augmented) instances dived into 6,531 training and 3,110 testing instances (i.e., 67.7 and 32.3% of the entire cohort, respectively).

2.7 | Identification of the most predictive features for expressive and receptive language score

To identify the most essential elements of the edge matrix, $A_{m,n}$, for prediction of language scores, we generated the regression activation maps using a gradient-based algorithm, called gradient-based class activation mapping (Grad-CAM) (Selvaraju et al., 2017; Simonyan, Vedaldi, & Zisserman, 2013; Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016). In Grad-CAM, the weights of the gradients, α_k are first calculated by taking the global average pooling of the partial derivatives of output score w.r.t the set of feature maps,

$$\alpha_k^l = \frac{1}{Z}\sum_m\sum_n \frac{\partial y_k}{\partial F^l_{mn}} \tag{3}$$

where y_k is the score of the kth study subject. F^l is the lth feature map after the last layer of convolution, and Z is the normalization factor. The final activation map $M_{\text{Grad-CAM}}$ for the kth patient is given as follows:

$$M^{k}_{Grad-CAM} = RELU\left(\sum_{l} \alpha^{l}_{k} F^{l}\right) \tag{4}$$

where the RELU operation removes the negative gradients.

2.8 | Performance metric and comparison with other methods

The mean absolute error (MAE), between the dilated CNN + RN-driven language score: y_k and the measured language score: t_k was calculated to evaluate the performance of the dilated CNN + RN. Also, we evaluated P(AE≤10) as a probability value if absolute error (AE) between y_k and t_k is less than 10 (i.e., *SD* of normative t-score) and also calculated Pearson's correlation coefficient (R) between y_k and t_k.

Intensive computational experiments were performed to compare the dilated CNN + RN model with current state-of-the-art models used for the prediction of language scores. In the same training and testing splits, we compared the prediction performances of the following models: (a) support vector regressor (SVR) with a radial basis function; (b) multilayer regressor (MLR): this constituted a three-FC layer network with 256 units in the first two layers with 50% dropout followed by 128 units in the third layer and a single linear unit in the final layer for prediction of expressive/receptive scores; (c) CNN + MLR: this constituted three convolutional-layers where each layer of CNN had 32 kernels of size 3 × 3 with stride 1 and ReLU activation. Maximum pooling with pool size 2 was used after the second layer, followed by the same MLR used in (b); (d) CNN + RN: three CNN layers as in (c) with maximum pooling after the first and second layer and Leaky ReLU activation with slope 0.2, were combined with the RN as presented in Figure 1; (e) dilated CNN + MLR: three convolutional layers with 32 kernels in each layer where the last convolutional layer has the kernel with a dilation factor of 2 and followed by the MLR used in (b).

3 | RESULTS

Nonverbal intelligence was not significantly correlated with either expressive (*R* = .33, *p*-value = .07) or receptive scores (*R* = .19, *p*-value = .31). In addition, neither externalizing behavior nor internalizing behavioral scores were significantly related to either expressive (*R* = .03, *p*-value = .88, *R* = .15, *p*-value =.43, respectively) or receptive scores (*R* = –.26, *p*-value = .16, *R* = .05, *p*-value = .80, respectively); therefore, neither nonverbal intelligence, nor behavioral metrics were included in further analyses.

Figure 2 shows the pDWIC-based expressive and receptive language networks, $\Omega_{expressive} = (R_{m\ =\ n\ =\ 1\text{-}17}, A_{m,n\ =\ 1–17})$ and $\Omega_{receptive} = (R_{m\ =\ n\ =\ 1\text{-}17}, A_{m,n\ =\ 1–17})$, where Pearson's correlation analysis between language scores and $A_{i,j\ =\ 1–116}$ of Ω determined $L_{expressive} = 17$ and $L_{receptive} = 17$ at the corrected *p*-value <.05.

The output after the final dilated CNN layer was 32 feature maps (size: 3 × 3 × 32). An object was sliced from the third dimension of the feature maps as shown in Figure 1. Thus, each object was of size 1 × 1 × 32 or 32 dimensional. Then, every pair of objects were concatenated and fed as an input to the RN. For feature maps of size 3 × 3, the total number of object combinations was 81 (3 × 3 × 3 × 3), and the size of each combination of an object pair was 64 (32 for each object). The function g of Equation (2) was then applied to each object combination to calculate the relation between every object pair in the dilated CNN + RN.

Figure 3 shows the convergence curves of the dilated CNN + RN to predict expressive and receptive language scores. Compared with other methods, the dilated CNN + RN to predict expressive/receptive scores provided the fastest convergence to global minima at 100 epochs, 0.0013/0.0008, which was 5.33/8.58, 3.86/6.16, 1.82/1.72, 2.52/2.13, 1.06/1.20 times smaller than the loss of SVR, MLR, CNN + MLR, dilated CNN + MLR, and CNN + RN at 100 epochs, indicating that the deep reasoning of high-level features abstracted from the dilated convolution layer are the most powerful to minimize prediction error in both expressive and language scores. Also, computational time per epoch was 0.05, 0.13, 0.20, 0.20, 1.0, and 1.0 second on a NVIDIA GTX 980 Ti GPU for SVR, MLR, CNN + MLR, dilated CNN + MLR, CNN + RN, and dilated CNN + RN, respectively.

Table 1 presents performances of dilated CNN + RN and other models for expressive and receptive language score prediction which was evaluated from the 9,641 instances of threefold cross validation. The dilated CNN + RN yielded the best performance compared with other models. MAE, P(AE≤10), and R of the dilated CNN + RN outperformed those of other methods when predicting expressive/receptive scores, yielding respective improvements in MAE, P(AE≤10), and R: 48.5 ± 28.26 (17.5–75.20)%/54.50 ± 20.38 (16.37–86.10)%, 0.42 ± 0.94 (0–2.11)%/0.61 ± 0.91 (0–2.04)%, 3.38 ± 4.63 (1.05–11.63)%/8.36 ± 15.26 (1.02–35.62)% (i.e., improvement [%] = 100 × |dilated CNN + RN—other method|/other method). These superior performances yielded >97% of P(AE≤10) to predict both expressive and receptive scores, suggesting that the dilated CNN + RN can provide a reliable means to predict language impairment and ultimately may help supplement or replace portions of early language assessments in clinical cases, which are often unreliable due to motivational and behavioral concerns, commonly present in small children, and particularly in small children with early language problems.

Also, compared with the CNN + RN, the dilated CNN + RN provided lower MAE values and higher correlation coefficients between actual and predicted language scores of three test folds (Figure 4, MAE/R/*p*-value of dilated CNN + RN = 0.278/0.999/<0.001 and 0.281/0.987/<0.001 for expressive and receptive scores, MAE/R/*p*-value of CNN + RN = 0.950/0.989/<0.001 and 0.704/0.980/<0.001 for expressive and receptive scores), yielding the best performance to predict expressive and receptive scores in three folds of 31 test subjects (*n* = 10, 10, and 11 test patients from first fold, second fold, and third fold, respectively). Note that we stacked the test cohorts of threefolds and generate 31 patients instead of overlapping the train and test patients. There was no overlap between train and test

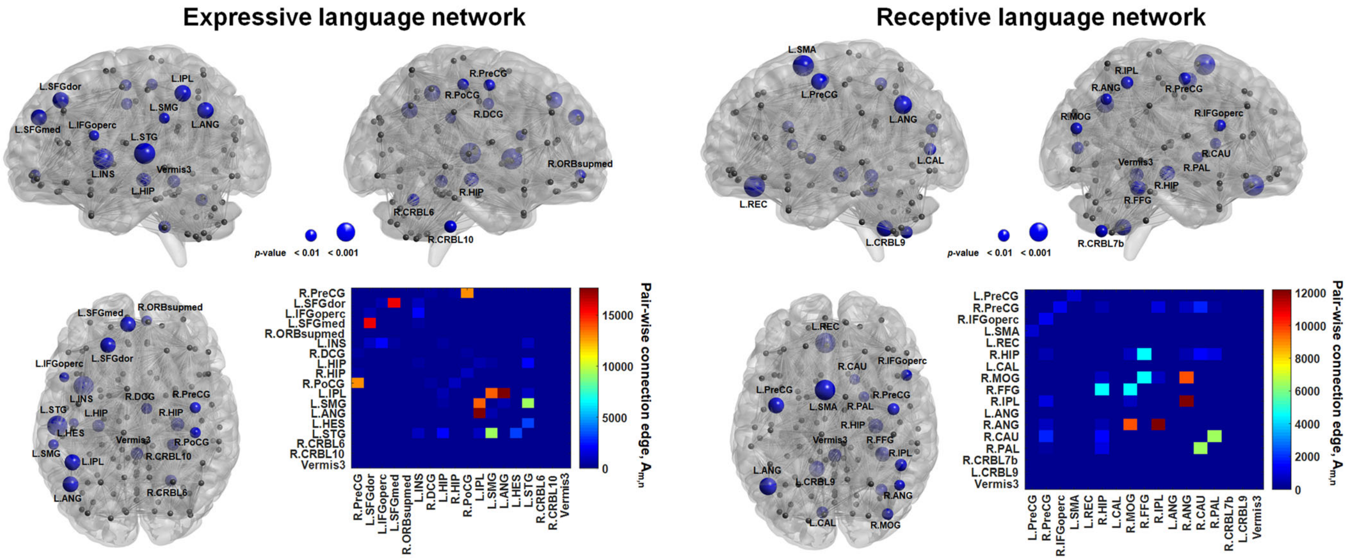

FIGURE 2 pDWIC-based language network, (left) expressive language network, $\Omega_{expressive}$. (right) receptive language network, $\Omega_{receptive}$, which were obtained from 31 young children (age: 4.25 ± 2.38 years, 20 male). Blue spheres indicate 17 nodes of $A_{m,n}$ showing statistically significant Pearson's correlation coefficient with CELF score at the corrected p-value <.05. Black spheres indicate the other 99 nodes of $A_{m,n}$ having no significant correlations. Each 2-D matrix shows pair-wise connection edges, $A_{m,n\,=\,1\text{–}17}$ at each language network. A complete list of regional names corresponding to node labels is available in Table S1

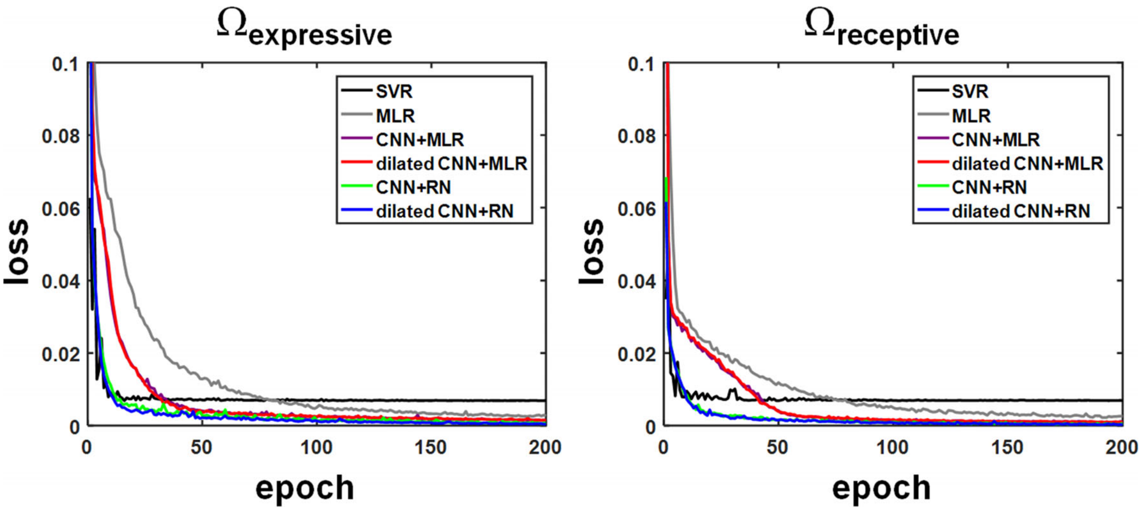

FIGURE 3 Convergence curves of the dilated CNN + RN (solid blue line) obtained from the augmented training instances, (left) expressive language network: $\Omega_{expressive}$, and (right) receptive language network: $\Omega_{receptive}$

patients as the 31 patient's predictions have been generated by stacking first fold (n = 10 test patients), second fold (n = 10 test patients) and third fold (n = 11 test patients).

Subsequent subgroup analyses of the stacked 31 test predictions (Table 2) found no significant interference of Δt (time interval between MRI and language assessment) and age of MRI on the absolute error of the dilated CNN + RN (i.e., unpaired t test p-value of AE in expressive/receptive language score = 0.268/0.503 and 0.695/0.402 for Δt ≤3 years vs. Δt > 3 years and age of MRI ≤3 years vs. age of MRI > 3 years), suggesting the stability of the dilated CNN + RN-based prediction without depending on different patient profiles, especially for children with LLE and younger than 3 years of age.

Figure 5 presents the activation maps of Equation (4) showing important pair-wise edges, $A_{m,n}$ of $\Omega_{expressive}$ and $\Omega_{receptive}$ that the dilated CNN + RN learned as the most predictive of expressive and receptive scores. Each activation map was averaged over the entire data set (n = 31 patients). Each 3D visualization showed corresponding edges (i.e., line segments) connecting AAL nodes listed in the 2D $M_{GRAD\text{-}CAM}$ maps of $\Omega_{expressive}$ and $\Omega_{receptive}$. Here, the greater weight indicates the presence of thicker edges contributing to a better score. For the prediction of the expressive language score,

TABLE 1 Performance comparison of the dilated CNN + RN with other methods for prediction of CELF score

	$\Omega_{expressive}$			$\Omega_{receptive}$		
Method	MAE	P(AE≤10)	R	MAE	P(AE≤10)	R
SVR	3.79 ± 0.71	0.97 ± 0.05	0.86 ± 0.07	3.31 ± 0.66	0.98 ± 0.04	0.73 ± 0.16
MLR	1.14 ± 0.81	0.97 ± 0.04	0.95 ± 0.04	0.64 ± 0.35	0.99 ± 0.02	0.96 ± 0.05
CNN + MLR	2.53 ± 0.88	0.97 ± 0.05	0.95 ± 0.06	1.63 ± 0.51	1.00 ± 0.00	0.98 ± 0.00
Dilated CNN + MLR	2.99 ± 0.81	0.95 ± 0.05	0.94 ± 0.06	1.54 ± 0.16	1.00 ± 0.00	0.98 ± 0.01
CNN + RN	1.15 ± 1.11	0.97 ± 0.05	0.95 ± 0.06	0.55 ± 0.36	1.00 ± 0.00	0.98 ± 0.01
Dilated CNN + RN	0.94 ± 0.93	0.97 ± 0.04	0.96 ± 0.06	0.46 ± 0.36	1.00 ± 0.00	0.99 ± 0.01

Note: For each method, the mean ± *SD* of MAE, P(AE≤10), and R were evaluated over threefold cross validation using 9,641 sample instances of three independent test cohorts: $\{A_{m,n}^{k}\}$ of $\Omega_{expressive}$ and $\Omega_{receptive}$.

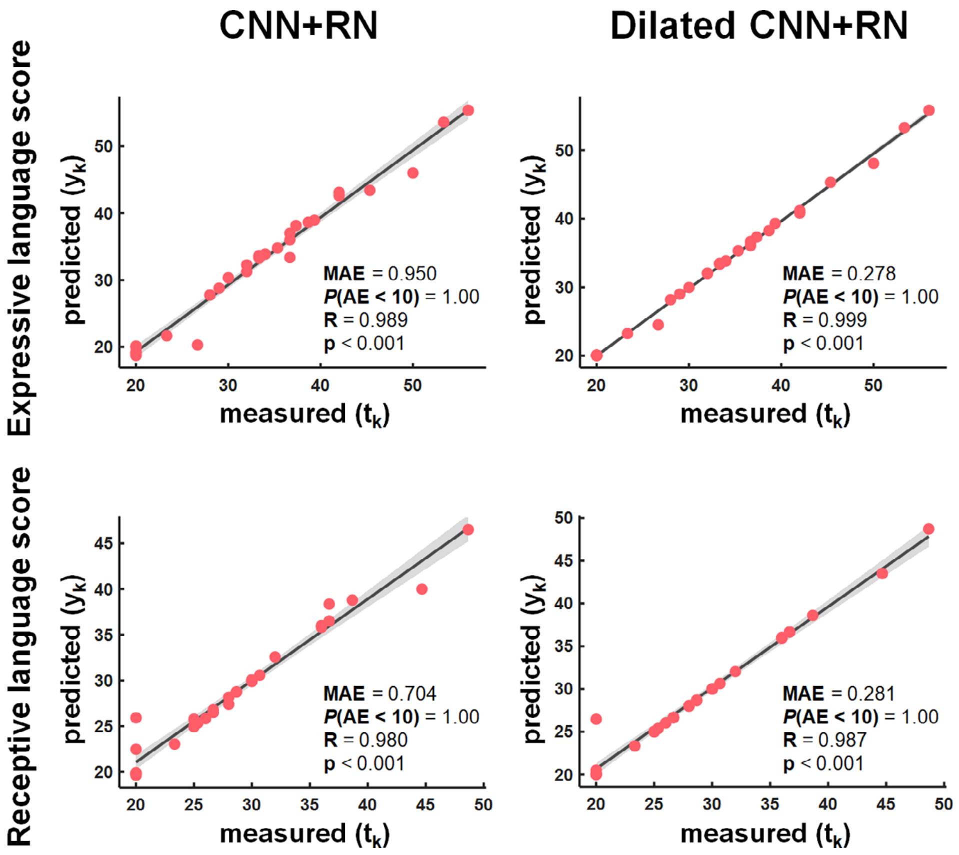

FIGURE 4 Significant linear correlations between the measured and predicted CELF language scores obtained from the three independent test folds (*n* = 31 patients), (left) CNN + RN and (right) dilated CNN + RN. The trained "CNN + RN" and "dilated CNN + RN" were applied to predict the measured scores of the corresponding test fold subject, t_k as plotted on the x-axis. Each y_k on the y-axis is the prediction of each t_k on the x-axis

TABLE 2 Mean absolute error (MAE), *SD* of absolute error (SDAE), prediction probability of the dilated CNN + RN driven language scores to achieve AE less than 10 (P(AE≤10)), and unpaired *t* test *p*-value evaluated from five different data sets of three independent test cohorts: whole data set (*n* = 31 patients), the time interval between MRI and neuropsychological assessment (Δt) ≤ 3 years (*n* = 19 patients), Δt > 3 years (*n* = 12 patients), age at MRI ≤3 years (*n* = 10 patients), and age at MRI > 3 year (*n* = 21 patients)

	Expressive language score			Receptive language score		
Group (*n*)	MAE ± SDAE	P(AE≤10)	*p*-value	MAE ± SDAE	P(AE≤10)	*p*-value
Whole (31)	0.28 ± 0.54	1.00		0.28 ± 1.16	1.00	
Δt≤3 year (19)	0.19 ± 0.46	1.00	.268	0.39 ± 1.48	1.00	.503
Δt > 3 year (12)	0.42 ± 0.65	1.00		0.10 ± 0.24	1.00	
Age of MRI ≤3 years old (10)	0.34 ± 0.67	1.00	.695	0.02 ± 0.02	1.00	.402
Age of MRI > 3 years old (21)	0.25 ± 0.49	1.00		0.40 ± 1.41	1.00	

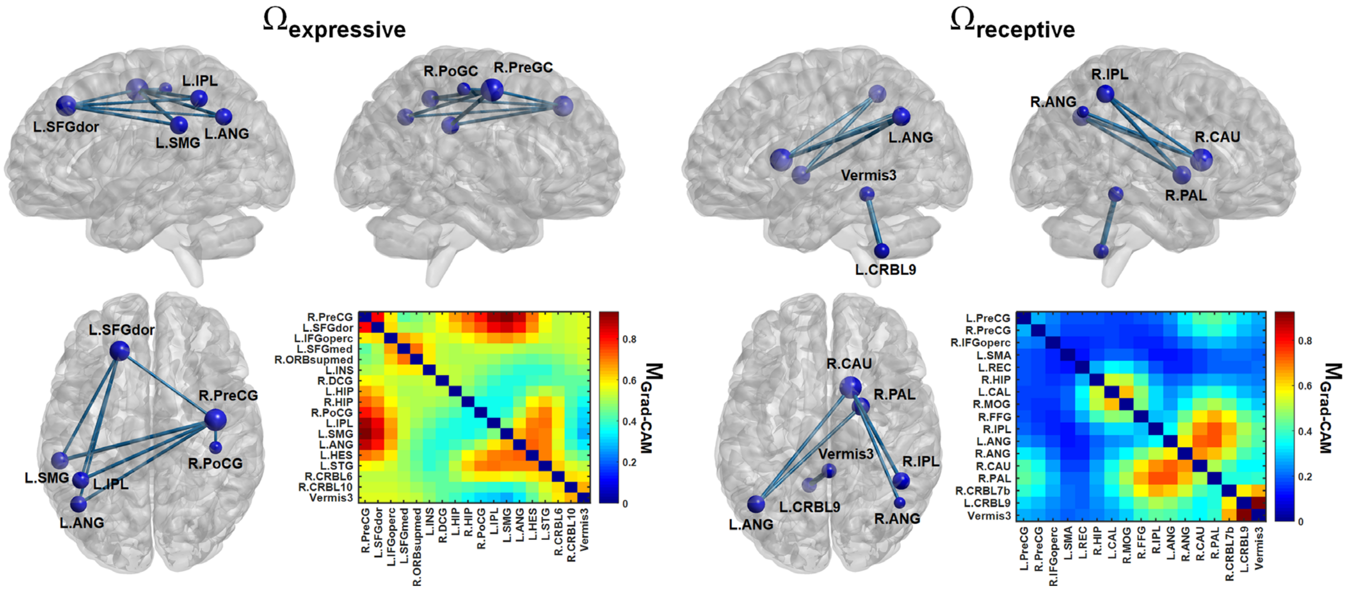

FIGURE 5 Activation maps ($M_{GRAD\text{-}CAM}$) showing AAL brain nodes, R_m of $A_{m,n}$ that the dilated CNN + RN learned as the most predictive of CELF scores, (left) expressive score using $\Omega_{expressive}$ and (right) receptive score using $\Omega_{receptive}$. Each 2D $M_{GRAD\text{-}CAM}$ presents an activation map showing pair-wise connection features, $A_{m,n}$, and their contributions (i.e., weights) to increase prediction accuracy. Higher weight (red-colored) indicates that a pair-wise connection is more predictive of the measured CELF scores. In 3D brain visualization, relatively small z-scores of activation weights (i.e., z-scores less than two times *SD* of each map) were omitted for clarity. The radius of each sphere indicates the sum of z-score at each node, while the thickness of the pair-wise edge represents z-score. A complete list of regional names corresponding to node labels is available in Table S1

the right precentral gyrus (R.PreCG), left supramarginal gyrus (L.SMG), angular gyrus (L.ANG), inferior parietal lobule (L.IPL), and superior frontal dorsal region (L.SFGdor) appeared to be prominent hubs of essential connections. That is, connection edges across the right PreCG, left SMG, ANG, and IPL were found to be the most predictive of positive expressive language outcomes. In contrast, the right caudate (R.CAU), pallidum (PAL), and left ANG regions were found to be prominent hubs of essential connections for the receptive language score. Connection edges connecting the right basal ganglia as well as left ANG were the most predictive of receptive language outcomes. In other words, weaker edges across these regions were predictive of worse language scores, suggesting the presence of receptive language impairment. There was an overlap in left ANG that is most predictive of both expressive and receptive language scores.

4 | DISCUSSION

The present study investigated the clinical feasibility of deep learning technology to predict the persistence of expressive and receptive language impairments in small children with late language acquisition (e.g., LLE). The dilated CNN + RN analysis outperformed other state-of-the-art methods to accurately and directly predict expressive and receptive language scores obtained at an older age. This finding supported our hypothesis that the comprehensive evaluation combining psychometry-driven DWIC language networks and deep reasoning of dilated convolutional features can be a breakthrough technology for the common MRI-negative LLE cases by providing neurological substrates and dissecting out this broad label into more distinct scientifically-based entities. Specifically, our dilated CNN + RN outperformed the CNN + RN to improve prediction errors up to 12 and 36% for expressive and receptive language scores, suggesting that the dilated convolutional features provide more efficiency for artificial reasoning at the same computational cost, especially in the case that the input matrix has the sparsely distributed clusters of edges in the nonlocal pattern.

Language functions are supported by widespread areas of the cortex and have multiple spatiotemporally distributed modules that serve various linguistic functions. For the differences in edges that are predictive of expressive vs. receptive scores, our Grad-CAM analysis found that there were commonalities between edges associated with expressive and receptive scores: they both involved bilateral structures and had one common structure (left angular gyrus). Significant information sharing was also identified by language-associated electrocorticography high-gamma modulation "within" frontal and temporoparietal language cortices, and "between" classical language areas in the dominant hemisphere (i.e., Broca's and Wernicke's areas) (Arya et al., 2019), suggesting that expressive and receptive language functions may communicate and share neural activity across multiple language regions. Interestingly, our result provided neurobiological evidence on a spatial redundancy (Schomers, Garagnani, & Pulvermüller, 2017) determined by our dilated CNN + RN to be most predictive of expressive and receptive scores (e.g., left ANG). However, information sharing between dilated CNN + RN-defined expressive and receptive language regions was not yet confirmed in the present study. Thus, further studies combining electrocorticography are warranted to better understand both specializations of the association cortices (Catani & ffytche, 2005) and axonal connections between dilated CNN + RN-defined hubs. Two notable differences are present in the structures involved in expressive and receptive language: the right hemisphere is more involved for receptive, and it also involves two subcortical structures (caudate and pallidum), while the expressive language network involved only cortical regions. Involvement of basal ganglia (usually caudate and putamen are mentioned) has been long known in language function, and reductions in the caudate volume are associated with language impairment (Tallal, Sainburg, & Jernigan, 1991). Badcock et al. (2012) also discussed specifically decreased right caudate volume in LI. However, it would be difficult to link the right caudate to receptive language exclusively, considering the role of the caudate nucleus in motor preparation and response planning. In addition, functional imaging of language functions has supported some involvement of the right-hemisphere in language development in typically developing persons (Müller et al., 1997). Empirical work has also indicated increased right (and decreased left) hemispheric involvement in children and adults (greater in children) with neural insult in cortical language regions, which have been suspected to reflect the neural reorganization of function (Müller et al., 1998). A recent review of fMRI studies (Deldar, Gevers-Montoro, Khatibi, & Ghazi-Saidi, 2020) also suggested a dynamic interaction between language and working memory, reflected in the involvement of subcortical structures, particularly the basal ganglia (caudate), and of widespread right hemispheric regions. Thus, our finding of the subcortical structures and right hemisphere involved in the receptive language network might reflect the language interaction in response to working memory function.

The advantage of using Grad-CAM over the conventional group analysis method is that Grad-CAM could be applied to the pre-trained dilated CNN + RN model to visualize the areas of the input connectome matrix that leads to accurate predictions. Conventional group analysis using Pearson's correlation with Bonferroni correction was first used to construct the input matrix with the most important edges for the prediction of specific language scores. To identify the most important edges in the input matrix, the Grad-CAM was then applied to the pre-trained deep learning model. That is, since a trained deep learning model is necessary for the prediction of CELF scores from unseen edge matrices, we used the Grad-CAM to identify the important connections in the input matrix. Thus, the connections which were common to both the Grad-CAM and the group analysis were the most important connections leading to an accurate CELF score. The findings of the present study support that Grad-CAM could be used as a reliable method to visualize important connections for future unseen test cases.

Moreover, Grad-CAM also has advantages over other deep learning-based visualization algorithms (Simonyan et al., 2013; Zhou, Wu, et al., 2016). For example, the saliency maps (Simonyan et al., 2013) calculate the partial derivative of the output class score w.r.t the input image pixels. Hence, it does not make use of the feature maps which describe the image at a higher abstract level. The issue of class-specific activation maps (CAM) (Zhou, Khosla, et al., 2016) is that it could only be applied to a specific set of CNN based architectures where global average pooling (GAP) operation is done before the final softmax layer for class score prediction (He, Zhang, Ren, & Sun, 2016). Compared with these methods, Grad-CAM (Selvaraju et al., 2017) is a more generalized activation map visualization algorithm, which could be applied to any CNN-based architecture. Thus, we used Grad-CAM to understand which connections in the input matrix are the most important in the prediction of the final regression or classification score.

Brain development is most dynamic in the first few years of life with cortical expansion and myelination of white matter tracts (Gogtay et al., 2004) - this includes the development of both cortex and white matter tracts/networks associated with language expression and reception. Many quantitative MRI studies (Deoni, Dean, Joelson, O'Regan, & Schneider, 2018; Lebel & Deoni, 2018; Zatorre, Fields, & Johansen-Berg, 2012) have demonstrated consistent, rapid microstructural white matter development over the first 3 years of life, including increased myelination and axonal packing. Diffusion MRI studies (Dubois et al., 2014; Fields, 2015; Qiu, Mori, & Miller, 2015) also reported changes attributed to continued white matter maturation, synaptogenesis, synaptic pruning, and remodeling during later childhood. As detailed in the emerging literature (Deoni et al., 2018), subtle, but differential, microstructural changes exist in most MRI signatures that may be associated with biological factors and environmental risks. We presume that the dilated CNN + RN, which deeply learns and intelligently reasons the hidden relationships between pair-wise edges in the pDWIC-based language networks might be powerful enough to pick up such subtle changes for prediction of the imaging-based language scores, although the nature of the reasoning remains unclear. More importantly, the present study provides detailed, evidence-based insight into the early white matter impairment that may underlie expressive and receptive language networks before acquiring language, possibly indicating the presence or extent of disorganized pruning and(or) poor myelination process in the early developmental period. For instance, a positive correlation between language scores and edge strengths (e.g., "right precentral gyrus," "right caudate") can indicate that these regions may function as new or additional hubs to compensate for altered language function. A recent neuroimaging study (Mitsuhashi et al., 2020) applying fMRI and DWI tractography to clarify the mechanisms of neural plasticity involved in language found increased fiber counts in right precentral gyrus in early bilingual acquisition, suggesting the increased neural connections between right precentral gyrus to right basal ganglia may be a key pattern of neural plasticity in language skills. Volumetric analyses (Burgaleta, Sanjuán, Ventura-Campos, Sebastian-Galles, & Ávila, 2016; Hervais-Adelman, Egorova, & Golestani, 2018) also revealed a significant relationship between multilingual experience and right caudate volume, as well as a marginally significant relationship with left caudate volume, suggesting right caudate as a key hub for language acquisition. Similarly, the left angular gyrus was suggested to be a key part of the perisylvian language network

(Catani, Jones, & ffytche, 2005) in multilingual children. Although Wernicke's original description was of a temporal lobe language area, the term Wernicke's area subsequently has been used to include inferior parietal areas, as well as posterior temporal areas encompassing angular gyrus, supramarginal gyrus, and inferior parietal lobule (BA 39 and 40) (Aboitiz & García, 1997). Thus, the increased axonal connectivity in these regions may reflect a substrate for language impairment or perhaps compensatory mechanism in an attempt to improve expressive language in late language acquisition. This, however, could only be proven by longitudinal studies.

In the proposed dilated CNN + RN, the predicted expressive language scores (32.8 ± 9.9, 20.0–55.8) had a higher *SD* than the predicted receptive language scores (29.2 ± 7.3, 20.0–48.7). It is likely because the measured expressive language scores (33.0 ± 10.1, 20–56) also had higher *SD* than the measured receptive language scores (29.2 ± 7.3, 20–48.7), evidencing that the dilated CNN + RN could have successfully learned the nonlinear relationship of each language function between the given edge matrix and the measured language score. That is, the nature of our sample cohort can affect the difference in *SD*. A recent study (Ryan, Gibbon, & O'shea, 2016) also found that the profile of superior expressive compared with receptive language is apparent in children with language delay and socioeconomic status (SES) factors such as parent–child interaction styles, level of maternal education, maternal sensitivity, income-to-needs ratio, and environmental variables may have a greater influence on receptive rather than expressive language. In addition, some data suggest a tendency for speech and language interventions to be more effective to improve expressive than receptive skills (Law, Garrett, & Nye, 2004)—this may be relevant because all of the children in the present study were receiving speech therapy services. Thus, our possible explanation for the lower *SD* of receptive language scores may be due to its higher sensitivity to the SES factors of our study cohort or the effect of speech-language therapy.

Despite our efforts to ensure methodological rigor, the present study has some unavoidable limitations. First, this retrospective work includes a small sample size (*n* = 31), which can be problematic for the individual machine learning model to learn heterogeneous natures of high dimensional features via a deep learning process. Although data augmentation was applied to alleviate this limitation (i.e., 6,531 sample instances of the training cohort used for training our dilated CNN + RN could prevent overfitting and resulted in a generalized model that could accurately predict the language scores in 3,110 sample instances of the independent test cohort), it was based on a randomly interpolated resampling procedure across the nearest neighbors, which may limit deep learning of the connectome features specific to a single institutional MRI cohort that is not yet generalized for multisite MRI cohorts. Thus, the reproducibility of each model should be evaluated using a new independent data set such as other institutional data. Also, the specific pattern of interhemispheric connectivity at the dilated CNN + RN-defined language hubs, which may reflect the effect of neural plasticity on axonal reorganization, has not been addressed by the present study. Our future study will focus on this reproducibility test to determine if the dilated CNN + RN can be translated to clinical cases in which current neuropschycometric language assessment fails to evaluate language abilities. While neither nonverbal intellect nor behavioral problems were correlated with language scores (and due to the sample size, which limited the number of variables to be included in data analyses), we did not statistically control for intellect/behavioral scores in the imaging-language data analyses. Thus, some of the findings could reflect an intellect/behavioral influence on the outcomes; further study, with larger samples and which controls for such extraneous concerns, will be needed to confirm the results of the present study. Due to retrospective nature of the study, we had no control over the interval between assessment and imaging. Finally, the present study did not consider combining clinical variables with DWIC features and the effect of MRI-assessment time interval to improve the prediction analyses. It is anticipated that features from both sources could be combined to improve the prediction accuracies further. As noted above, the sample only included children with LLE, and which limits our understanding of the meaning of the findings—future study including children with typically developing language skills is necessary to fully understand the fuller meaning of the findings in the present study.

In summary, this study provides additional data that may further elucidate the anatomical blueprint of the atypically developed language networks in children with persistent language concerns. Our approach may lead to the refinement of imaging and psychometric language phenotype relationships in a detailed manner, and ultimately may allow for earlier and more reliable identification of persisting language impairment in small children, allowing for earlier onset of intervention programs, and further may allow for the identification of the type of language impairment, which could lead to more targeted intervention in such children.

ACKNOWLEDGMENT

This work was supported by the US National Institutes of Health (NIH) grants R01NS089659 (to J. J.) and US National Science Foundation (NSF) grant CNS-1637312 (to M. D.).

CONFLICT OF INTERESTS

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Jeong-Won Jeong https://orcid.org/0000-0003-4498-0939

Csaba Juhász https://orcid.org/0000-0002-5067-5554

REFERENCES

Aboitiz, F., & García, V. R. (1997). The evolutionary origin of the language areas in the human brain. A neuroanatomical perspective. *Brain Research Reviews*, *25*, 381–396.

Andersson, J. L., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage*, *20*, 870–888.

Arya, R., Ervin, B., Wilson, J. A., Byars, A. W., Rozhkov, L., Buroker, J., & Holland, K. D. (2019). Development of information sharing in language neocortex in childhood-onset drug-resistant epilepsy. *Epilepsia*, *60*, 393–405.

Badcock, N. A., Bishop, D. V. M., Hardiman, M. J., Barry, J. G., & Watkins, K. E. (2012). Co-localisation of abnormal brain structure and function in specific language impairment. *Brain and Language*, *120*(3), 310–320.

Banerjee, S, Dong, M, Lee, M., O'Hara, N. B., Asano, E., & Jeong, J. W. (2020). Prediction of language impairment using deep relational reasoning. *In IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1680–1684.

Banerjee, S., Dong, M., Lee, M. H., O'Hara, N. B., Juhász, C., Asano, E., & Jeong, J. W. (2021). Deep relational reasoning for the prediction of language impairment and postoperative seizure outcome using preoperative DWI connectome data of children with focal epilepsy. *IEEE Transactions on Medical Imaging*, *40*(3), 792–804.

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & the CATALISE-2 consortium. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *The Journal of Child Psychology and Psychiatry*, *58*, 1068–1080.

Brown, C. J., Miller, S. P., Booth, B. G., Poskitt, K. J., Chau, V., Synnes, A. R., Hamarneh, G. (2015). Prediction of motor function in very preterm infants using connectome features and local synthetic instances. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 69–76.

Burgaleta, M., Sanjuán, A., Ventura-Campos, N., Sebastian-Galles, N., & Ávila, C. (2016). Bilingualism at the core of the brain. Structural differences between bilinguals and monolinguals revealed by subcortical shape analysis. *NeuroImage*, *125*, 437–445.

Catani, M., & ffytche, D. H. (2005). The rises and falls of disconnection syndromes. *Brain*, *128*, 2224–2239.

Catani, M., Jones, D. K., & ffytche, D. H. (2005). Perisylvian language networks of the human brain. *Annals of Neurology*, *57*, 8–16.

Chai, L. R., Mattar, M. G., Blank, I. A., Fedorenko, E., & Bassett, D. S. (2016). Functional network dynamics of the language system. *Cerebral Cortex*, *26*, 4148–4159.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Cheng, H., Wang, Y., Sheng, J., Kronenberger, W. G., Mathews, V. P., Hummer, T. A., & Saykin, A. J. (2012). Characteristics and variability of structural networks derived from diffusion tensor imaging. *NeuroImage*, *61*, 1153–1164.

Dale, P. S., Price, T. S., Bishop, D. V., & Plomin, R. (2003). Outcomes of early language delay: I. predicting persistent and transient language difficulties at 3 and 4 years. *Journal of Speech, Language, and Hearing Research*, *46*, 544–560.

Deldar, Z., Gevers-Montoro, C., Khatibi, A., & Ghazi-Saidi, L. (2020). The interaction between language and working memory: A systematic review of fMRI studies in the past two decades. *AIMS Neuroscience*, *8* (1), 1–32.

Deoni, S., Dean, D. r., Joelson, S., O'Regan, J., & Schneider, N. (2018). Early nutrition influences developmental myelination and cognition in infants and young children. *NeuroImage*, *178*, 649–659.

Downing, J. E., & Perino, D. M. (1992). Functional versus standardized assessment procedures: Implications for educational programming. *Mental Retardation*, *30*, 289–295.

Dubois, J., Dehaene-Lambertz, G., Kulikova, S., Poupon, C., Hüppi, P. S., & Hertz-Pannier, L. (2014). The early development of brain white matter: A review of imaging studies in fetuses, newborns and infants. *Neuroscience*, *276*, 48–71.

Ellis Weismer, S., Murray-Branch, J., & Miller, J. F. (1994). A prospective longitudinal study of language development in late talkers. *Journal of Speech, Language, and Hearing Research*, *37*, 852–867.

Fields, R. D. (2015). A new mechanism of nervous system plasticity: Activity-dependent myelination. *Nature Reviews Neuroscience*, *16*, 756–767.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., & Thompson, P. M. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 8174–8179.

Hagler, D. J., Saygin, A. P., & Sereno, M. I. (2006). Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *NeuroImage*, *33*(4), 1093–1103.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778.

Hervais-Adelman, A., Egorova, N., & Golestani, N. (2018). Beyond bilingualism: Multilingual experience correlates with caudate volume. *Brain Structure and Function*, *223*, 3495–3502.

Hussain, Z., Gimenerz, F., Yi, D., & Rubin, D. (2017). Differential data augmentation techniques for medical imaging classification tasks. *AMIA Annual Symposium Proceedings Archive*, *2017*, 979–984.

Jeong, J. W., Asano, E., Juhasz, C., Behen, M. E., & Chugani, H. T. (2016). Postoperative axonal changes in the contralateral hemisphere in children with medically refractory epilepsy: A longitudinal diffusion tensor imaging connectome analysis. *Human Brain Mapping*, *37*, 3946–3956.

Jeong, J. W., Sundaram, S., Behen, M. E., & Chugani, H. T. (2016a). Differentiation of speech delay and global developmental delay in children using DTI tractography-based connectome. *American Journal of Neuroradiology*, *37*, 1170–1177.

Jeong, J. W., Sundaram, S., Behen, M. E., & Chugani, H. T. (2016b). Relationship between genotype and arcuate fasciculus morphology in six young children with global developmental delay: Preliminary DTI stuy. *Journal of Magnetic Resonance Imaging*, *44*, 1504–1512.

Kawahara, J., Brown, C. J., Miller, S. P., Booth, B. G., Chau, V., Grunau, R. E., & Hamarneh, G. (2017). BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, *146*, 1038–1049.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*:1412.6980. https://arxiv.org/abs/1412.6980

Law, J., Garrett, Z., & Nye, C. (2004). The efficacy of treatment for children with developmental speech and language delay/disorder: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *47*(4), 924–943.

Lebel, C., & Deoni, S. (2018). The development of brain white matter microstructure. *NeuroImage*, *182*, 207–218.

Lee, M. H., O'Hara, N. B., Behen, M. E., & Jeong, J. W. (2020). Altered efficiency of white matter connections for language function in children with language disorder. *Brain and Language*, *203*, 104743.

Mitsuhashi, T., Sugano, H., Asano, K., Nakajima, T., Nakajima, M., Okura, H., & Arai, H. (2020). Functional MRI and structural connectome analysis of language networks in Japanese-English bilinguals. *Neuroscience*, *431*, 17–24.

Müller, R. A., Rothermel, R. D., Behen, M. E., Muzik, O., Mangner, T. J., Chakraborty, P. K., & Chugani, H. T. (1998). Brain organization of language after early unilateral lesion: A PET study. *Brain and Language*, *62* (3), 422–451.

Müller, R. A., Rothermel, R. D., Behen, M. E., Muzik, O., Mangner, T. J., & Chugani, H. T. (1997). Receptive and expressive language activations for sentences: A PET study. *Neuroreport*, *8*(17), 3767–3770.

Nelson, H. D., Nygren, P., Walker, M., & Panoscha, R. (2006). Screening for speech and language delay in preschool children: Systematic evidence review for the US preventive services task force. *Pediatrics*, *117*, e298–e319.

Paul, R., Hernandez, R., Taylor, L., & Johnson, K. (1996). Narrative development in late talkers: Early school age. *Journal of Speech and Hearing Research*, *39*, 1295–1303.

Qiu, A., Mori, S., & Miller, M. I. (2015). Diffusion tensor imaging for understanding brain development in early life. *Annual Review of Psychology*, *66*, 853–876.

Rescorla, L. A. (1989). The language development survey: A screening tool for delayed language in toddlers. *Journal of Speech and Hearing Disorders*, *54*, 587–599.

Rescorla, L. A. (2000). Do late-talking toddlers turn out to have reading difficulties a decade later? *Annals of Dyslexia*, *50*, 87–102.

Rescorla, L. A., & Alley, A. (2001). Validation of the language development survey (LDS): A parent report tool for identifying language delay in toddlers. *Journal of Speech, Language, and Hearing Research*, *44*, 434–445.

Rice, M. L., Taylor, C. L., & Zubrick, S. R. (2008). Language outcomes of 7-year-old children with or without a history of late language emergence at 24 months. *Journal of Speech, Language, and Hearing Research*, *51*, 394–407.

Roulstone, S., Loader, S., Northstone, K., & Beveridge, M. (2002). The speech and language of children aged 25 months: Descriptive data from the Avon longitudinal study of parents and children. *Early Child Development and Care*, *172*, 259–268.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, *52*, 1059–1069.

Ryan, A., Gibbon, F. E., & O'shea, A. (2016). Expressive and receptive language skills in preschool children from a socially disadvantaged area. *International Journal of Speech-Language Pathology*, *18*(1), 41–52.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems* 4967–4976.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications*. CA. Jerome M. Sattler: San Diego.

Schomers, M. R., Garagnani, M., & Pulvermüller, F. (2017). Neurocomputational consequences of evolutionary connectivity changes in perisylvian language cortex. *Journal of Neuroscience*, *37*, 3045–3055.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Semel, E., Wiig, E. H., & Secord, W. A. (2006). *Clinical evaluation of language fundamentals-preschool-second edition (CELF-P2): Australian and New Zealand* (Standardised ed.). NSW: Australia. Harcourt Assessment Inc.

Shevell, M., Ashwal, S., Donley, D., Flint, J., Gingold, M., Hirtz, D., & Sheth, R. D. (2003). Practice parameter: Evaluation of the child with global developmental delay: Report of the quality standards Subcommittee of the American Academy of neurology and the practice Committee of the Child Neurology Society. *Neurology*, *60*, 367–380.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*:1312.6034. https://arxiv.org/abs/1312.6034

Sundaram, S. K., Sivaswamy, L., Makki, M. I., Behen, M. E., & Chugani, H. T. (2008). Absence of arcuate fasciculus in children with global developmental delay of unknown etiology: A diffusion tensor imaging study. *The Journal of Pediatrics*, *152*, 250–255.

Tallal, P., Sainburg, R. L., & Jernigan, T. (1991). The neuropathology of developmental dysphasia: Behavioral, morphological, and physiological evidence for a pervasive temporal processing disorder. *Reading and Writing*, *3*(3–4), 363–377.

Tounier, J. D., Calamante, F., & Connelly, A. (2010). Improved probabilistic streamlines tractography by 2nd order integration over fiber orientation distributions. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, pp. 1670.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*, 273–289.

Verly, M., Gerrits, R., Sleurs, C., Lagae, L., Sunaert, S., Zink, I., & Rommel, N. (2019). The mis-wired language network in children with developmental language disorder: Insights from DTI tractography. *Brain Imaging and Behavior*, *13*, 973–984.

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv*:1511.07122.

Zatorre, R. J., Fields, R. D., & Johansen-Berg, H. (2012). Plasticity in gray and white: Neuroimaging changes in brain structure during learning. *Nature Neuroscience*, *15*, 528–536.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Zhou, S., Wu, J. N., Wu, Y., & Zhou, X. (2016). Exploiting local structures with the kronecker layer in convolutional networks. *arXiv*:1512.09194. https://arxiv.org/abs/1512.09194

Zubrick, S. R., Taylor, C. L., Rice, M. L., & Slegers, D. W. (2007). Late language emergence at 24 months: An epidemiological study of prevalence, predictors, and covariates. *Journal of Speech, Language, and Hearing Research*, *50*, 1562–1592.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Jeong J-W, Banerjee S, Lee M-H, et al. Deep reasoning neural network analysis to predict language deficits from psychometry-driven DWI connectome of young children with persistent language concerns. *Hum Brain Mapp*. 2021;42:3326–3338. https://doi.org/10.1002/hbm.25437