Received: 30 August 2021 | Revised: 11 March 2022 | Accepted: 6 May 2022

DOI: 10.1002/hbm.25953

RESEARCH ARTICLE

WILEY

Generalizable predictive modeling of semantic processing ability from functional brain connectivity

Danting Meng[1,2] | Suiping Wang[1] | Patrick C. M. Wong[3,4] | Gangyi Feng[3,4]

[1]Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents (South China Normal University), Ministry of Education, Guangzhou, China

[2]Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou, China

[3]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong SAR, China

[4]Brain and Mind Institute, The Chinese University of Hong Kong, Hong Kong SAR, China

Correspondence

Gangyi Feng, Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong SAR, China.
Email: g.feng@cuhk.edu.hk

Suiping Wang, Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents (South China Normal University), Ministry of Education, Guangzhou, 510631, China.
Email: wangsuiping@m.scnu.edu.cn

Funding information

Research Grants Council, University Grants Committee, Grant/Award Numbers: 14614221, 14619518; Direct Grant for Research, The Chinese University of Hong Kong, Grant/Award Number: 4051137; National Natural Science Foundation of China, Grant/Award Number: 32171051

Abstract

Semantic processing (SP) is one of the critical abilities of humans for representing and manipulating conceptual and meaningful information. Neuroimaging studies of SP typically collapse data from many subjects, but its neural organization and behavioral performance vary between individuals. It is not yet understood whether and how the individual variabilities in neural network organizations contribute to the individual differences in SP behaviors. We aim to identify the neural signatures underlying SP variabilities by analyzing functional connectivity (FC) patterns based on a large-sample Human Connectome Project (HCP) dataset and rigorous predictive modeling. We used a two-stage predictive modeling approach to build an internally cross-validated model and to test the model's generalizability with unseen data from different HCP samples and other out-of-sample datasets. FC patterns within a putative semantic brain network were significantly predictive of individual SP scores summarized from five SP-related behavioral tests. This cross-validated model can be used to predict unseen HCP data. The model generalizability was enhanced in the language task compared with other tasks used during scanning and was better for females than males. The model constructed from the HCP dataset can be partially generalized to two independent cohorts that participated in different semantic tasks. FCs connecting to the Perisylvian language network show the most reliable contributions to predictive modeling and the out-of-sample generalization. These findings contribute to our understanding of the neural sources of individual differences in SP, which potentially lay the foundation for personalized education for healthy individuals and intervention for SP and language deficits patients.

KEYWORDS

functional connectivity, individual differences, model generalization, predictive modeling, semantic processing

1 | INTRODUCTION

Making sense of the outside world is critical for human survival and development. The ability to store, retrieve, and manipulate meaningful information (e.g., concepts) is central to many cognitive functions and is also a defining characteristic of human brains (Berwick et al., 2013; Nation & Snowling, 1998; Ratcliff et al., 2010). This so-called semantic processing (SP) ability is a gift for humans in general. However, there

are extensive interindividual differences in SP evidenced by various behavioral tasks, ranging from object recognition (Lewellen et al., 1993), categorization decision (Plaut & Booth, 2000; Schilling et al., 1998; Yap et al., 2012) to language comprehension, and production (Just & Carpenter, 1987; Walker et al., 1994). These behavioral SP measures have also been demonstrated to be associated with various cognitive components, such as executive control (Allen et al., 2012), attention (McGlinchey-Berroth et al., 1993), selection (Nation & Snowling, 1998), and inhibition abilities (Cain, 2006). These findings suggest that SP may comprise multiple cognitive components where interindividual variabilities in these components may contribute to the variability in SP. Although these behavioral findings provide insights into our understanding of the composition of the SP variability, it is not yet clear to which extent the individual differences in neural organizations contribute to the individual differences in SP behaviors.

One candidate neural source underlying variability in SP is interregional functional connectivity (FC) patterns within a putative semantic network (Binder et al., 2009; Binder & Fernandino, 2015). The multifaceted essence of SP in cognitive composition suggests that the neural implementation of SP requires joint efforts of multiple brain regions, not only reflecting in regional activities but also more in interregional connectivity. Consistent with this hypothesis, findings derived from a large number of group-level neuroimaging studies have shown that SP is robustly related to distributed brain regions, including the left inferior frontal gyrus, left superior temporal gyrus/sulcus, left middle temporal gyrus, left anterior temporal lobe, left angular gyrus, left inferior parietal lobule, medial prefrontal cortex, and posterior cingulate cortex (Binder et al., 2009; Binder & Desai, 2011; Vigneau et al., 2006). These SP-related regions can be divided into three subnetworks; classical Perisylvian language network (PSN), frontoparietal network (FPN), and default mode network (DMN) based on their resting-state FC profiles (Xu et al., 2016). The functional associations of these subnetworks are distinct and may relate to different aspects of SP. For example, PSN regions have been proposed to be related to high-level linguistic processes (Fedorenko, 2014; Fedorenko et al., 2011; Feng et al., 2016), language learning (Feng, Li, et al., 2021; Forkel et al., 2014; Xiang et al., 2012), and language recovery after stroke (Dehaene-Lambertz et al., 2006; Griffis et al., 2017; López-Barroso et al., 2013; Ojemann, 1991; Saur et al., 2006). FPN regions relate to semantic control processes (Feng et al., 2016; Geranmayeh et al., 2012, 2014, 2017; Wirth et al., 2011), and DMN regions may relate to the social concept representation and processing (Binder et al., 2009; Binder & Fernandino, 2015) as well as integrating or simulating multimodal experiences (see Xu et al., 2017 for a review).

While this large body of research paints a convincing picture of the relationship between the semantic network and SP behaviors at the group level, it does not adequately acknowledge the tremendous individual differences in SP. Few studies investigate whether participants' variabilities in neural patterns contribute to predicting individual differences in SP behaviors and, if so, how. With small sample sizes and correlational approaches, previous studies show that interindividual variability in FC between SP-related regions was associated with individual differences in SP behaviors. In particular, the strength of the connectivity between a network hub in the middle temporal gyrus and other areas explains interindividual differences in SP performance (Krieger-Redwood et al., 2016; Mollo et al., 2016; Vatansever et al., 2017; Wei et al., 2012). These studies provide initial evidence supporting neural connectivity as one primary source of individual differences in SP. At the same time, there are significant limitations in the findings due to the methodology constraint and solely focusing on individual connectivity strengths.

Small sample sizes and traditional correlational approaches could potentially inflate the neural–behavioral correlations, limiting the generalization of a finding from one population to other unseen samples. Questions were raised on the reliability and replicability of the correlational findings. Most of the previous neuroimaging studies in SP relied on datasets with a limited number of subjects, which have low statistical power in general, leading to inflated effect size estimates, and poor replicability (Dubois & Adolphs, 2016; Schönbrodt & Perugini, 2013). For example, for behavioral studies, statistical simulation with the Monte-Carlo procedure has demonstrated that stable correlation estimation should approach a minimal sample size of 250 for a significant degree of confidence (Schönbrodt & Perugini, 2013). For fMRI studies, limited available subjects would reduce the stability, test–retest reliability, and replicability of an activation effect estimate across experimental paradigms used (Bossier et al., 2020; Kühberger et al., 2014).

Moreover, traditional correlational methods often overestimate the neural–behavioral relationships and do not ensure the generalizability of the established relationship from one sample population to out-of-sample subjects (Dubois & Adolphs, 2016; Lo et al., 2015; Whelan & Garavan, 2014). This out-of-sample generalization ability is rarely demonstrated in behavioral and neuroimaging studies, partly due to the small sample sizes and lack of use of machine-learning approaches. For example, previous studies show that machine learning algorithms, cross-validation, and randomization procedure can prevent overfitting data and promote model generalization and replication (Shmueli, 2010; Yarkoni & Westfall, 2017).

To identify the FC patterns underlying individual differences in SP while overcoming the methodological limitations, here we used the predictive modeling approach (Finn et al., 2015; Rosenberg et al., 2013; Shen et al., 2017) with two-step cross-validation and generalization procedures to construct and validate SP prediction models based on a large number of subjects ($N = 868$) from the Human Connectome Project (HCP; Van Essen et al., 2013). First, to estimate the individual differences in SP while minimizing the bias in selecting a single SP task and reducing confoundment from other task-specific cognitive components, we extracted a latent core SP factor from five SP-related behavioral tests with confirmatory factor analysis. Two other cognitive components (i.e., cognitive and motor controls) were also estimated from other offline test tasks presumably unrelated to SP to examine the sensitivity and specificity of the models in predicting SP. Second, we constructed cross-validated semantic models to predict individual latent SP scores with half of the HCP

samples. We identified the most predictive functional connections to build a prediction model. The model was then generalized to another half of the unseen HCP samples and two independent datasets with different populations and SP tasks. These in-sample and out-of-sample model predictions were evaluated by bootstrapping and permutation procedures to assess the statistical significance of the model's predictive power, reliability, and generalizability. Moreover, previous studies have identified factors that modulate the model predictability of cognitive traits with FC patterns (e.g., task and gender) (Greene et al., 2018; Gao et al., 2019; R. Jiang et al., 2020). Here, we further explored the extent to which the populational factors (i.e., gender and age) and different fMRI task states modulate the SP model prediction and generalization.

2 | MATERIALS AND METHODS

2.1 | Dataset description and sample population

Three datasets were used in the current study to construct prediction models and estimate model generalization performances. The three datasets include the HCP 1200 Subjects Release (Van Essen et al., 2013), the semantic lexical decision (SLD) dataset (Feng et al., 2016), and the Alice story comprehension (ASC) dataset (Bhattasali et al., 2020). The HCP dataset was used to construct and cross-validate models and assess in-sample model generalization performance. The two independent datasets (i.e., SLD and ASC) were used to estimate the out-of-sample model generalization performance.

2.1.1 | HCP dataset

The HCP dataset includes behavioral and 3T MR imaging data from 1206 healthy young adult subjects. These subjects were asked to complete a resting-state session and a range of cognitive tasks during fMRI scanning (Barch et al., 2013). They also completed a battery of cognitive assessments outside the scanner. Detailed descriptions of these fMRI tasks and behavioral tests were listed in Methods in Data S1 and Table S1. We applied the following criteria to exclude subjects from the FC analysis and predictive modeling: (i) Subjects must have completed all resting-state and task-based fMRI scanning (including language, working memory, gambling, motor, social cognition, relation, and emotion processing tasks) as well as the cognitive assessments of interest (see Table S1 for the detailed test description). Subjects with any missing data in any scans and tests were discarded ($N = 253$); (ii) Imaging quality control. We excluded subjects whose data was collected during the period of known intermittent problems with head coil leading to temporal instability in acquisitions (i.e., QC_Issue = C; $N = 75$); (iii) We also removed subjects whose one or more functional scans contained any significant coil- or movement-related artifact that manifests prominently in the "minimally preprocessed" data (i.e., QC_Issue = D; $N = 10$). Finally, 868 subjects (407 males, age range 22–35 years old) were included in our analyses. The HCP scan protocol was approved by the local Institutional Review Board at Washington University in St. Louis.

2.1.2 | Independent datasets

The two independent datasets were included to evaluate the out-of-sample model generalization. These two datasets were chosen because they both have semantic tasks during fMRI scanning and represent two types of research protocols. The SLD uses a classical semantic priming paradigm (i.e., semantic-unrelated word pairs versus related pairs) with rigorous experimental controls. The ASC uses a more ecological setting probing semantic processes (e.g., semantic access and integration) during language comprehension. The SLD dataset has resting-state and task-based fMRI scans, consisting of 26 healthy Chinese participants (11 males, 18–28 years old; Feng et al., 2016). All participants were right-handed undergraduate or graduate Chinese students with normal or corrected-to-normal vision and no prior history of neuropsychiatric disorders. They were asked to fixate on a cross during the resting-state scans and lie still. For the task-based scanning, the participants were asked to complete a lexical decision task (i.e., judging whether the second word is real or not) for a list of Chinese word pairs. Word pairs include semantic-related (e.g., "Bread–Cake"), unrelated (e.g., "Driver–Cake"), and nonword pairs.

The ASC dataset only includes task-based fMRI scans (Bhattasali et al., 2020). This dataset contains 26 healthy native English speakers (11 males, 18–24 years old, right-handed). All participants were right-handed, with normal or corrected-to-normal vision, and no prior history of neuropsychiatric disorders. During the scanning, participants listened passively to an audio storybook of the first chapter of *Alice's Adventure in Wonderland* (duration = 12.4 min) read by Kristen McQuillan. Participants answered 12 multiple-choice questions related to the story after scanning.

2.2 | Imaging data acquisition

2.2.1 | HCP dataset

Structural T1-weighted images were acquired using a magnetization-prepared rapid acquisition gradient-echo (MPRAGE) sequence (TR = 2400 ms, TE = 2.14 ms, flip angle = 8°, FOV = 228 × 224 × 180 mm, 0.7-mm isotropic voxels). Whole-brain gradient echo-planar imaging (EPI) data were acquired with a 32 channel head coil on a modified 3T Siemens Skyra (TR = 720 ms, TE = 33.1 ms, flip angle = 52°, FOV = 208 × 180 × 144 mm^3, 2-mm isotropic voxels, multiband acceleration factor = 8) with 72 oblique axial slices that alternated between phase encoding in the right to left direction in one run and the left to right direction in the other run (Uğurbil et al., 2013; Van Essen et al., 2012). Imaging data with different phase-encoding directions were used for FC analysis. The

resting-state and task-state sessions were performed separately (see more details in Methods in Data S1).

2.2.2 | SLD dataset

This dataset was acquired using a Siemens Trio 3T MRI system with a 32-channel head coil. High-resolution T1-weighted structural images were acquired using an MPRAGE sequence (TR = 1900 ms, TE = 2.53 ms, flip angle = 9°, 176 slices, 1-mm isotropic voxels). The functional MRI data were recorded by a T2*-weighted EPI pulse sequence (TR = 2000 ms, TE = 20 ms, flip angle = 90°, field of view = 224 × 224 mm, in-plane resolution = 3.5 × 3.5 mm, 38 slices, slice thickness = 3.5 mm with 1.1 mm gap). Both resting- and task-state fMRI data were acquired.

2.2.3 | ASC dataset

This dataset was acquired using a GE Discovery MR750 3T MRI scanner with a 32-channel head coil. High-resolution T1-weighted structural images were collected with an MPRAGE sequence (1-mm isotropic voxels). For functional data acquisition, 10 participants were scanned with a T2*-weighted EPI sequence (TR = 2000 ms, TE = 27 ms, flip angle = 77°, field of view = 216 × 216 mm, 44 slices, 3-mm isotropic voxels, acceleration factor = 2). Sixteen subjects were scanned with a three-echo EPI sequence (TR = 2000 ms, TE = 27.5 ms, field of view = 240 × 240 mm, in-plane resolution = 3.75 × 3.75 mm, 33 slices, slice thickness = 3.8 mm with 0.05 mm gap). Only task-state fMRI data were acquired.

2.3 | Estimation of latent SP ability

We selected five offline behavioral tests in the HCP dataset to estimate a latent core SP factor. These tests were derived from the NIH Toolbox Cognition Battery (Gershon et al., 2013). We select tests that require participants to store, retrieve, or manipulate meaningful objects or conceptual information explicitly or implicitly to a certain extent. The five tasks include the Picture Vocabulary (PicVoc) (measures the object representation, semantic access, and vocabulary retrieval ability), the Oral Reading Recognition (ReadEng) (measures the semantic retrieval of words and reading decoding ability), List Sorting Working Memory (ListSort) (measures temporal storage of visually and orally presented objects), Picture Sequence Memory (PicSeq) (measures storage capacity of a series of visual objects and events), and Pattern Comparison Processing Speed (ProcSpeed) (measures the object representation and processing speed of object discernment) (see Table S1 for a detailed description of each test).

We selected these tests because they are related to SP to a certain extent. For example, PicVoc and ReadEng may be more related to the traditional-defined semantic processes due to the tasks requiring participants to access actively (e.g., in PicVoc) and manipulate (e.g., in ReadEng) semantic information explicitly. The other three tests (i.e., PicSeq, ListSort, and ProcSpeed) involve semantic processes implicitly, where the participants need to access the semantic information of the stimuli (e.g., concrete objects or concepts) to complete the tasks. For example, PicSeq requires participants to memorize the order of sequentially presented objects and events (Bauer et al., 2013). The objects and events are thematically related, and the participants need to understand the semantic relationships of the objects and events to better memorize and recall the test items. Like PicSeq, ListSort and ProcSpeed require information to be accessed at the semantic level during the tasks to a certain extent. At the same time, these tests may vary their focus in assessing different cognitive components (e.g., working memory and pattern extraction).

We also included two tests related to general cognitive control (CC) and four tests related to motor control (MC) to estimate the latent CC and MC factors, respectively. The two CC-related tests include dimensional change card sort (CardSort), which measures executive function and cognitive flexibility, and flanker task (Flanker), which measures selection and inhibition ability. The four MC-related tests include the 2-min walk endurance (Endurance) (measures endurance), 4-m walk (GaitSpeed) (measures locomotion), 9-hole pegboard (Dexterity) (measures dexterity), and grip strength dynamometry (Strength) (measures strength). These tests do not include meaningful object items or conceptual or language stimuli.

We adopted a confirmatory factor analysis (CFA) with those behavioral test variables to estimate the latent SP, CC, and MC factors. CFA is a multivariate statistical procedure often used to test how well the observed variables support the theoretical constructs of interest (Harrington, 2009). CFA attempts to reproduce the observed covariances between test items of interest with a more concise set of latent factors. CFA was used to confirm the putative latent SP factors using a set of test variables, which partition the test variances into two broad types: the variances due to a common latent SP factor and variances due to task-specific settings. We specified a hypothesized three-factor model. The five SP-related test variables contributed to a latent SP factor, two CC-related test variables contributed to a latent CC factor, and four MC-related test variables contributed to a latent MC factor.

Scores from the tests were normalized and age-adjusted using the age-appropriate band of Toolbox Norming Sample (bands of ages 18–29, or 30–35), where a score of 100 indicates performance that was at the national average and a score of 115 or 85 indicates performance 1 SD above or below the national average for participants' age band. We conducted the CFA model fitting using the R package Lavaan (Rosseel, 2012) with the maximum likelihood function to iteratively minimize the differences between the model-implied variance–covariance matrix and the sample variance–covariance matrix. The latent factor scores were generated for all participants after the model fitting.

2.4 | SP scores of the independent datasets

We used the semantic priming effect in lexical decision time as individuals' SP scores for the SLD dataset. The strength of semantic

priming was calculated by subtracting the lexical decision time of the semantic-related condition (i.e., a target word paired with a semantically related prime word) from the semantic-unrelated condition (i.e., a target word paired with an unrelated prime word) for each participant. Semantic priming was first reported by Meyer and Schvaneveldt (1971). Lexical decision time to a word is facilitated by prior presentation of its semantically related words. For example, prior exposure of "doctor" would facilitate subsequent recognition of "nurse" (faster in recognition time). We used this semantic priming in recognition time as an SP index to reflect individual participants' semantic access/activation processing ability. Generally, more semantic priming reflects more facilitation in semantic processes (e.g., semantic access) of the target words, thus reflecting tighter relationships between the related words in participants' mental semantic/conceptual representations. Trials with error responses and response times deviating from the mean of all trials by more than three standard deviations were removed. The semantic priming scores across subjects range from −66.7 to 87.0 ms (mean = 21.0 ms).

For the ASC dataset, the performance of a story comprehension test was used as individuals' SP scores. Participants were required to answer 12 questions after listening to an audio storybook. To answer these questions correctly, participants need to accurately understand the content of the storybook (i.e., speech signals) during scanning, where they were required to extract the meaning of individual words and cohesively combine them (Hagoort, 2005; Werning et al., 2012). SP involved in this task is defined as accessing the meaning of individual words and integrating the individual semantic units into larger representations. Relative to the semantic priming effect, this comprehension SP score is a more ecological measure of SP ability and requires less experimental control. Higher scores implicate the participants have better semantic and/or language processing ability. These comprehension scores range from 5 to 12 (mean = 9.81, SD = 1.67).

2.5 | Imaging data analysis

2.5.1 | Preprocessing

All three datasets were processed with the same analysis pipeline. The HCP data was downloaded in its minimally preprocessed form (i.e., after motion correction, B0 distortion correction, co-registration to T1-weighted images, and normalization to the MNI152 space; see Glasser et al. (2013) for detailed preprocessing parameters). The other two independent datasets were preprocessed with the same procedure as the HCP using Data Processing Assistant for Resting-State fMRI (DPARSF) (Yan & Zang, 2010). Before calculating FC matrices, all the preprocessed datasets were resliced into 3.5 mm^3 voxel size. We then applied band-pass filtering (0.01–0.08 Hz) to remove physiological noises, and linear detrend to remove slow drifts. We also regressed out variances of nuisance variables including 12 head motion parameters, global signals, white-matter signals averaged from the deep cerebral white matter, and cerebrospinal fluid signals averaged from the ventricles to reduce non-neuronal contributions to variable covariance. To avoid unwanted frequency components leaking back into the data, we applied the regression with all the nuisance regressors in a single one-step model (Jo et al., 2010; Jo et al., 2013). In addition, due to the controversy about including global signals in the regression analysis of the resting-state data (Gavrilescu et al., 2002; Saad et al., 2013), we also analyzed the data without global signal regression to examine the extent to which this preprocessing step influenced the predictive modeling and model generalization.

2.6 | FC matrix construction

We used a semantic brain template of 60 nodes to derive FC patterns within the semantic network for each participant. This template is constructed based on a meta-analysis deriving from 120 SP-related fMRI activation studies from Binder et al. (2009). These network nodes are frequently reported in a range of semantic task manipulations (e.g., words vs. pseudo words, semantic vs. phonological tasks, high vs. low meaningfulness conditions, etc.). All the regions are listed in Table S2. We defined a sphere with a radius of 6 mm for each node as a region of interest (ROI) and ensured no spatial overlapping among these ROIs. Representative mean time series of each ROI were extracted by averaging all voxels within that region. The interregional FC was estimated by calculating the pairwise Pearson correlation coefficient of the time series between each pair of ROIs. The FC was then normalized with Fisher's *r*-to-*z* transformation. Finally, a 60 × 60 symmetric connectivity matrix was obtained for each participant in each scanning session. For the HCP dataset, each participant had eight FC matrixes, separately calculated from the resting-state, language, working memory, gambling, motor, social cognition, relation processing, and emotion processing task fMRI data. For the SLD dataset, each participant has two FC matrixes (i.e., the resting state and the SLD task). For the ASC dataset, each participant has an FC matrix (i.e., the story comprehension task).

2.7 | Predictive model construction procedure (Phase 1)

We randomly split the HCP sample into two parts ($N = 439$ each) for predictive model construction and in-sample generalization estimation, respectively (see the overview of the analysis schema in Figure 1). We combined the 10-fold cross-validation (CV) with bootstrapping and permutation procedures to estimate CV prediction performances and construct a cross-validated SP prediction model (Feng et al., 2018; Feng et al., 2021). First, we randomly split the HCP model-construction sample ($N = 439$) into 10-folds. Nine folds of the subjects (i.e., training set) were used for model training, and the held-out fold (i.e., testing set) was used for model validation. To avoid overfitting the model with a large number of FCs, we employed a feature selection procedure to select the most informative features (i.e., network edges) during modeling training. For each training set,

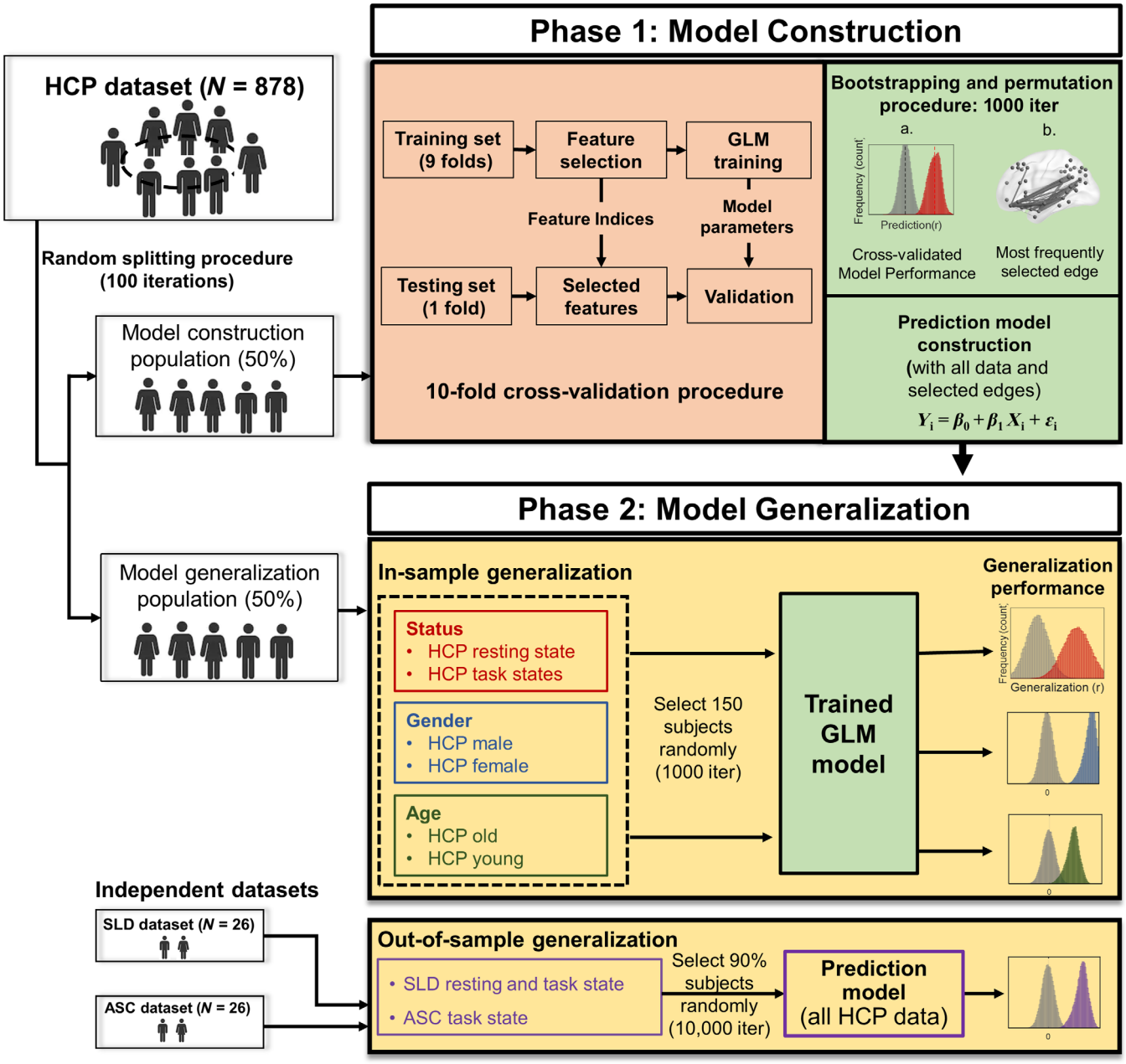

FIGURE 1 Schema of the predictive modeling and generalization evaluation procedure. The HCP participants were randomly split into two groups (50% each) for model construction (Phase 1) and generalization evaluation (Phase 2), respectively. This random splitting procedure was repeated 100 times to minimize sampling bias. For each splitting, at the model construction phase, we used 10-fold cross-validation (light red box), bootstrapping, and permutation procedures to construct and validate models (green box). At the end of Phase 1, a prediction model was built with all the Phase 1 samples and selected edges derived from the edgewise permutation test. At Phase 2, we generalized the prediction model to various independent HCP data and subpopulations, including different task states, gender, and age groups, to estimate the in-sample generalization performance. Another two independent datasets were used to assess out-of-sample model generalization (the bottom box). We randomly selected 90% of participants to calculate generalization performance, and this procedure was repeated 10,000 times for the nonparametric permutation test. iter = iteration

we computed the partial Pearson correlations between FC strengths and SP scores for each edge, where age and gender were controlled (Cai et al., 2020; He et al., 2020). We discarded the nonsignificant edges with a threshold of $p = .005$ while selecting the positively and negatively correlated edges separately as predictive features to train prediction models (Beaty et al., 2018; Rosenberg et al., 2016). The two types of edges were modeled separately because their connection patterns could contribute differently to the model prediction based on previous findings (Feng et al., 2021; Rosenberg et al., 2016). We also combined these two types of edges in model construction and generalization to examine whether the model performance would outperform models with only one type of edge. Different feature-

selection thresholds (e.g., 0.001 and 1% of the total number of edges) were also used to examine the reliability of the model prediction. Next, to further reduce the number of feature dimensions, we used principal component analysis (PCA) to summarize the main principal components (PCs) and selected the PCs that were significantly correlated with the SP scores with a threshold of $p = .05$. For positive prediction models, the selected components vary from 1 to 16 (mean = 5.29) across iterations. For negative predictive models, the selected components range from 1 to 16 (mean = 6.47) across iterations. These feature selection and dimension reduction procedures were conducted only in each training set independent of the held-out testing set. Finally, a linear prediction model was constructed with the selected PCs as predictive features and the latent SP factor as the dependent variable. This trained model was then applied to the held-out 10% unseen participants to predict their SP scores. This cross-validation process was repeated 10 times to predict all the participants' SP scores. Pearson correlations between the observed and predicted SP scores were used to assess the model prediction performance (i.e., $r_{[observed,predicted]}$).

We employed bootstrapping and permutation test procedures to evaluate the reliability and statistical significance of the model predictions. In the bootstrapping procedure, we repeated the 10-fold CV procedure 1000 times. Each CV repetition would result in a slightly different prediction performance $r_{[observed,predicted]}$ due to the sampling differences in partitioning the training and testing sets. Therefore, a prediction distribution was generated after 1000 CV iterations. To test whether the bootstrapping-based prediction distribution occurred by chance, we adopted a permutation test procedure (i.e., randomization test) where all participants' SP scores and their FCs were scrambled before being used for predictive modeling. Specifically, the FCs and the SP scores were permuted independently to generate a fully randomized data matrix. The 10-fold cross-validation procedure was repeated 1000 times with the randomized matrix to generate a null (chance) prediction distribution. Statistical significance of the model predictions was determined by comparing the medium $r_{[observed,predicted]}$ of the bootstrapping distribution with the permutation-based null distribution. The 95th percentile points of each null distribution were used as the critical values for a one-tailed test against the null hypothesis with $p = .05$. In addition, we obtained a set of significantly selected edges based on the feature selection procedure with an edgewise permutation test with a threshold of $p = .001$.

2.8 | Model generalization and evaluation (Phase 2)

At the model generalization phase, we built prediction models with all the significantly selected edges (positive- and negative-correlation edges separately) and model-construction samples to estimate generalization performance. Data with different fMRI states (i.e., resting and other task states) and subpopulation groups (i.e., age and gender) were selected from another half of the HCP sample (i.e., model-generalization samples) to examine to what extent these factors influence the model generalization. To this end, we randomly selected 150 participants' resting-state and task-state (including seven tasks) data as a generalization set for each generalization iteration. We chose 150 participants to ensure each subpopulation group had sufficient and equal participants for model generalization estimation. The models constructed in Phase 1 were then applied to the generalization set individually to estimate generalization prediction performances. This in-sample generalization procedure was repeated 1000 times to obtain a generalization prediction distribution for each model built in Phase 1.

Similarly, to estimate to which extent age and gender modulate the generalization, we randomly selected 150 females and males and 150 adults younger and older than 30 years respectively for generalization sets. The choice of 30 years old to split the data into two groups was to ensure each age group had an equal number of participants based on HCP's age classification (four age ranges: 22–25, 26–30, 31–35, and 36+). The same generalization procedure was applied to each of these subpopulations. The Pearson correlation between observed and predicted SP scores was used to assess model generalizability. To minimize sampling bias, we repeated the random splitting procedure (i.e., randomly splitting the HCP into construction and generalization samples) 100 times. A model was constructed for each split; therefore, 100 models were built at Phase 1. As a result, the accumulative iteration for cross-validation (Phase 1) and generalization (Phase 2) procedures was 100,000 times.

To estimate the out-of-sample model generalization performance, we used all the HCP participants to build a prediction model and generalize the model onto two independent datasets (i.e., SLD and ASC). Both datasets are previously published. These datasets were collected with different experimental settings (e.g., different semantic tasks, data collection procedures, MRI data acquisition, language use, etc.). We only used the language task fMRI data from the HCP dataset and significant edges selected from this task to build the prediction model for out-of-sample generalization because the language task data yielded the best CV and in-sample generalization performances. We then applied this HCP model to the two datasets to generate predicted SP scores. Pearson correlations between observed and predicted SP scores were used to assess the out-of-sample generalizability. We repeated this out-of-sample generalization and the corresponding permutation test procedures 10,000 times (with 90% of the samples each). Statistical significance was determined by comparing the bootstrapping distribution and the permutation-based null distribution. All the prediction analyses were performed with customized MATLAB scripts.

We used a Linux desktop workstation with two Intel Xeon CPUs (16-core processors each) to preprocess the data and build and validate the prediction models. It took about 12–60 h for each fMRI task (~60 h for HCP) for preprocessing and extracting the FC patterns for all participants. Thanks to the parallel computing technique, the prediction model construction and validation process were highly accelerated. It took around 6 h for the prediction model construction and about 2 h for generalization. Also, to facilitate the iterative

computational processes (i.e., bootstrapping and permutation procedures), we used two feature-selection techniques (Pearson's correlation and PCA) to remove noninformative features and reduce data dimensions. Thus, the model construction and validation were significantly accelerated.

2.9 | Predictive modeling with different brain network modules

To further estimate the prediction contributions of different intra- and internetwork connections, we conducted predictive modeling with the same bootstrapping and permutation procedures for each module of subnetwork connections. First, we classified the 60 semantic network nodes into three subnetworks based on previously defined network labels (Xu et al., 2016). Xu et al. performed a modularity analysis with the spectral optimization algorithm to detect network communities. They found three sub-networks, Perisylvian network (PSN), frontoparietal network (FPN), and default mode network (DMN), based on the resting-state FC patterns. We then separated the edges into six divisions based on their intra- and internetwork connectivity. Specifically, the six divisions (i.e., connection modules) include three modules of intranetwork connections (PSN–PSN, FPN–FPN, and DMN–DMN) and three modules of internetwork connections (PSN–FPN, PSN–DMN, and FPN–DMN). The same predictive modeling procedures were conducted for each connection module.

2.10 | Control analyses

We further applied our trained SP models to predict CC and MC scores to examine whether the SP prediction model is semantic-domain-specific in explaining individual differences in SP or domain-general so that the models can be used to predict other cognitive traits (i.e., CC and MC) that do not require manipulation and processing of semantic information. Moreover, we tested whether a domain-general model can be used to predict SP scores. To do so, we constructed two prediction models for the CC and MC scores with half of the HCP samples. We then estimated how well these two models could be used to predict SP scores with another half of the unseen HCP samples.

3 | RESULTS

3.1 | The latent core SP factor

The confirmatory factor analysis (CFA) was used to extract a latent variable to reflect SP from five behavioral variables (see Figure 2a for the distributions of the five test scores). Another two latent variables for cognitive control (CC) and motor control (MC) were also estimated for comparisons and control analyses. The three-factor CFA model has a robust model fit to the data ($\chi^2_{[41]} = 463.264$, $p = 1.00 \times 10^{-4}$, $RMSEA = 0.109$, $CFI = 0.750$). The five selected tests contributed differently to the SP latent variable (Figure 2b, left panel) as expected due to these tasks involving SP differently. In particular, the Picture Vocabulary (PicVoc: $R^2 = .755$), Oral Reading Recognition (ReadEng: $R^2 = .841$), and List Sorting (ListSort: $R^2 = .212$) tests contributed more to the SP latent scores than the other two tests (i.e., Picture Sequence Memory and Pattern Comparison Processing Speed tests). The SP latent variable shows a moderate-to-low correlation with the other two latent variables ($R^2_{[SP,CC]} = .169$; $R^2_{[SP,MC]} = .118$; see Figure 2b, right panel). The distribution of the SP scores across the entire HCP population is shown in Figure 2c. These SP scores were used for predictive modeling and model generalization estimation.

3.2 | Cross-validation performance in predicting latent SP scores

At the model construction phase, the SP prediction models built with FC features of the language task (LT) were significantly predictive of individual SP scores (see Figure 3a left panel for the predicted vs. observed scores). This cross-validation (CV) model performance was not only statistically significant but also reliable across CV repetitions, which was demonstrated by the bootstrapping and permutation procedures with 100,000 iterations (the positive-predictive model: $r_{pos[observed,predicted]} = .322$, $p = 2.00 \times 10^{-5}$; the negative-predictive model: $r_{neg[observed,predicted]} = 0.317$, $p = 2.00 \times 10^{-5}$; Bonferroni-corrected; Figure 3a, right panel).

We also reconducted the CV predictive modeling with FC data that the global signals were not regressed out from each ROI's time series. We found that the CV prediction performances without global signal regression (see Figure S1A for the results) are consistent with the results shown in Figure 3. The contributing edges were largely overlapped (Figure S1B), which demonstrated that global signals do not significantly influence the prediction performance of our model. In addition, when we applied the same CV predictive modeling to the PicVoc and ReadEng raw scores (the top two behavioral metrics contributing to the latent SP factor estimation), we found the results were consistent with that of SP scores (see Figures S2 and S3 for details).

The model performance with both types of edges was also significantly better than chance ($r_{all[observed,predicted]} = 0.318$, $p = 2.00 \times 10^{-5}$) but it did not outperform the positive-predictive model (model comparison: $p = .533$; nonparametric permutation test) or the negative-predictive model ($p = .494$). The CV prediction performance of the LT data outperformed models of the resting state and other tasks (see Figure 3b; LT vs. resting state [Rest]: $p = .005$; LT vs. relational task [RT]: $p = .005$; LT vs. gambling task [GT]: $p = 7.07 \times 10^{-5}$; LT vs. emotion task [ET]: $p = 7.03 \times 10^{-5}$; LT vs. working memory task [WM]: $p = .014$; LT vs. motor task [MT]: $p = .153$; LT vs. social task [ST]: $p = .191$; nonparametric tests with Bonferroni corrected p).

We identified connections (i.e., edges) that were significantly selected during the CV model training with the edgewise permutation

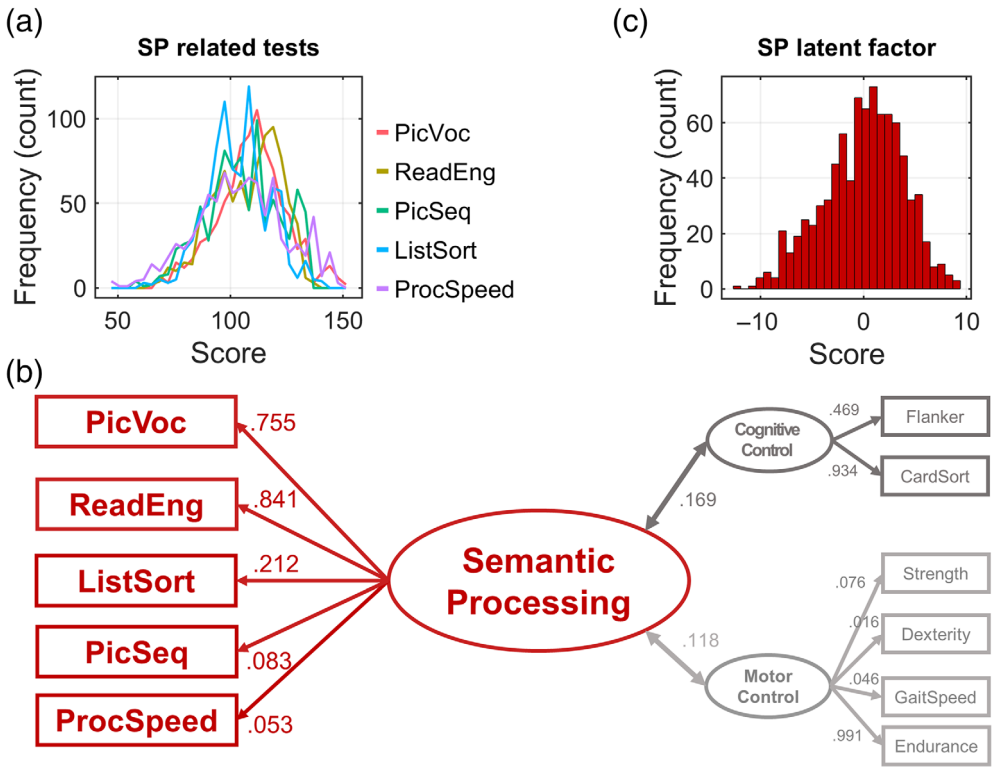

FIGURE 2 Confirmatory factor analysis (CFA) was used to estimate the latent semantic processing (SP) component and another two putative control components. (a) Test score distributions for the five SP-related tests. The five tests include oral reading recognition (ReadEng), picture vocabulary (PicVoc), list sorting (ListSort), picture sequence memory (PicSeq), and pattern comparison processing speed (ProcSpeed) (see detailed descriptions of the tests in Table S1). These score distributions were shown in different colors for different tests. (b) A three-factor CFA model (SP, CC, and MC components) was constructed and estimated. The number beside each line denotes each test variable's contribution (R^2) to explaining the latent factors. (c) The score distribution of the latent SP factor was displayed in the histogram.

test. We classified the significantly selected edges (i.e., predictive edges) into two types (i.e., positively and negatively predictive edges). The positively predictive edges indicate that increased FC strengths are associated with enhanced SP ability. These positive predictive edges are mainly those PSN intranetwork connections (selection rate = 38.4%) and PSN–DMN (32.1%) and PSN–FPN (14.3%) internetwork connections (see Figure 3c, left panel), notably including the connections between the orbital part of the inferior frontal gyrus (IFG), temporal pole (TP), middle temporal gyrus (MTG), and angular gyrus (AG). In contrast, the negative predictive edges indicate that increased FC strengths are associated with poorer SP ability. We found that the negative predictive edges were distributed across network modules, mainly those PSN–DMN (31.9%), PSN–FPN (26.3%), DMN–DMN (18.7%), and FPN–DMN (17.0%) connections, mainly including edges between an FPN node, inferior parietal lobule (IPL), and PSN temporal nodes, as well as FPN intranetwork connections (e.g., an edge between IFG and IPL) (Figure 3c, right panel; all the significantly predictive edges were listed in Table S3).

To further examine the contributions of the edges in the model construction while minimizing the influence of the highly correlated FC edges, we conducted the predictive modeling for the SP scores with the ridge regression approach (Gao et al., 2019). The ridge regression assigns a coefficient to each selected edge and shrinks the regression coefficients of correlated edges to deal with the collinearity problem. We found that the CV prediction performances and contributing edges revealed by the ridge regression approach were consistent with that of the present approach (see Figure S4 for details).

With predictive modeling for each module of connections (i.e., PSN–FPN, PSN–DMN, PSN–PSN, FPN–DMN, DMN–DMN, and FPN–FPN; see Figure 4a for the connection patterns), we found that different modules showed varying degrees of predictive power (Figure 4b). The significant positively predictive edges are those module connections between PSN and the other two subnetworks as well as the PSN intranetwork connections ($r_{PSN-FPN}$ = .220, $p = 6.00 \times 10^{-5}$; $r_{PSN-DMN}$ = .220, $p = 6.00 \times 10^{-5}$; $r_{PSN-PSN}$ = .219, $p = 1.52 \times 10^{-4}$; $r_{FPN-DMN}$ = .176, p = .002; $r_{DMN-DMN}$ = .154, p = .007; $r_{FPN-FPN}$ = .075, p = .300; Bonferroni corrected p; Figure 4b). The significant negatively predictive edges are those internetwork connections between PSN and DMN, PSN and FPN, and FPN and DMN as well as the DMN and FPN intranetwork connections ($r_{PSN-FPN}$ = .235, $p = 6.00 \times 10^{-5}$; $r_{PSN-DMN}$ = .199, $p = 4.80 \times 10^{-4}$; $r_{PSN-PSN}$ = .066, p = .587; $r_{FPN-DMN}$ = .223, $p = 6.09 \times 10^{-5}$; $r_{DMN-DMN}$ = .203, $p = 3.06 \times 10^{-4}$; $r_{FPN-FPN}$ = .151, p = .007; Bonferroni corrected p) (Figure 4b). Detailed predictive edges for each module were displayed in Figure 4c. These

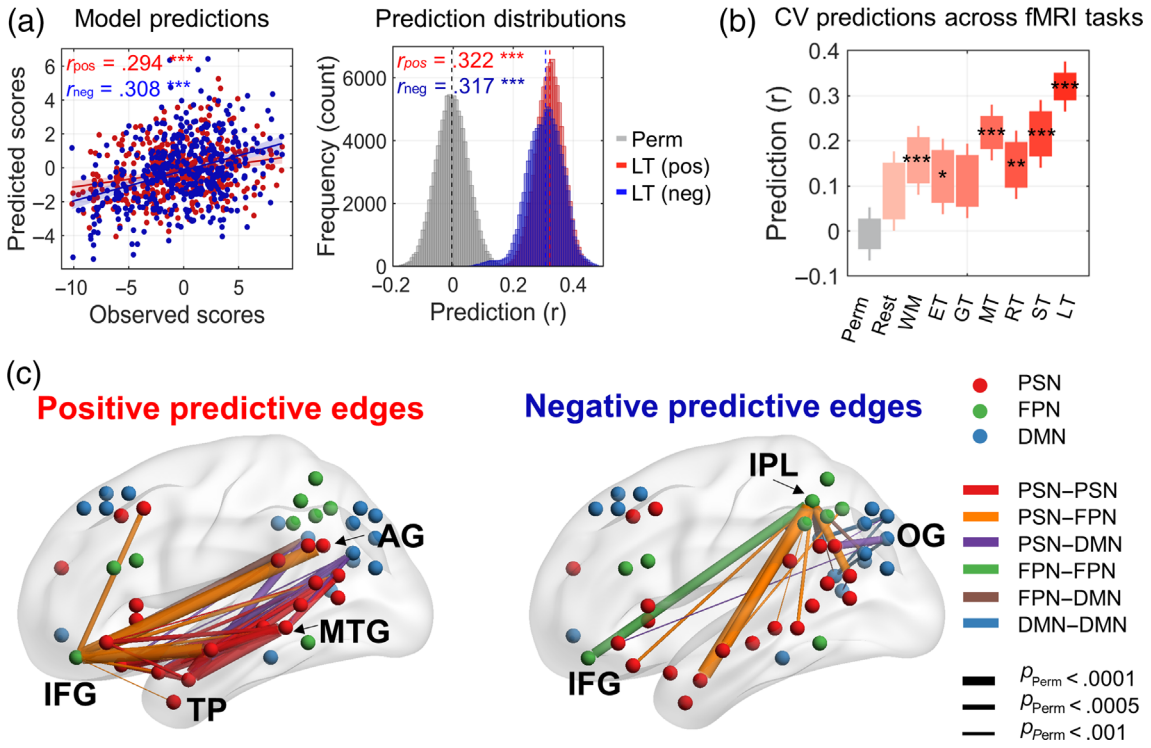

FIGURE 3 Cross-validated model prediction performance and the significantly contributing network connections. (a) SP prediction performance at the model construction phase. Left panel, scatter plot shows linear correlations between predicted and observed SP scores; right panel, bootstrapping- and permutation-based prediction distributions derived from the positively (red) and negatively predictive (blue) FCs, respectively, with language task (LT) data. Perm = permutation-based distribution. (b) Prediction distributions across fMRI states. Each box in the boxplot represents the quartile of each prediction distribution. Permutation test: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .005$, Bonferroni corrected. Data abbreviations: ET, emotion task; GT, gambling task; LT, language task; MT, motor task; Rest, resting state; RT, relational task; ST, social task; WM, working memory task. (c) Significantly predictive edges that contribute to the model prediction. The thickness of the edges denotes the statistical significance derived from the edgewise permutation test. Left panel, significant positive-predictive edges; right panel, significant negative-predictive edges. Node colors denote network modules; edge colors denote inter- and intranetwork connection modules. Node abbreviations: AG, angular gyrus; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; MTG, middle temporal gyrus; OG, occipital gyrus; TP, temporal pole.

module-based prediction results are consistent with the overall edge-wise prediction results described in Figure 3c.

3.3 | Model generalization in predicting independent HCP samples

We constructed predictive models with all the samples used at Phase 1 and the significantly selected edges derived from the edgewise permutation test ($p < .001$) to estimate model generalization. We used the language task's fMRI data to build the models for generalization because the CV performance of the language task data outperformed the resting state and other tasks (see Figure 3b). The predictive models' in-sample and out-of-sample generalization performances were assessed with another half of the unseen HCP dataset and the two independent datasets, respectively.

For in-sample generalization, different task states (i.e., resting-state and seven tasks) and subpopulations (i.e., two age groups and two gender groups) of the unseen HCP dataset were used to estimate to which extent these variables affect the generalization performance. First, we generalized our trained model from the language task data to the data derived from different fMRI states. We found that the positively predictive model significantly generalized to unseen individuals for the language task data (LT: median $r = .249$, $p = .0068$; permutation test; the same for the following tests) and the motor task (MT: $r = .208$, $p = .037$), but not for the other states (Rest: $r = .105$, $p = .792$; WM: $r = .157$, $p = .200$; ET: $r = .132$, $p = .405$; GT: $r = .139$, $p = .341$; RT: $r = .092$, $p = .999$; ST: $r = .154$, $p = .226$; Bonferroni corrected p). The generalization performance of the language task was significantly better than the resting state ($p = .016$; uncorrected). However, no significant difference was found between the language task and most of the other task states in generalization (LT vs. WM: $p = .090$; LT vs. ET: $p = .044$; LT vs. GT: $p = .055$; LT vs. MT: $p = .188$; LT vs. RT: $p = .017$; LT vs. ST: $p = .067$; uncorrected p). These results indicated that the language task not only selectively enhances CV prediction but also in-sample model generalization (Figure 5a, left panel). To further explore which module of connections contributed to this task effect in generalization, we conducted the model generalization analysis for each network module individually. We found that mainly those edges connecting to the PSN yielded

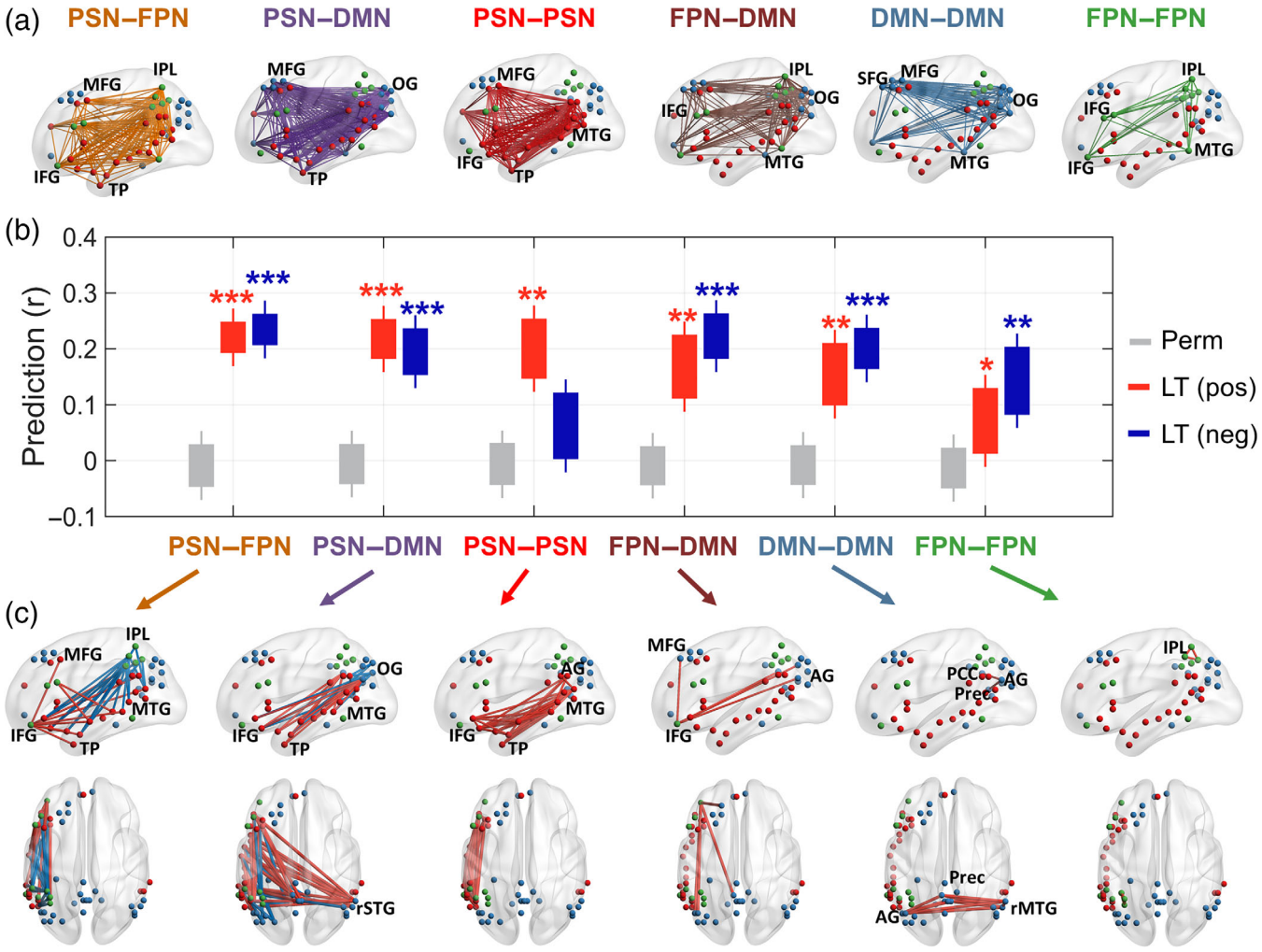

FIGURE 4 Predictive modeling (Phase 1) for each module of connections. (a) Six connection modules. Key regions in each module were labeled. (b) Prediction powers for different modules and types of edge. The red and blue boxes represent the predictive powers of positive- and negative-predictive models, respectively. The order of the modules was sorted by the mean prediction *r*-value of the positive models in descending order. Perm, permutation-based distribution. $^{**}p < .01$; $^{***}p < .005$, Bonferroni corrected. (c) Significantly predictive edges for each connection module. The red-colored and the blue-colored edges represent the significant positively and negatively predictive connections, respectively. Edgewise permutation test, $p < .005$. Node abbreviations: AG, angular gyrus; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; MFG, middle frontal gyrus; MTG, middle temporal gyrus; OG, occipital gyrus; PCC, posterior cingulum cortex; Prec, precuneus; rMTG, right middle temporal gyrus; rSTG, right superior temporal gyrus; TP, temporal pole.

significant model generalization ($r_{PSN-PSN} = .248$, $p = .005$; $r_{PSN-FPN} = .239$, $p = .007$; $r_{FPN-DMN} = .199$, $p = .020$; Bonferroni corrected *p*). While the PSN intranetwork connections (i.e., PSN–PSN) contributed most to the generalization, the model generalization of the PSN–PSN connections in the language task also outperformed that in the resting state (LT vs. Rest: $p = .039$, uncorrected; Figure 5a, right panel).

We further examined the in-sample generalization performance to different gender and age groups of the HCP dataset for the positive predictive models (see Figure 5b,c). We selected these subpopulations' resting-state and language-task data as the target for generalization. We examined to which extent age and gender variables modulate the generalizability of the SP prediction model. For the language task data, the trained model significantly generalized to predict SP scores of both females ($r = .359$, $p = 2.00 \times 10^{-5}$, Bonferroni corrected) and males ($r = .174$, $p = .030$; Bonferroni corrected). The model's generalizability was significantly higher for females than males only for the language task data ($p = .039$; Bonferroni corrected) (Figure 5b, left panel). Those edges connecting to the PSN nodes yielded significant model generalization only for females ($r_{PSN-PSN} = .332$, $p = 1.20 \times 10^{-4}$; $r_{PSN-FPN} = .279$, $p = .002$; $r_{PSN-DMN} = .240$, $p = .009$; $r_{FPN-DMN} = .190$, $p = .045$; Bonferroni corrected). The gender effect in generalization was most prominent in the PSN intranetwork connections (males vs. females: $p = .014$, uncorrected) (Figure 5b, right panel), similar to the task effect. No significant generalization prediction was observed for resting-state data for either gender group (female: $r = .140$, $p = .080$; male: $r = .040$, $p = .624$; corrected). We further tested whether the homogeneity in

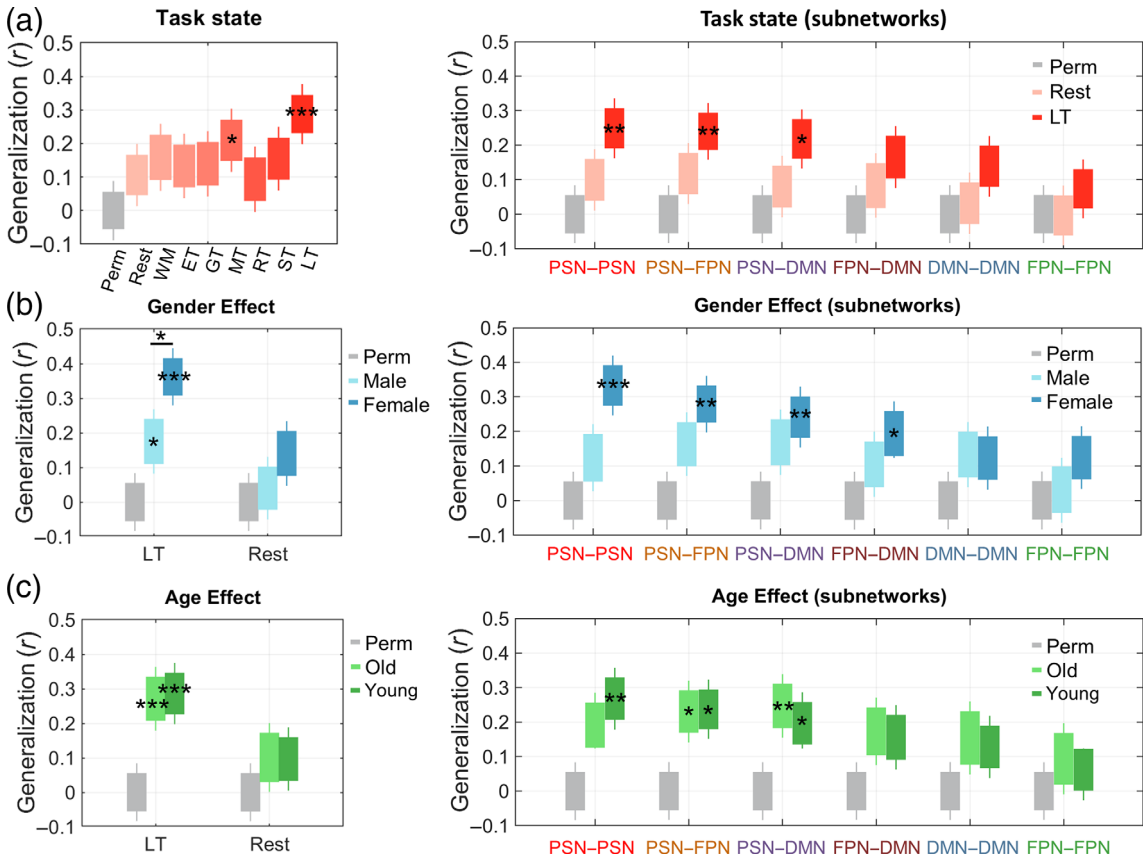

FIGURE 5 The in-sample generalization performance of the positive predictive models was modulated by task and gender. (a) Model generalization performance was estimated with independent HCP samples across different task states. ET, emotion task; GT, gambling task; LT, language task; MT, motor task; Perm, permutation; Rest, resting state; RT, relational task; ST, social task; WM, working memory task. (b) Predictive model generalization performance was estimated with independent HCP samples of the two gender groups, respectively. (c) model generalization performance for the two age groups. Right panels in (a–c) show generalization evaluation for each of the six connection modules. These connection modules are sorted according to the significance of the task state effect in descending order. Boxplots represent 25th and 75th percentiles (box) and range (whiskers) for the distributions. Asterisks indicate the significance of the generalizability (vs. null distributions): $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .005$; Bonferroni corrected.

FC strengths within the females was significantly less than that within the males. We randomly selected 100 males and females and calculated their standard deviations (SD; a proxy of inhomogeneity) of the FC strengths of the significantly predictive edges (displayed in Figure 3b). This process was repeated 10,000 times to estimate SD distributions. We found that the SD of the FC strength for females was significantly smaller than that for the males ($p = .001$, nonparametric test), suggesting that females' SP-related FCs are more homogeneous than males.

Moreover, for both age groups, significant generalization predictions were found only for the language task data (young: $r = .285$, $p = 2.60 \times 10^{-4}$; old: $r = .269$, $p = 5.80 \times 10^{-4}$, corrected p), consistent with the overall generalization results. However, no significant difference in generalization was found between the younger and older participants, neither for the language task (corrected $p = .874$) nor for resting-state data (corrected $p = .873$) (Figure 5c).

For the generalizability of the negative-predictive models, both task status and gender effects were found (Figure S5). First, significant in-sample generalization (vs. the chance distribution) was found only for the language task data ($r = .316$, $p = 3.20 \times 10^{-4}$, Bonferroni corrected), but not for any of the other fMRI states ($ps > .1$). The generalization performance of the language task data was significantly better than the resting state ($p = .014$, Bonferroni corrected) and most of the other task states (LT vs. ST: $p = .018$; LT vs. ET: $p = .037$; LT vs. GT: $p = .070$; LT vs. MT: $p = .041$; LT vs. RT: $p = .039$; LT vs. WM: $p = .160$; Bonferroni corrected; see Figure S5A, left panel). We further found that most of the subnetwork connections (i.e., PSN–FPN, PSN–DMN, DMN–DMN, FPN–DMN, and FPN–FPN) except the PSN intranetwork connections (i.e., PSN–PSN) contributed to the task enhancement in model generalization (LT vs. Rest: $p_{PSN-FPN} = .023$, $p_{PSN-DMN} = .013$, $p_{DMN-DMN} = .024$, $p_{FPN-DMN} = .016$, and $p_{FPN-FPN} = .031$; uncorrected p) (Figure S5B, right panel). However, these comparisons were not significant after multiple comparisons correction (LT vs. Rest: $p_{PSN-FPN} = .138$, $p_{PSN-DMN} = .075$, $p_{DMN-DMN} = .142$, $p_{FPN-DMN} = .096$, and $p_{FPN-FPN} = .183$). For the generalization in different gender and age groups, the model's generalizability was significantly higher for females than males only for the language task data ($p = .043$, uncorrected) (Figure S5B, left panel).

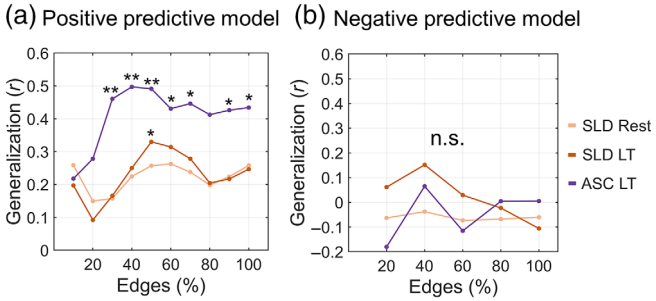

FIGURE 6 Out-of-sample generalization performance for two independent datasets. (a) The out-of-sample generalization of the positive-predictive model built from the HCP dataset. The line chart illustrates the out-of-sample generalization performance across a different percentage of predictive features (i.e., edges). (b) The out-of-sample generalization performance of the negative-predictive model. Dataset abbreviations: ASC LT, language-task data from the ASC dataset; n.s., not significant; SLD LT, language-task data from the SLD dataset; SLD rest, resting-state data from the SLD dataset. Edges (%), percentage of significant edges used to build an HCP model for generalization. Permutation test (vs. null distributions): $^{*}p < .05$; $^{**}p < .01$; uncorrected *p*.

However, this gender effect in generalization did not reach significance in any module of the connections (Figure S5B, right panel). However, the quantity differences between females and males were observed across modules except for PSN's intranetwork connections. Also, no significant difference in generalization was found between the younger and older groups, neither for the language-task ($p = .660$, corrected) nor resting-state ($p = .495$, corrected) data (Figure S5C).

3.4 | Out-of-sample model generalization evaluation

To further examine to which extent the cross-validated models can be generalized to independent datasets (i.e., out-of-sample generalization) collected in different populations and with different task settings and SP measures, we generalized our HCP model onto two independent datasets to predict their semantic performances: the semantic priming scores and the story comprehension scores. The out-of-sample model generalization was often unsuccessful in previous attempts. Therefore, we considered this analysis as an exploratory test and selected a wide range of percentages of significant edges as FC features to build HCP models to examine whether the number of connections modulates the generalization performance. We found that only the positive-predictive model can significantly predict unseen individuals' reading comprehension scores in the ASC dataset across a wide range of edge selection percentages (e.g., 50% edge: $r = .491$, $p = .005$, uncorrected) (Figure 6a). However, the positive HCP model can only marginally significantly predict semantic priming scores in the SLD dataset when 50% of the predictive edges are used ($r = .330$, $p = .058$, uncorrected). No significant out-of-sample generalization was found when we applied the negative-predictive models (Figure 6b). We also tried to combine both the positive and negative models for generalization. However, the model generalization performance did not outperform the positive-predictive model alone.

3.5 | Control analyses

To further examine to which extent the HCP model is domain-specific (i.e., selectively predicting SP scores instead of other components), we applied the built SP model to predict CC and MC scores. We found that the positive predictive model was not significantly predictive of CC or MC scores (Figure S6A, upper panel). The negative-predictive model was weakly predictive of CC but not MC's (Figure S6A, lower panel). These results indicate that the generalizable positive-predictive SP model is semantic-domain-specific in predicting individual SP abilities.

In addition, we further examined whether the CC and MC prediction models built with half of the HCP sample can be used to predict the SP scores from another half of the participants. We ran the predictive modeling analysis (including the Phase 1 model construction and Phase 2 in-sample generalization) with the CC and MC scores. The models trained with the CC and MC scores cannot significantly predict held-out participants' SP scores (Figure S6B,C). These findings provide converging evidence supporting the specificity of the SP model and the FC patterns underlying individual differences in SP.

4 | DISCUSSION

The present study used the connectome-based predictive modeling approaches and datasets with sizable samples to examine the functional network organizations underlying individual differences in SP ability. We constructed predictive models with rigorous cross-validation, bootstrapping, and permutation procedures. We then assessed the model's generalizability with unseen in-sample and out-of-sample datasets. We demonstrated the robust relationships between individual differences in SP and variabilities in FC organization while overcoming the low effect size inherent by small sample sizes and traditional correlational approaches. We identified a cluster of intra- and internetwork connections where their variabilities contributed significantly to predicting individual SP scores. Increased FCs both within the Perisylvian network (PSN) and between PSN and other subnetworks are predictive of superior SP ability. In contrast, increased FCs between a frontoparietal network (FPN) node, inferior parietal lobule, and other subnetworks are predictive of poorer SP. These predictive relationships were enhanced when subjects participated in a language task comparing resting-state and other tasks. This task-specific enhancement in prediction is more prominent for females than males. Also, the SP prediction model built with HCP samples has the potential to generalize to independent datasets that used very different neuroimaging and behavioral paradigms. These findings—that connectome-based models predict different measures of SP across different populations—provide significant insights into

our understanding of the neural network organizations underlying individual differences in SP and reveal detailed effects on how task and demographic factors modulate model generalization performances.

The current models, which are constructed by the HCP dataset and capable of generalizing to different in-sample and out-of-sample datasets, make significant progress toward identifying neuromarkers of SP. SP is not a unitary ability; instead, it has been related to various cognitive and control components. Therefore, it is challenging to select an unbiased behavioral test that can only reflect the core composition of SP while minimizing the test- or task-specific components so that the built model can be generalized to other semantic task contexts. Here we used a new analytic strategy by selecting five different SP-related behavioral measures and using confirmatory factor analysis to extract a core latent SP component instead of one task. This approach could minimize biases in task selection and maximize the representation of the latent SP measure. Also, we included two putative domain-general latent factors, cognitive control (CC) and motor control (MC), where their test materials do not involve any semantic stimuli, and their tasks do not require any SP-related processes. Additional control analyses confirmed that neither the CC nor the MC model was predictive of SP scores, while the SP models were not consistently predictive of CC or MC scores. These findings suggest that the current SP models and the underlying connectivity organization are mainly specific to the SP component that requires the manipulation and process of semantic or conceptual information. Nevertheless, we cannot rule out the possibility that some of the connectivity organization patterns (e.g., PSN) found here may also reflect joint contributions of multiple cognitive components (e.g., semantics-phonology or semantics-working memory) tightly linked to the SP. Future studies should be conducted with more component-specific behavioral tests to isolate their underlying FC organization.

In addition to demonstrating the prediction performance of the SP model, we further reveal that FCs connecting to the PSN nodes play a critical role in the model prediction and generalization, especially the interplay between the FPN and PSN. Increased FCs within PSN and between PSN and FPN are associated with superior SP ability, especially between the inferior frontal gyrus (IFG) and distributed temporoparietal PSN nodes. Previous studies have shown that both structural and FC properties of the PSN were associated with semantic task performance (Bookheimer, 2002; Saur et al., 2008). The predictive PSN intranetwork connections mainly included the connections between the left IFG and the left middle temporal regions (e.g., anterior to posterior). These regions are the main constituents of the canonical language system that is more activated by language tasks than control tasks (Fedorenko et al., 2011; Fedorenko & Thompson-Schill, 2014). For example, the left middle temporal gyrus (LMTG) has been proposed as a hub in the semantic and language network. LMTG has widely distributed connections with other language areas (Turken & Dronkers, 2011). The FCs between the LMTG and other regions, including the left IFG and dorsal lateral and medial prefrontal cortex are associated with individual differences in semantic behaviors (Jackson et al., 2016; Wei et al., 2012). Also, our results are consistent with previous studies on aphasia patients and stroke patients, where a selective disruption of IFG and impaired connections between IFG and left anterior superior temporal regions were associated with semantic impairments (Meinzer et al., 2011). The predictive FCs found in PSN and FPN could potentially be neural indicators of language/semantic impairments/disorders and neural predictors for future intervention since these connections have been demonstrated to be tightly linked to various language functions (e.g., semantic and syntactic processes) (Badre et al., 2005; Krieger-Redwood et al., 2016; Papoutsi et al., 2011; Snijders et al., 2010; Tyler & Marslen-Wilson, 2007; Vatansever et al., 2017).

It is worth noting that the Perislyvian regions and their predictive connections found here may not just reflect individual differences in SP. They may also reflect the interactions between semantics and other language and cognitive components, such as semantics-phonology and semantics-working memory interactions. The predictive network hubs include the orbital IFG and inferior parietal regions. These frontoparietal regions were previously related to working memory and phonological processing of language stimuli (Ardila et al., 2016; Price & Devlin, 2011). Also, the ReadEng task, weighted highest in the latent SP factor, has been shown to require phonological decoding processing (Dickens et al., 2019). Therefore, the predictive PSN connections may also reflect interaction components (e.g., semantic-phonology or semantic-working memory). Future works should be conducted with additional tests designed for measuring phonology and working memory to test this possibility.

We also show that increased FCs within FPN and between FPN and other subnetworks is predictive of poorer SP ability. One critical node in FPN, the left inferior parietal lobue (LIPL), significantly contributes to the negative predictions. Increased FC strengths between the LIPL and a range of PSN nodes across the inferior frontal and middle temporal regions are associated with decreased SP ability. LIPL has been characterized as a provincial hub and internetwork connector. This region has been proposed as connecting the semantic control system in FPN with a putative language-based semantic system in PSN and a memory-based simulation system in DMN (Xu et al., 2016). Increased connectivity between the LIPL and the other two systems may be related to decreased efficiency in SP. For example, in a challenging semantic retrieval task, people with superior SP ability can solve the task relatively quickly and efficiently where they may only rely on increased focal FCs between the frontotemporal PSN regions. However, people with poor SP ability may need additional assistance from other semantic systems, especially the FPN control system to solve the difficult task and the DMN memory-based system for retrieving additional information. Thus, increased FCs between the internetwork connector LIPL and the other two systems may be a compensation mechanism for those with poor SP. They need to retrieve additional task-related SP information to make a decision. This compensation interpretation may also explain why increased FCs between PSN posterior temporal regions and DMN occipital regions are associated with poorer SP ability. Future studies should further explore and test this possibility.

We further demonstrate that the built SP models can be generalized to unseen independent in-sample and out-of-sample data. We also explore to which extent task states and demographic factors (i.e., gender and age) modulate the generalization performance. We found that both task and gender were critical factors affecting how well a trained SP model derived from one sample can be generalized to unseen samples. In particular, using a language task significantly enhances the generalization performance over resting-state and other task-state data. This language-task enhancement effect is consistent with previous findings from group-level FC studies where they found increased FCs between the frontoparietal cognitive-control regions and language areas when subjects performed language tasks (Cole et al., 2014; Di et al., 2013; T. Jiang et al., 2004; Smith et al., 2009). For individual prediction findings, previous studies show that cognitive tasks enhance individual differences of fluid intelligence in FC patterns, such that predictive models built from task fMRI data outperform models built from resting-state data (Finn et al., 2017; Greene et al., 2018; R. Jiang et al., 2020). Consistent with and moving beyond this finding, we demonstrate here that only language tasks can amplify individual differences in FC patterns more tightly related to individuals' behavioral semantic performance than other tasks. We demonstrate that this language-task-specific amplification effect can be observed not only for cross-validation prediction but also for model generalization. One possibility of this task-amplification effect in model prediction is that the language-task-induced functionally relevant FC changes over resting state and other task states, which subserve behavioral performance of the semantic tasks not only at the group level but also at individual differences level. During a language task, the FC patterns reorganize based on the task demand for optimal processing of the language stimuli in hand (Feng et al., 2015), especially semantic information. Therefore, the individual differences in FC during language tasks are best associated with behavioral test scores that involve similar semantic processes. In contrast, resting-state FCs are unconstrained, and it is likely to involve many SP-irrelevant components, for example, mind wandering (Godwin et al., 2017), arousal (Koike et al., 2011), attention (Bonnelle et al., 2011), and different levels of conscious thoughts (Smallwood et al., 2012).

In addition to the task effect, we also demonstrate sex differences in model generalization. We found that model generalization to females was more robust than males. This gender effect may be due to sex differences in SP-related network organization, task-induced neural activation patterns, or both (Baxter et al., 2003; Biswal et al., 2010; Satterthwaite et al., 2015; Scheinost et al., 2015). For example, previous studies have revealed that females had more focal activation in the left hemisphere and greater right posterior temporal and insula activations during SP-related tasks than males (Baxter et al., 2003). In FC patterns, detectable sex differences were found across network modules and seed-based connectivity strengths (Biswal et al., 2010). Consistent with these findings, recent FC studies demonstrate that multivariate resting-state FC patterns are associated with individuals' cognitive profiles of "male" and "female" (Satterthwaite et al., 2015). Extending these previous findings, we demonstrate that predicting unseen females' SP scores is better than that of males with their FC patterns. The sex differences in neurocognitive measures reported in previous studies may not fully explain the underlying cause of the generalization differences found here. One potential source of the sex differences in model generalization is FC homogeneity. We show that FCs between females are significantly more homogeneous (i.e., less interindividual variability) than males. The differences in FC homogeneity between females and males may be one potential factor in the gender differences in model generalization performance. This speculation suggest that researchers may consider sample homogeneity when building prediction models. The group-specific model may yield superior prediction performance than group-general models. Future studies should be conducted to systematically investigate how feature homogeneity modulates the generalization of a prediction model.

Moving beyond cross-validation prediction, we adopted in-sample and out-of-sample generalization estimation with independent datasets. Overestimation of prediction performance is commonly found with traditional correlational approaches and small sample sizes. At the same time, there is a failure to maintain the independence of training and testing datasets. To ensure data independence, we ensure that model validation and estimation were true tests of the models' ability to generalize to unseen subjects at every analysis step. We defined the model predictions based on the levels of generalization to unseen subjects (i.e., level 1: cross-validation; level 2: in-sample generalization; level 3: out-of-sample generalization) with an increasing level of data independence. For prediction with the out-of-sample generalization procedure, the model built with HCP samples and the positive SP-related FCs has the potential to be generalized to unseen datasets with different settings, such as different populations, tasks, languages, data acquisition, and MRI scanners. This finding implies that the SP prediction model captures the core and maybe universal relationships between individual SP variability and FC organizations.

Nevertheless, there is a limitation to the out-of-sample generalization estimation in the current study. We only included two previously published datasets as out-of-sample datasets. While it is successful in generalization for one independent dataset, it is not relatively reliable in predicting SP scores in another. The sample sizes of the independent datasets are relatively small, which could limit our examination of the out-of-sample model generalization. Future studies need to assess the out-of-sample generalization performance with more and sizable datasets and further examine what factors modulate the generalization across datasets.

The model generalization performance based on combined positive and negative edges did not significantly outperform the models with only positive- or negative-predictive connections. It may imply that simply adding up the two types of features linearly to train a model is not an effective (although efficient in computation) way to boost the model generalization further. Further studies may need to solve this technical challenge when combining different predictive features in a model to improve generalization. Another possibility is that the negative connections may not be essential and unique in explaining individual differences in SP behaviors; therefore, they do not contribute to the model generalization when adding these

negative edges. Consistent with this speculation, the negative-predictive model was not successfully generalized to the out-of-sample datasets.

5 | CONCLUSIONS

In summary, we show that FC patterns play a critical role in explaining the individual differences in SP ability. The SP prediction model constructed from the HCP dataset has the potential to generalize to independent cohorts with different experimental settings, suggesting potentially robust model reliability and generalization. FCs connecting to the Perisylvian network show the most reliable contributions to predictive modeling and model generalization. These findings contribute to our understanding of the neural sources of individual differences in SP, which potentially lay the foundation for personalized education and improve intervention outcomes for SP and language deficits patients.

ACKNOWLEDGMENTS

Data collection and sharing for this project was provided by the MGH-USC Human Connectome Project (HCP; Principal Investigators: Bruce Rosen, M.D., Ph.D., Arthur W. Toga, Ph.D., Van J. Weeden, MD). HCP funding was provided by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). HCP data are disseminated by the Laboratory of NeuroImaging at the University of Southern California.

CONFLICT OF INTEREST

Patrick C. M. Wong is a founder of a company in Hong Kong supported by a Hong Kong SAR government startup scheme for universities.

DATA AVAILABILITY STATEMENT

MATLAB scripts for the main predictive modeling analysis can be found at the OSF website (https://osf.io/b9h2x/). MATLAB scripts written to perform the preprocessing and additional control analyses are available from the authors upon request. The HCP data supporting the findings of this study are publicly available on the ConnectomeDB database (https://www.humanconnectome.org). The meta-data and functional connectivity matrices can be found at https://osf.io/b9h2x/.

ORCID

Suiping Wang https://orcid.org/0000-0001-8726-169X
Patrick C. M. Wong https://orcid.org/0000-0002-6105-5027
Gangyi Feng https://orcid.org/0000-0003-2239-5296

REFERENCES

Allen, C. M., Martin, R. C., & Martin, N. (2012). Relations between short-term memory deficits, semantic processing, and executive function. *Aphasiology*, *26*, 428–461.

Ardila, A., Bernal, B., & Rosselli, M. (2016). How localized are language brain areas? A review of Brodmann areas involvement in oral language. *Archives of Clinical Neuropsychology*, *31*, 112–122.

Badre, D., Poldrack, R. A., Pare-Blagoev, E. J., Insler, R. Z., & Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron*, *47*, 907–918.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., … WU-Minn HCP Consortium. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, *80*, 169–189.

Bauer, P. J., Dikmen, S. S., Heaton, R. K., Mungas, D., Slotkin, J., & Beaumont, J. L. (2013). III. NIH toolbox cognition battery (CB): Measuring episodic memory. *Monographs of the Society for Research in Child Development*, *78*, 34–48.

Baxter, L. C., Saykin, A. J., Flashman, L. A., Johnson, S. C., Guerin, S. J., Babcock, D., & Wishart, H. A. (2003). Sex differences in semantic language processing: A functional MRI study. *Brain and Language*, *84*, 264–272.

Beaty, R. E., Kenett, Y. N., Christensen, A. P., Rosenberg, M. D., Benedek, M., Chen, Q., Fink, A., Qiu, J., Kwapil, T. R., Kane, M. J., & Silvia, P. J. (2018). Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 1087–1092.

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences*, *17*, 89–98.

Bhattasali, S., Brennan, J., Luh, W.-M., Franzluebbers, B., & Hale, J. (2020). The Alice Datasets: fMRI & EEG Observations of Natural Language Comprehension. Paper presented at the Proceedings of The 12th Language Resources and Evaluation Conference.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*, 527–536.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*, 2767–2796.

Binder, J. R., & Fernandino, L. (2015). *Semantic processing brain mapping: An encyclopedic reference* (Vol. 3). Academic Press.

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., Dogonowski, A. M., Ernst, M., Fair, D., Hampson, M., Hoptman, M. J., Hyde, J. S., Kiviniemi, V. J., Kötter, R., Li, S. J., … Colcombe, S. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, *107*, 4734–4739.

Bonnelle, V., Leech, R., Kinnunen, K. M., Ham, T. E., Beckmann, C. F., De Boissezon, X., Greenwood, R. J., & Sharp, D. J. (2011). Default mode network connectivity predicts sustained attention deficits after traumatic brain injury. *Journal of Neuroscience*, *31*, 13442–13451.

Bookheimer, S. (2002). Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience*, *25*, 151–188.

Bossier, H., Roels, S. P., Seurinck, R., Banaschewski, T., Barker, G. J., Bokde, A. L., Quinlan, E. B., Desrivières, S., Flor, H., Grigis, A., Garavan, H., Gowland, P., Heinz, A., Ittermann, B., Martinot, J. L., Artiges, E., Nees, F., Orfanos, D. P., Poustka, L., … IMAGEN Consortium. (2020). The empirical replicability of task-based fMRI as a function of sample size. *NeuroImage*, *212*, 116601.

Cai, H., Zhu, J., & Yu, Y. (2020). Robust prediction of individual personality from brain functional connectome. *Social Cognitive and Affective Neuroscience*, *15*, 359–369.

Cain, K. (2006). Individual differences in children's memory and reading comprehension: An investigation of semantic and inhibitory deficits. *Memory*, *14*, 553–569.

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., & Petersen, S. E. (2014). Intrinsic and task-evoked network architectures of the human brain. *Neuron*, *83*, 238–251.

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Meriaux, S., Roche, A., Sigman, M., & Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 14240–14245.

Di, X., Gohel, S., Kim, E. H., & Biswal, B. B. (2013). Task vs. rest—Different network configurations between the coactivation and the resting-state brain networks. *Frontiers in Human Neuroscience*, *7*, 493.

Dickens, J. V., Fama, M. E., DeMarco, A. T., Lacey, E. H., Friedman, R. B., & Turkeltaub, P. E. (2019). Localization of phonological and semantic contributions to reading. *Journal of Neuroscience*, *39*, 5361–5368.

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences*, *20*, 425–443.

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology*, *5*, 335.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 16428–16433.

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*, 120–126.

Feng, G., Chen, H. C., Zhu, Z., He, Y., & Wang, S. (2015). Dynamic brain architectures in local brain activity and functional network efficiency associate with efficient reading in bilinguals. *NeuroImage*, *119*, 103–118.

Feng, G., Chen, Q., Zhu, Z., & Wang, S. (2016). Separate brain circuits support integrative and semantic priming in the human language system. *Cerebral Cortex*, *26*, 3169–3182.

Feng, G., Ingvalson, E. M., Grieco-Calub, T. M., Roberts, M. Y., Ryan, M. E., Birmingham, P., Burrowes, D., Young, N. M., & Wong, P. C. M. (2018). Neural preservation underlies speech improvement from auditory deprivation in young cochlear implant recipients. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, E1022–E1031.

Feng, G., Li, Y., Hsu, S. M., Wong, P. C. M., Chou, T. L., & Chandrasekaran, B. (2021). Emerging native-similar neural representations underlie non-native speech category learning success. *Neurobiology of Language*, *2*, 1–82.

Feng, G., Ou, J., Gan, Z., Jia, X., Meng, D., Wang, S., & Wong, P. C. M. (2021). Neural fingerprints underlying individual language learning profiles. *Journal of Neuroscience*, *41*, 7372–7387.

Finn, E. S., Scheinost, D., Finn, D. M., Shen, X., Papademetris, X., & Constable, R. T. (2017). Can brain state be manipulated to emphasize individual differences in functional connectivity? *NeuroImage*, *160*, 140–151.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*, 1664–1671.

Forkel, S. J., Thiebaut de Schotten, M., Dell'Acqua, F., Kalra, L., Murphy, D. G., Williams, S. C., & Catani, M. (2014). Anatomical predictors of aphasia recovery: A tractography study of bilateral perisylvian language networks. *Brain*, *137*, 2027–2039.

Gao, S., Greene, A. S., Constable, R. T., & Scheinost, D. (2019). Combining multiple connectomes improves predictive modeling of phenotypic measures. *NeuroImage*, *201*, 116038.

Gavrilescu, M., Shaw, M. E., Stuart, G. W., Eckersley, P., Svalbe, I. D., & Egan, G. F. (2002). Simulation of the effects of global normalization procedures in functional MRI. *NeuroImage*, *17*, 532–542.

Geranmayeh, F., Brownsett, S. L., Leech, R., Beckmann, C. F., Woodhead, Z., & Wise, R. J. (2012). The contribution of the inferior parietal cortex to spoken language production. *Brain and Language*, *121*, 47–57.

Geranmayeh, F., Chau, T. W., Wise, R. J. S., Leech, R., & Hampshire, A. (2017). Domain-general subregions of the medial prefrontal cortex contribute to recovery of language after stroke. *Brain*, *140*, 1947–1958.

Geranmayeh, F., Wise, R. J. S., Mehta, A., & Leech, R. (2014). Overlapping networks engaged during spoken language production and its cognitive control. *Journal of Neuroscience*, *34*, 8728–8740.

Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH toolbox for assessment of neurological and behavioral function. *Neurology*, *80*, S2–S6.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D., Jenkinson, M., & WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, *80*, 105–124.

Godwin, C. A., Hunter, M. A., Bezdek, M. A., Lieberman, G., Elkin-Frankston, S., Romero, V. L., Witkiewitz, K., Clark, V. P., & Schumacher, E. H. (2017). Functional connectivity within and between intrinsic brain networks correlates with trait mind wandering. *Neuropsychologia*, *103*, 140–153.

Greene, A. S., Gao, S., Scheinost, D., & Constable, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications*, *9*, 1–13.

Griffis, J. C., Nenert, R., Allendorfer, J. B., Vannest, J., Holland, S., Dietz, A., & Szaflarski, J. P. (2017). The canonical semantic network supports residual language function in chronic post-stroke aphasia. *Human Brain Mapping*, *38*, 1636–1658.

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*, 416–423.

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford University Press.

He, N., Rolls, E. T., Zhao, W., & Guo, S. (2020). Predicting human inhibitory control from brain structural MRI. *Brain Imaging and Behavior*, *14*, 2148–2158.

Jackson, R. L., Hoffman, P., Pobric, G., & Lambon Ralph, M. A. (2016). The semantic network at work and rest: Differential connectivity of anterior temporal lobe subregions. *The Journal of Neuroscience*, *36*, 1490–1501.

Jiang, R., Calhoun, V. D., Fan, L., Zuo, N., Jung, R., Qi, S., Lin, D., Li, J., Zhuo, C., Song, M., Fu, Z., Jiang, T., & Sui, J. (2020). Gender differences in connectome-based predictions of individualized intelligence quotient and sub-domain scores. *Cerebral Cortex*, *30*, 888–900.

Jiang, T., He, Y., Zang, Y., & Weng, X. (2004). Modulation of functional connectivity during the resting state and the motor task. *Human Brain Mapping*, *22*, 63–71.

Jo, H. J., Gotts, S. J., Reynolds, R. C., Bandettini, P. A., Martin, A., Cox, R. W., & Saad, Z. S. (2013). Effective preprocessing procedures virtually eliminate distance-dependent motion artifacts in resting state FMRI. *Journal of Applied Mathematics*, *2013*.

Jo, H. J., Saad, Z. S., Simmons, W. K., Milbury, L. A., & Cox, R. W. (2010). Mapping sources of correlation in resting state FMRI, with artifact detection and removal. *NeuroImage*, *52*, 571–582.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Allyn & Bacon.

Koike, T., Kan, S., Misaki, M., & Miyauchi, S. (2011). Connectivity pattern changes in default-mode network with deep non-REM and REM sleep. *Neuroscience Research*, *69*, 322–330.

Krieger-Redwood, K., Jefferies, E., Karapanagiotidis, T., Seymour, R., Nunes, A., Ang, J. W. A., Majernikova, V., Mollo, G., & Smallwood, J. (2016). Down but not out in posterior cingulate cortex: Deactivation yet functional coupling with prefrontal cortex during demanding semantic cognition. *NeuroImage*, *141*, 366–377.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS One*, *9*, e105825.

Lewellen, M. J., Goldinger, S. D., Pisoni, D. B., & Greene, B. G. (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General*, *122*, 316–330.

Lo, A., Chernoff, H., Zheng, T., & Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, *112*, 13892–13897.

López-Barroso, D., Catani, M., Ripollés, P., Dell'Acqua, F., Rodríguez-Fornells, A., & de Diego-Balaguer, R. (2013). Word learning is mediated by the left arcuate fasciculus. *Proceedings of the National Academy of Sciences*, *110*, 13168–13173.

McGlinchey-Berroth, R., Milberg, W. P., Verfaellie, M., Alexander, M., & Kilduff, P. T. (1993). Semantic processing in the neglected visual field: Evidence from a lexical decision task. *Cognitive Neuropsychology*, *10*, 79–108.

Meinzer, M., Harnish, S., Conway, T., & Crosson, B. (2011). Recent developments in functional and structural imaging of aphasia recovery after stroke. *Aphasiology*, *25*, 271–290.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Mollo, G., Karapanagiotidis, T., Bernhardt, B. C., Murphy, C. E., Smallwood, J., & Jefferies, E. (2016). An individual differences analysis of the neurocognitive architecture of the semantic system at rest. *Brain and Cognition*, *109*, 112–123.

Nation, K., & Snowling, M. J. (1998). Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language*, *39*, 85–101.

Ojemann, G. A. (1991). Cortical organization of language. *Journal of Neuroscience*, *11*, 2281–2287.

Papoutsi, M., Stamatakis, E. A., Griffiths, J., Marslen-Wilson, W. D., & Tyler, L. K. (2011). Is left fronto-temporal connectivity essential for syntax? Effective connectivity, tractography and performance in left-hemisphere damaged patients. *NeuroImage*, *58*, 656–664.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*, 786–823.

Price, C. J., & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, *15*, 246–253.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*, 127–157.

Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R. T., & Chun, M. M. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. *Nature Neuroscience*, *19*, 165–171.

Rosenberg, M. D., Noonan, S. K., DeGutis, J., & Esterman, M. (2013). Sustaining visual attention in the face of distraction: A novel gradual-onset continuous performance task. *Attention, Perception, & Psychophysics*, *75*, 426–439.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*, 1–36.

Saad, Z. S., Reynolds, R. C., Jo, H. J., Gotts, S. J., Chen, G., Martin, A., & Cox, R. W. (2013). Correcting brain-wide correlation differences in resting-state FMRI. *Brain Connectivity*, *3*, 339–352.

Satterthwaite, T. D., Wolf, D. H., Roalf, D. R., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E. D., Elliott, M. A., Smith, A., Hakonarson, H., Verma, R., Davatzikos, C., Gur, R. E., & Gur, R. C. (2015). Linked sex differences in cognition and functional connectivity in youth. *Cerebral Cortex*, *25*, 2383–2394.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M. S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., & Weiller, C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 18035–18040.

Saur, D., Lange, R., Baumgaertner, A., Schraknepper, V., Willmes, K., Rijntjes, M., & Weiller, C. (2006). Dynamics of language reorganization after stroke. *Brain*, *129*, 1371–1384.

Scheinost, D., Finn, E. S., Tokoglu, F., Shen, X., Papademetris, X., Hampson, M., & Constable, R. T. (2015). Sex differences in normal age trajectories of functional brain networks. *Human Brain Mapping*, *36*, 1524–1535.

Schilling, H. E., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, *26*, 1270–1281.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612.

Shen, X., Finn, E. S., Scheinost, D., Rosenberg, M. D., Chun, M. M., Papademetris, X., & Constable, R. T. (2017). Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nature Protocols*, *12*, 506–518.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

Smallwood, J., Brown, K., Baird, B., & Schooler, J. W. (2012). Cooperation between the default mode network and the frontal–parietal network in the production of an internal train of thought. *Brain Research*, *1428*, 60–70.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, *106*, 13040–13045.

Snijders, T. M., Petersson, K. M., & Hagoort, P. (2010). Effective connectivity of cortical and subcortical regions during unification of sentence structure. *NeuroImage*, *52*, 1633–1644.

Turken, A. U., & Dronkers, N. F. (2011). The neural architecture of the language comprehension network: Converging evidence from lesion and connectivity analyses. *Frontiers in Systems Neuroscience*, *5*, 1.

Tyler, L. K., & Marslen-Wilson, W. (2007). Fronto-temporal brain systems supporting spoken language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*, 1037–1054.

Uğurbil, K., Xu, J., Auerbach, E. J., Moeller, S., Vu, A. T., Duarte-Carvajalino, J. M., Lenglet, C., Wu, X., Schmitter, S., Van de Moortele, P., Strupp, J., Sapiro, G., De Martino, F., Wang, D., Harel, N., Garwood, M., Chen, L., Feinberg, D. A., Smith, S. M., … WU-Minn HCP Consortium. (2013). Pushing spatial and temporal resolution for functional and diffusion MRI in the human connectome project. *NeuroImage*, *80*, 80–104.

Van Essen, D., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., … WU-Minn HCP Consortium. (2012). The human connectome project: A data acquisition perspective. *NeuroImage*, *62*, 2222–2231.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, *80*, 62–79.

Vatansever, D., Bzdok, D., Wang, H. T., Mollo, G., Sormaz, M., Murphy, C., Karapanagiotidis, T., Smallwood, J., & Jefferies, E. (2017). Varieties of semantic cognition revealed through simultaneous decomposition of intrinsic brain connectivity and behaviour. *NeuroImage*, *158*, 1–11.

Vigneau, M., Beaucousin, V., Hervé, P. Y., Duffau, H., Crivello, F., Houdé, O., Mazoyer, B., & Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *NeuroImage*, *30*, 1414–1432.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development*, *65*, 606–621.

Wei, T., Liang, X., He, Y., Zang, Y., Han, Z., Caramazza, A., & Bi, Y. (2012). Predicting conceptual processing capacity from spontaneous neuronal activity of the left middle temporal gyrus. *Journal of Neuroscience*, *32*, 481–489.

Werning, M. E., Hinzen, W. E., & Machery, E. E. (2012). *The Oxford handbook of compositionality*. Oxford University Press.

Whelan, R., & Garavan, H. (2014). When optimism hurts: Inflated predictions in psychiatric neuroimaging. *Biological Psychiatry*, *75*, 746–748.

Wirth, M., Jann, K., Dierks, T., Federspiel, A., Wiest, R., & Horn, H. (2011). Semantic memory involvement in the default mode network: A functional neuroimaging study using independent component analysis. *NeuroImage*, *54*, 3057–3066.

Xiang, H., Dediu, D., Roberts, L., Oort, E. V., Norris, D. G., & Hagoort, P. (2012). The structural connectivity underpinning language aptitude, working memory, and IQ in the perisylvian language network. *Language Learning*, *62*, 110–130.

Xu, Y., He, Y., & Bi, Y. (2017). A tri-network model of human semantic processing. *Frontiers in Psychology*, *8*, 1538.

Xu, Y., Lin, Q., Han, Z., He, Y., & Bi, Y. (2016). Intrinsic functional network architecture of human semantic processing: Modules and hubs. *NeuroImage*, *132*, 542–555.

Yan, C., & Zang, Y. (2010). DPARSF: A MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. *Frontiers in Systems Neuroscience*, *4*, 13.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology. Human Perception and Performance*, *38*, 53–79.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Meng, D., Wang, S., Wong, P. C. M., & Feng, G. (2022). Generalizable predictive modeling of semantic processing ability from functional brain connectivity. *Human Brain Mapping*, *43*(14), 4274–4292. https://doi.org/10.1002/hbm.25953