Rongtao Jiang, Choong-Wan Woo, Shile Qi, Jing Wu, and Jing Sui

Interpreting Brain Biomarkers

Challenges and solutions in interpreting machine learning-based predictive neuroimaging

Predictive modeling of neuroimaging data (predictive neuroimaging) for evaluating individual differences in various behavioral phenotypes and clinical outcomes is of growing interest. However, the field is experiencing challenges regarding the interpretability of results. Approaches to defining the specific contribution of functional connections, regions, and networks in prediction models are urgently needed, potentially helping to explore underlying mechanisms. In this article, we systematically review methods and applications for interpreting brain signatures derived from predictive neuroimaging, based on a survey of 326 research articles. Strengths, limitations, and suitable conditions for major interpretation strategies are also deliberated. An in-depth discussion of common issues in the existing literature and corresponding recommendations to address these pitfalls are provided. We highly recommend exhaustive validation of the reliability and interpretability of biomarkers across multiple data sets and contexts, which could translate technical advances in neuroimaging into concrete improvements in precision medicine.

Introduction

The past few decades have witnessed significant improvements toward a cumulative understanding of neural mechanisms underlying high-order cognitive functioning [1] by investigating how these constructs map to the brain [2] and are impaired in complex brain disorders. These advances have led to compelling insights into human brain function. Specifically, the research paradigm shift from group-level inference to individual-level prediction is exceedingly impressive, with analytical tools transferring from mass-univariate correlation to multivariate data mining in parallel [3].

Tremendous effort has been devoted to forecasting individual differences on a continuum for both health and disease by using regression-based predictive modeling approaches (hereafter referred to as *predictive neuroimaging*) in an extensive battery of behavioral phenotypes [4], [5] and clinical outcomes [6], [7]. Nevertheless, the field is experiencing immense challenges in translating neuroimaging findings

Digital Object Identifier 10.1109/MSP.2022.3155951
Date of current version: 28 June 2022

into concrete improvements in real-world settings [8]. One of the key factors that may lead to translational failure is the low or fair interpretability of prediction models, where interpretability means identifying the unique contribution of individual brain features to the models decoding predictions, thereby attempting to identify the underlying neurosubstrates of the decoded target variable [9]. Although interpretability has attracted substantial attention from other research fields, it is often an overlooked issue in predictive neuroimaging, compared to classification and diagnosis [9]–[11] in many research hotspots based on deep learning models [12]–[16]. Few previous studies have provided a systematical review summarizing the strategies and recommendations for interpreting regression-based predictive neuroimaging markers. Therefore, we provide a detailed review of approaches and applications for interpreting brain signatures and, more importantly, offer guidelines on how to use them in predictive neuroimaging.

Approaches to defining the specific contribution of functional connections, regions, and networks in prediction models are urgently needed.

Our primary focus is connectome-based prediction, due to its ability to leverage functionally coherent but spatially distributed whole-brain patterns [17] and to yield more reproducible biomarkers [18]. We first outline multiple essential aspects that distinguish predictive neuroimaging from traditional brain mapping studies. Then, based on 326 research studies, we systematically summarize methodological solutions to interrogating predictor contribution, discuss the strengths and limitations for each of them, and provide recommendations and cautions against scenarios that may potentially result in bias and misleading outcomes when interpreting brain findings. Moreover, in an experimental analysis, we compare these interpretation approaches by applying each representative method to the same data. Finally, some encouraging and challenging directions are presented, which show promise to make the black box of this active field transparent.

Common issues in current neuroimaging study

Group mean versus individual differences

For more than two decades, neuroimaging research has predominately focused on revealing group differences. However, focusing on group effects may ignore the rich information that makes individuals unique and obscure true neural signals. Specifically, voxels showing large variations across individuals do not necessarily correspond well to those showing large mean activation [19] [see Figure 1(a) and Supplementary File S1, available at https://doi.org/10.1109/MSP.2022.3155951), implying that brain regions with weak average activation may also carry valuable individualized information. In contrast, predictive neuroimaging emphasizes both intra- and intersubject variabilities. For example, functional connectivity can serve as a unique and reliable fingerprint distinguishing one person from others [18] and is capable of predicting individual cognitive abilities and specific symptoms [5], [20]. Convergent evidence suggests that intersubject variability in multimodal brain measurements shapes the substantial variance in human behavior [3], [21], [22].

Inference versus prediction

Conventional brain mapping investigations usually aim at making inferences about which brain regions are involved in a manipulated mental process by assessing the probability of $P(\text{brain}|\text{behavior})$ [20]. In this framework, behavioral outcomes are independent variables, and neural measures are dependent variables [Figure 1(b)]. Instead, predictive neuroimaging is focused on evaluating how well behavioral outcomes can be predicted from measured brain features, i.e., $P(\text{behavior}|\text{brain})$. Traditional analyses are often evaluated based on the "goodness of fit" to an entire data set, which increases the likelihood of overfitting [23] [Figure 1(c)]. Moreover, they have a heavy reliance on in-sample population inference, leaving the generalizability of an established relationship to out-of-sample data that are largely unknown. In contrast, predictive neuroimaging employs cross validation to mitigate overfitting and increases the likelihood that an established relationship will hold in independent data, offering more translational implications [24].

Furthermore, the emphasis of correlational study is on examining whether an association reaches significance beyond the chance level and whether the direction of effect matches existing evidence [25]. However, a statistically significant in-sample correlation is descriptive and may be insufficient to guarantee robust and useful generalization [26]. When the sample size is small, the correlational results are sensitive to outliers, whereas working with very large data sets can also lead to serious problems [1], e.g., generating extremely small p values but with tiny effect sizes [Figure 1(d)]. In comparison, predictive neuroimaging quantitatively predicts the value of a continuously behavioral dimension, which is able to better characterize the full range of target metrics [27]. Trained within a cross-validation framework, models built using predictive neuroimaging can be directly applied to brain features from out-of-sample individuals, enabling a model to generalize to more accurately predict behavioral scores.

Univariate analysis versus multivariate integrated model

Brain mapping studies typically analyze brain–behavior associations across myriad isolated brain features (i.e., voxels and regions) in parallel [28]. Performing massive statistical comparisons can increase false positives. In addition, a correction for multiple tests can lead to false negatives when the feature dimension is much larger than the sample size [20]. Another issue is that univariate analysis focuses on information from circumscribed voxels and brain regions. However, decades of research have shown that there exists an intricate interplay among distinct brain regions, and the generation of mental traits is not constrained to any a priori region but engages multiple interacting systems spanning the whole brain [29]. Consequently, many behavioral constructs cannot be decoded

from isolated brain regions. A predictive neuroimaging approach can tap into rich multimodal information by jointly combining individual features that have selective relationships with a target outcome within an integrated model [30].

Why interpretability is an overlooked ingredient in predictive neuroimaging

Although specific implementations vary across studies, the workflow for predictive neuroimaging analyses generally includes similar steps [24] (see Supplementary File S2, available at https://doi.org/10.1109/MSP.2022.3155951). The interpretability step, however, has received much less attention in predictive neuroimaging. One potential reason is that the neuroimaging community tends to reward higher prediction performance over neurobiologically meaningful interpretation. Specifically, existing studies often incorporate prediction accuracy as the gold standard to evaluate model quality, no matter whether the research objective is to develop novel algorithms or determine the involved neural circuits [9]. Another reason is that many researchers in predictive neuroimaging

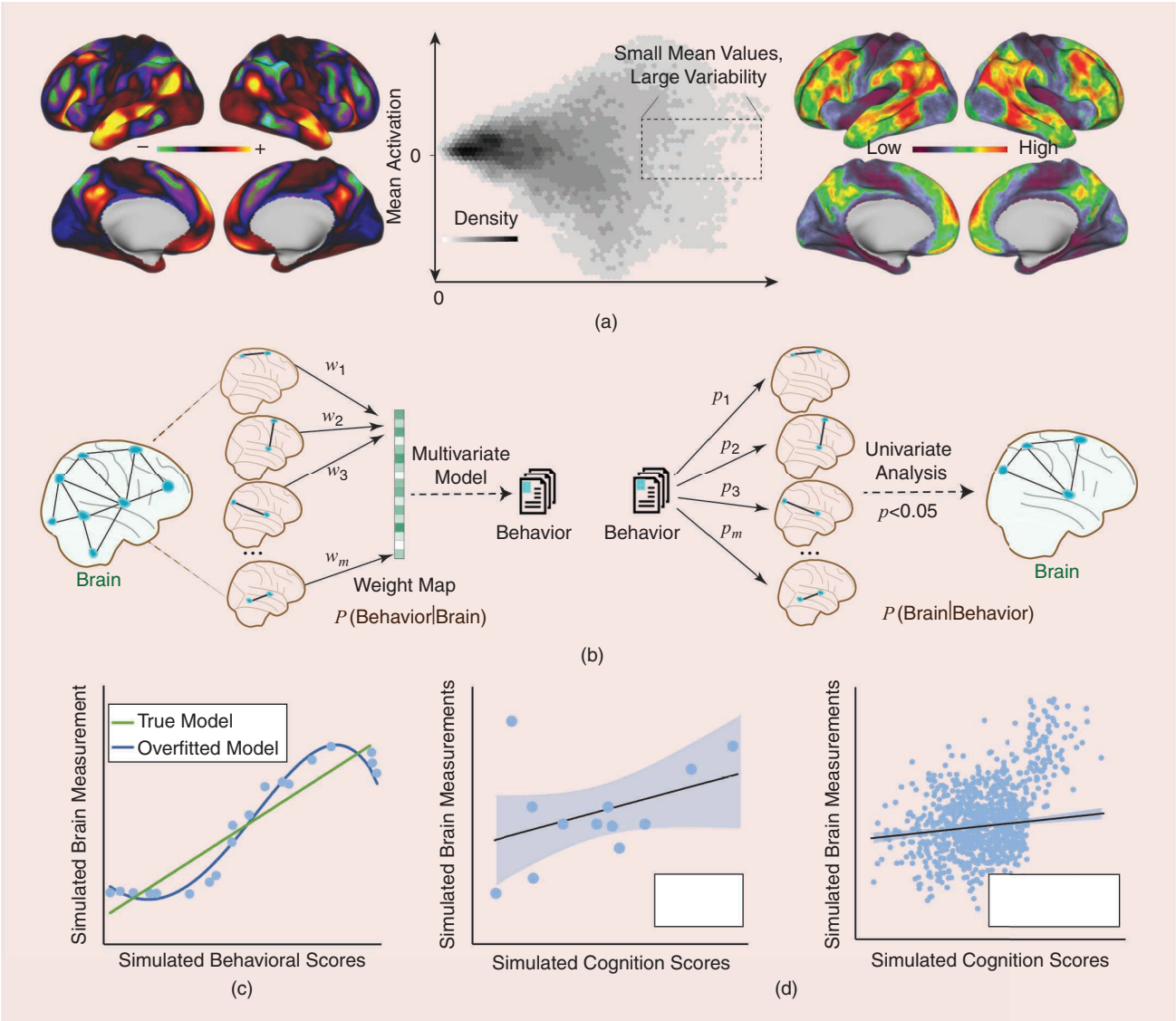

FIGURE 1. The key aspects distinguishing predictive neuroimaging from traditional brain mapping studies. (a) The abundance of information is encoded in individual differences in brain measurements. Voxels showing large variations across individuals do not necessarily correspond well to those showing a large mean activation (for example, voxels in the box have a small mean activation but large variations). The data in this plot are from the publicly available Human Connectome Project data set. Top left: the average whole-brain activation across 922 subjects performing a language task. Top right: the corresponding variability in voxel activation across 922 subjects. Top center: a scatter plot showing the correlation between averaged activation and the variability. (b) Conventional brain mapping focuses on making an inference about which brain regions are involved in a manipulated mental process; predictive neuroimaging makes an inference about how well behavioral outcomes can be forecast from measured brain features. (c) Evaluating models based on the "goodness of fit" to the entire data set risks overfitting. (d) When a sample size is small, correlational results are sensitive to outliers. Correlational analysis in large samples can generate associations with extremely small *p* values but tiny effect sizes. Please note that the plots in (c) and (d) were generated using toy data and thus are provided for illustrative purposes only.

are not domain experts in neuroscience but experts in machine learning, and thus they are more enthusiastic about and better at developing effective models than interpreting results.

However, in addition to pursuing higher predictive performance, determining which specific connections, regions, and functional networks contribute to prediction may significantly advance our knowledge of how the brain implements cognition and, more importantly, facilitate the translation of neuroimaging findings into clinical practice [20], [31]. Moreover, machine learning methods tend to be treated as a black box, which results in focusing on the highest possible predictive performance rather than mechanism understanding [2]. This may lead to the current dilemma of researchers treating interpretation as a secondary goal, e.g., explaining feature importance in their own way and attempting to link with neurobiological significance in a relatively shallow manner without taking full advantage of interpretable models. In this regard, the arbitrary interpretation of models may make it hard to reveal the neural underpinnings of behavioral traits [9].

Although interpretability has attracted substantial attention from other research fields, it is often an overlooked issue in predictive neuroimaging.

Approaches to build interpretable models in predictive neuroimaging

As outlined in Table 1, we systematically describe the three most popular interpretation strategies in the context of regression-based predictive neuroimaging by reviewing 326 relevant articles published since 2010, via keyword searches on Google Scholar and PubMed [5]. The search strategy is provided in Supplementary File S1. Key points, such as the interpretation method, imaging modality, sample size, prediction algorithm, and validation strategy, are listed in Supplementary Table S1 and Supplementary Figure S1. Moreover, to better demonstrate the utility of the interpretation strategies, we construct predictive models for working memory performance based on Human Connectome Project (HCP) activation maps and extract the most-predictive features using each of the three interpretation strategies (see Figure 2; details can be found in Supplementary File S3, available at https://doi.org/10.1109/MSP.2022.3155951).

Beta weights-based quantification metrics

The simplest way to determine feature importance from a regression model is to extract the beta weight from each standardized predictor. This mathematically assigns the expected credit that each predictor receives in forecasting the outcome variable given a one-unit change in it while holding the other independent variables constant [32]. Consequently, it is reasonable to assume that predictors with larger beta weights have greater contributions. A crucial consideration in using such a quantification method is that prediction models are frequently placed within a cross-validation framework, which means that identified predictive features and their beta weights may vary across folds. To cope with this problem, researchers favor the practice of computing an overall mean contribution for each predictor [33] or using the full data to train the final model and extracting the beta weights. For example, in predicting brain maturity and executive functions, Cui et al. [34] applied a twofold cross validation with 100 repetitions and summarized the feature contribution by averaging the beta weights from all 200 prediction models [Figure 3(a)]. In our example data, the mean weight map was highly similar to that from each cross-validation loop but with a relatively low variability due to the effect of averaging [Figure 2(a)].

Although this approach is widely used, overreliance on beta weights can suffer from serious limitations. On the one hand, equating large weights with greater importance may fail for nonlinear models [10]. On the other hand, this quantification

Table 1. A summary of interpretability approaches in predictive neuroimaging.

Interpretability Approach	Description	Cautions and Recommendations	Reference
Beta weight-based metrics	Using regression coefficients as representative of individual feature importance	• Input features should be scaled in a standardized manner. • Techniques should be adopted to alleviate multicollinearity among neuroimaging features (e.g., relative importance analysis). • Averaged beta weight should be used across multiple repetitions of cross validation to enhance interpretability. • Report the reliability of beta weights across different cross-validation folds to provide an overall measure of stability.	[34], [39]–[42], [45], [69]
Stability-based metrics	Determining feature contribution by counting the number of occurrences over multiple different prediction models built on cross validation or resampling data, regardless of the magnitude of regression coefficients	• An additional thresholding technique is required to retain the most-predictive features when there is a vast number of candidates. • It is recommended to demonstrate consensus features for interpretation; they are defined as those with an occurrence rate of 100%. • Combining beta weights- and stability-based quantification metrics can yield better interpretability.	[4], [29], [50], [52], [63], [70]
Prediction performance-based metrics	Evaluating feature importance by examining whether certain brain regions individually carry more predictive information than others (specificity analysis) or to what extent excluding certain features degrades the overall performance (virtual lesion analysis)	• Before concluding the unique contribution of any single network, it is necessary to test against null models to examine the possibility of whether network size influences prediction accuracy more than network identity. • Specificity analysis or virtual lesion analysis relates to increased computational loads. • The unique predictive power of an individual network may have high reliance on the parcellation scheme of the brain space.	[4], [18], [33], [51], [57]–[59], [61]

strategy is applicable only in situations where there exists no strong correlation among predictors. However, neuroimaging features can be highly intercorrelated, resulting in a statistical phenomenon called *multicollinearity* [35]. In this case, beta weights are heavily influenced by covariance among predictors, and the squared coefficients do not naturally decompose overall prediction R^2 (variance explained). Importantly, in the context of multivariate classifiers, Haufe et al. implied

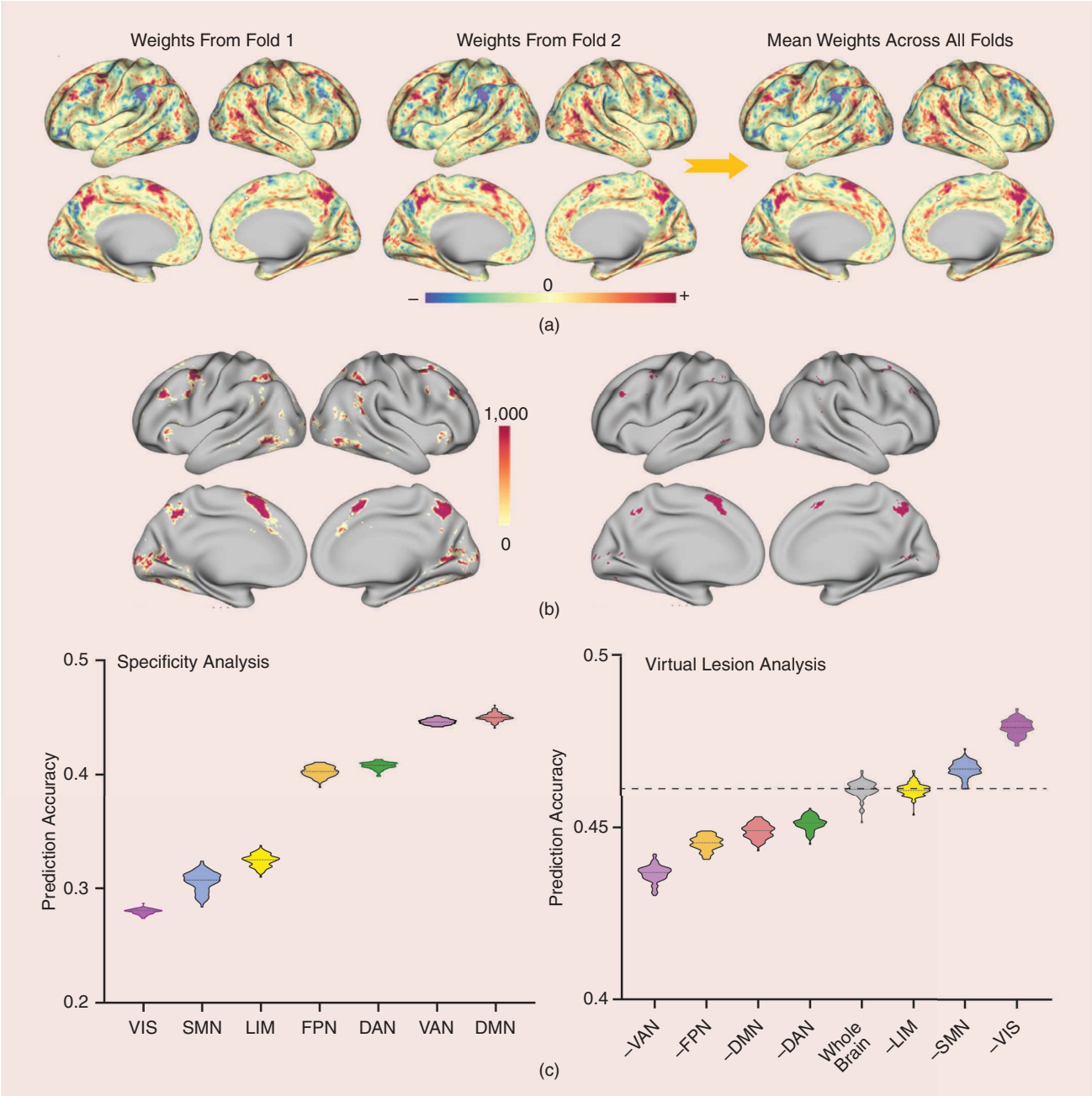

FIGURE 2. A comparison of interpretation approaches by applying each representative method to the same data. Leveraging whole-brain activation maps from the two-back condition of the working memory task in the HCP, we constructed predictive models for working memory performance and extracted the most-predictive features on the basis of each of the three interpretation strategies. (a) Beta weights (left) for whole-brain features derived from two example cross-validation loops and the mean beta weights (right) averaged across all models. (b) Across 100 rounds of 10-fold cross validation, a total of 6,063 distinct voxels appeared in all predictive models, while the consensus features included 614 voxels, representing approximately 1% of the brain's total. (c) In specificity analysis, ventral attention and default mode networks emerged as the top two most-predictive networks. In virtual lesion analysis, ventral attention and frontoparietal networks degraded the prediction performance the most upon removal, reflecting their great contribution in prediction. Although these interpretation strategies vary in multiple aspects, predictive biomarkers derived from different quantification approaches are more likely to be consistent with one another when the model is reliable enough. More details can be found in Supplementary File S3. DAN: dorsal attention network; DMN: default mode network; FPN: frontoparietal network; LIM: limbic network; SMN: somatomotor network; VAN: ventral attention network; VIS: visual network.

that interpreting model weights (filters) as activation patterns (truth) can lead to erroneous conclusions unless the individual features are uncorrelated [10]. This study proposed a framework for interpreting linear multivariate models by considering covariance structure and demonstrated its effectiveness in both simulation and real functional magnetic resonance imaging (fMRI) and electroencephalography data.

Multicollinearity may yield unstable regression coefficients, and sometimes even a minor change in covariance structure could dramatically alter the beta weights, complicating feature interpretation [36]. Fortunately, four different approaches can be adopted to alleviate this problem. The first is to employ prediction methods that have good resilience to multicollinearity data, such as the least absolute shrinkage and selection operator (LASSO) and ridge regression. These methods still work well when data have many more features than instances. Specifically, ridge regression deals with multicollinearity by assigning similar coefficients to correlated features but may come at the cost of increased model complexity [37]. The LASSO arbitrarily retains a representative predictor from a group of correlated ones and drops the others to avoid multicollinearity [38]. A prominent concern is that this can lead to the exclusion of some important features. Practically, researchers often incorporate both model-selected features and their tightly correlated ones to make interpretations [39], [40].

Focusing on group effects may ignore the rich information that makes individuals unique and obscure true neural signals.

The second approach is to project neuroimaging features into a small set of separable (i.e., orthogonal or independent) latent components by using dimensionality reduction techniques and then feed the components into a prediction model. Since the predictors are uncorrelated, variable importance can be determined by directly inspecting the derived beta weights. A classic example is the recently proposed Brain Basis Set prediction model [41], which transforms high-dimensional functional connections into a small suite of latent components by using principal component analysis (PCA) and then fits a multiple linear regression model to predict neurocognitive scores through expression scores of these components. Accordingly, feature importance is determined by multiplying feature component maps with their respective beta coefficients from the prediction model. This method is strongly recommended when there is a need to visualize the contribution of all features that are highly correlated. A potential problem is that the optimal number of latent components needs to be determined by additional experiments.

The third approach does not use the original beta weight values but performs permutation tests to assess statistical significance [42], [43]. For example, in a recent fMRI study using partial least-squares regression to predict reading comprehension abilities [42], after extracting the regression coefficients β, a permutation test was employed to create a null weight distribution β_{perm} for each feature. The most strongly predictive features were then determined as those whose β value significantly differed from the empirical distribution β_{perm} obtained from 10,000 permutations [Figure 3(a)]. A prominent strength of this approach is that it provides the statistical significance of individual features. However, it is computationally intensive since a massive number of permutation tests needs to be performed.

The fourth approach, named *relative importance analysis*, is capable of decomposing the overall R^2 into nonnegative contributions [44]. This posthoc technique has the advantage of not changing the feature selection and model building processes; rather, it applies mathematical techniques to control for multicollinearity. This means that the quantification of feature importance is independent of model construction, and consequently we can separately achieve model interpretation and model building. Importantly, this approach suits multicollinear data. In a recent study based on United Kingdom biobank data, [45], the correlation-adjusted marginal correlation (CAR) score was adopted to assist interpretation, which used Mahalanobis decorrelation to adjust the multicollinearity among explanatory variables [46]. Detailed implementations of such relative importance metrics can be found in a series of R packages, including "relaimpo," "hier.part," and "care" [36], [46], [47].

Stability-based quantification metrics

Stability-based quantification metrics count the number of occurrences of a given predictor across multiple different prediction models built on cross validation or resampling data, regardless of the magnitude of the regression coefficients. A significant strength of this approach is reduced sensitivity to multicollinearity and applicability even to nonlinear models. For example, Liu et al. predicted fluid intelligence and cognitive flexibility scores based on functional connectivity for a sample of 105 healthy participants within a leave-one-out cross-validation framework [48]. To quantify feature contribution, this study counted the number of times each functional connectivity was selected across all 105 folds. Another study leveraged a bootstrapping strategy to build a total of 100 predictive models based on resampled data, and features exceeding a frequency percentage of 70% were determined to be the most-predictive ones [49]. Stability-based quantification metrics are often adopted by predictive frameworks that incorporate a built-in (e.g., the LASSO) or separate feature selection step to achieve dimension reduction and are not applicable to frameworks that include all available features in the final models.

One limitation is that this may lead to the inclusion of a large number of candidate features, thereby requiring additional thresholding to select the most-predictive ones. A conservative solution is to use only consensus features for interpretation. The features are defined as those with an identification rate of 100%; i.e., they are shared across every iteration of cross validation [50], [51] [Figure 3(b)]. Features in the consensus set are considered to have equal contributions to prediction and should be interpreted as a whole. This compact set of features has the highest stability and reduced susceptibility to potential confounds [27]. Establishing models using this parsimonious set of features has been demonstrated to afford robust

generalizability across multiple independent data sets [4], [29], [52]. Therefore, this interpretation approach is best used when there is a need to characterize a complex behavioral trait with a condensed set of brain signatures and, more importantly, use the signatures to establish reliable and generalizable predictive models. This is especially helpful for neuroimaging data that include only a small number of subjects since quantifying feature importance via consensus features is attached to high reliability and generalizability.

The neuroimaging community tends to reward higher prediction performance over neurobiologically meaningful interpretation.

However, when prediction models significantly differ across distinct cross-validation loops, there may be little overlap among identified features, resulting in few and even no consensus ones [53]. Another limitation is that reducing any complex behavioral trait to a handful of brain features risks oversimplification [4], [54] and thus may miss information that is crucial for understanding underlying mechanisms, rendering biological interpretability difficult. As shown in Figure 2(b), a total of 6,063 distinct voxels appeared in all predictive models with an identification occurrence ranging from one to 1,000. The consensus features included only 614 voxels, representing approximately 1% of the brain's total. These 614 features were assumed to have the greatest contribution to prediction because they were repeatedly identified by all 1,000 distinct

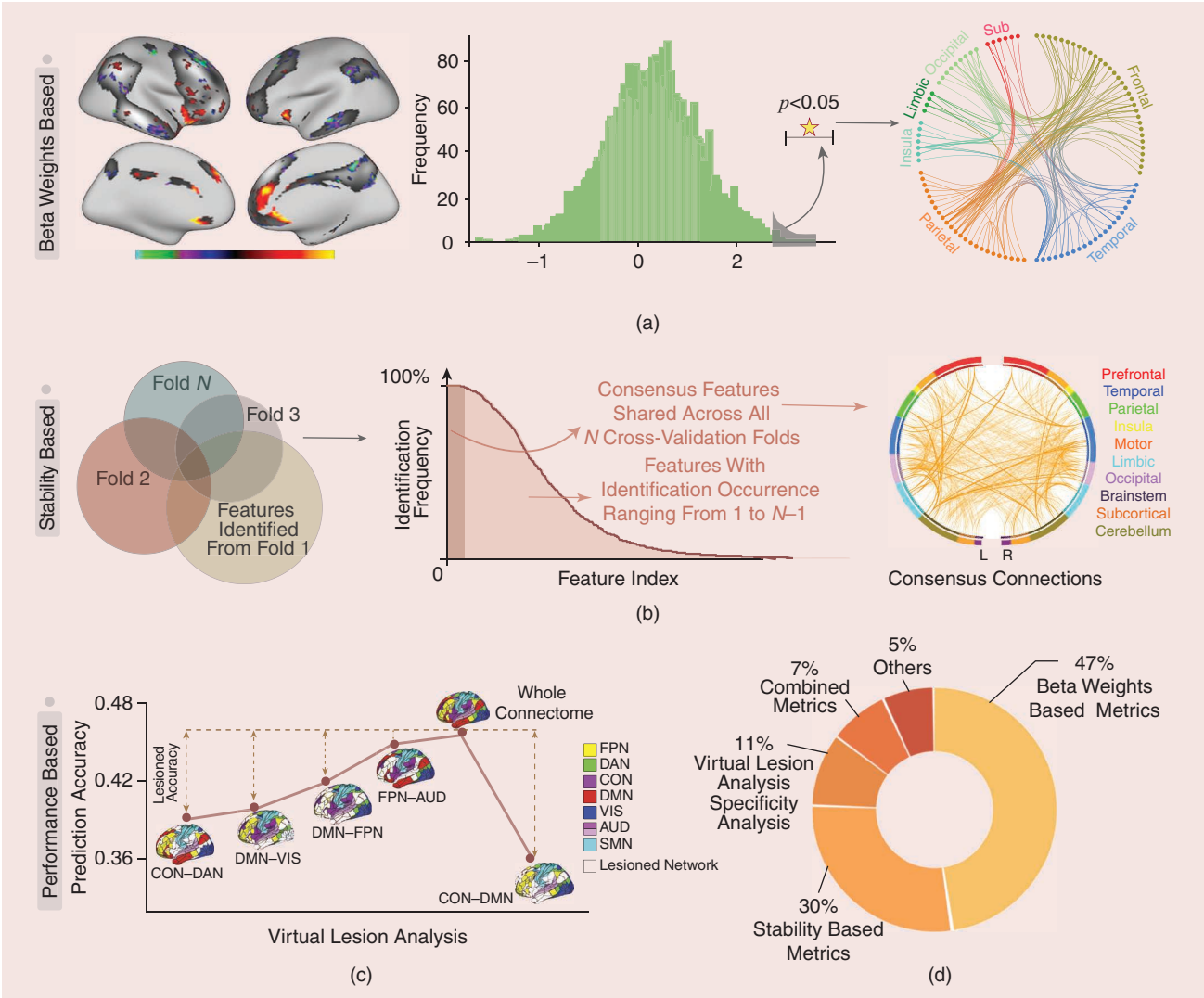

FIGURE 3. The approaches to build interpretable models in predictive neuroimaging. (a) Cui et al. (left) summarized feature contribution by averaging beta weights from all repeated cross-validation models. Jiang et al. (right) adopted a permutation test to assess the statistical significance of the beta weight for each feature. The most strongly predictive features were defined as those whose true β value significantly differed from the permutation-derived β_{perm}. (b) In predicting attention performance, Rosenberg et al. counted the number of occurrences when each feature was selected across N cross-validation loops and finally identified 757 consensus features with an occurrence rate of 100%. (c) Using virtual lesion analysis, Dubois et al. iteratively removed features from between any two functional networks from the whole-brain connectome and reran the predictive framework to isolate each network pair's contribution to prediction. (d) The proportion of each interpretation strategy among all reviewed papers. (Adapted and modified from [4], [34], [42], and [59].) CON: cingulo-opercular network; AUD: auditory network.

models. However, such a complex and multifactorial phenotype (working memory) is unlikely to be driven by this small set of voxels; therefore, interpreting by using only consensus features may risk oversimplification.

Another group of studies builds an overall predictive model on all subjects, with the algorithm parameters determined through cross validation [55], [56], and then extracts all features from the fitted model for further interpretation and visualization. Despite reduced computational cost, this approach is inherently explanatory and should be used only for preliminary interpretation, due to an increased likelihood of overfitting [56]. However, when the validity of these identified brain signatures is verified by multiple external data sets, interpreting predictive models through this method is highly recommended.

The selection of an appropriate approach can be a thorny issue since there is no optimal solution that applies to all conditions.

Moreover, beta weights- and stability-based quantification metrics can be combined to obtain better interpretation, especially when the employed predictive framework retains all features in the final model but there is a need to demonstrate only a small number of the most-predictive ones. In particular, to extract the most-predictive functional connections of brain maturity, Dosenbach et al. selected a constant number of the 200 highest-ranked features from each cross-validation fold according to their magnitude of beta weights and then identified 156 shared connections as consensus features [50].

Prediction performance-based quantification metrics

Prediction performance-based quantification metrics place less emphasis on constructed models and participated predictors. Instead, they evaluate feature importance by examining whether certain brain regions and networks individually carry more predictive information than others [33], [57], [58] and to what extent excluding certain features degrades overall performance [4], [24], [51], [59].

For connectome-based predictive neuroimaging, virtual lesion analysis and specificity analysis are two representative methods that are developed on the basis of prediction performance. Specifically, virtual lesion analysis works by iteratively removing connections in a certain network from the whole-brain connectome to isolate its contribution to prediction [60]. Putatively, the magnitude of change in prediction accuracy upon the removal of a specific network reflects its unique contribution. One study predicted general intelligence by using the whole-brain connectome, with an accuracy or $r = 0.457$ [59], and then employed virtual lesion analysis to elucidate the predictive power of connections from between any two networks. The results demonstrated that removing the connections between cingulo–opercular and default model networks yielded the lowest prediction accuracy ($r = 0.37$), indicating the great contribution of these two networks in intelligence prediction [Figure 3(c)].

Specificity analysis restricts model building to brain connectivity from one functional network and attributes greater contribution power to networks that achieve higher prediction accuracy [24]. For example, in predicting symptom severity for patients with obsessive–compulsive disorder, Reggente et al. divided whole-brain regions into eight functional networks and built eight prediction models using connections from each network [58]. The results showed that only the default model and visual networks achieved significant accuracies, while no other network reached statistical significance.

Compared with other types of quantification metrics, a prominent strength of prediction performance-based measures is that they provide a straightforward way to directly pinpoint brain regions with the highest contribution and require no additional technique to summarize those low-level features (e.g., edges) to high-level representations (e.g., networks) for better interpretation and visualization. Although accounting for only 11% of all reviewed papers, this interpretation strategy is usually combined with other approaches, serving as a validation and complementary analysis to confirm identified brain signatures [4]. Moreover, it can also be used in nonlinear models since feature importance does not rely on model-learned feature weights.

Nevertheless, an insidious problem comes from the fact that the network size may influence the prediction accuracy more than the network identity. As an example, Nielsen et al. grouped whole-brain nodes into 13 functional systems and used within-connections from each of these networks to predict individual brain maturity [61]. The results demonstrated that all networks can predict age; however, prediction accuracies varied as a function of network size. Additional analyses suggested that none of these networks achieved better predictions than models built on a matched number of randomly selected connections. This study highlighted the necessity of testing against null models before concluding the unique contribution of any single network [31], [61].

Another limitation is an increased computational load. For a parcellation scheme of m networks, at least $C_m^2 + m$ different models need to be constructed to examine the predictive power for any between-network and within-network connections in a virtual lesion or specificity analysis. Moreover, this approach requires a priori specification of how whole brain nodes are divided into different functional systems, based on which virtual lesion analysis or specificity analysis can be performed to characterize each system's unique contribution. In this respect, neurobiological insight can be acquired only from the level of predefined functional systems, not facilitating finer-grained representations. For example, in our experimental analysis, we grouped whole-brain voxels into seven canonical networks and could not make any interpretations beyond these functional networks [Figure 2(c)].

Future considerations for building interpretable neuroimaging biomarkers

Recommendations for interpreting predictive neuroimaging results

The preceding interpretation approaches vary in multiple aspects and sometimes may provide different answers to a problem.

In this sense, the selection of an appropriate approach can be a thorny issue since there is no optimal solution that applies to all conditions, and different methods may have their own strengths and weakness. The selection of an interpretation strategy can depend on research aims. Specifically, if we aim to determine which specific functional network contributes more to prediction than others, the prediction performance-based approach may be a good choice; if we would like to determine the contribution of whole-brain features quantitatively, the beta weights-based metrics may be more appropriate, and if we want to derive a compact set of features for further validation, consensus features may be optimal. Nevertheless, what we can do is follow best practices to better implement a selected approach.

For interpreting neuroimaging results from a prediction model, we provide the following recommendations:

1) When extreme multicollinearity exists among predictors, avoid using beta weights to interpret results.
2) Stability-based quantification metrics, particularly consensus features, are preferred when there is a need for constructing robust and generalizable prediction models with a compact site of neuroimaging features.
3) Prediction performance-based quantification metrics are suitable for ascertaining the unique contribution of individual functional networks and brain regions and can be used as a complement to confirm results from other interpretation strategies.
4) Try to perform *k*-fold cross validation with as many partition repetitions as possible to dilute the influence of the random division of data folds, and use the averaged beta weights to increase the stability of feature importance.
5) Report the reliability of beta weights across different cross-validation folds to provide an overall measure of quantification stability.
6) Utilize multiple interpretation techniques to validate and examine the convergence among them, instead of being limited to a single approach (Figure 4).
7) When dealing with high-dimensional neuroimaging features, effective feature extraction techniques are preferred since they result in a small set of more informative representations (e.g., PCA).
8) Establish an appropriate null model for examining whether identified features perform better than chance to unambiguously claim their unique utility [31].
9) Normalize edge counts and weight sums to account for network size when summarizing individual connections to network representations for visualization when using beta weights- and stability-based quantification metrics [17].

Validating the biological plausibility of identified brain signatures

Not until a brain signature is externally validated across different contexts can it become a usable biomarker [62]. However, validating the biological plausibility of brain signatures is exceedingly challenging, given that the underlying substrates for any phenotype that are theoretically agnostic as the "ground truth" about which a specific set of neuroimaging features defines this construct are unknown. Therefore, it is impossible to explicitly

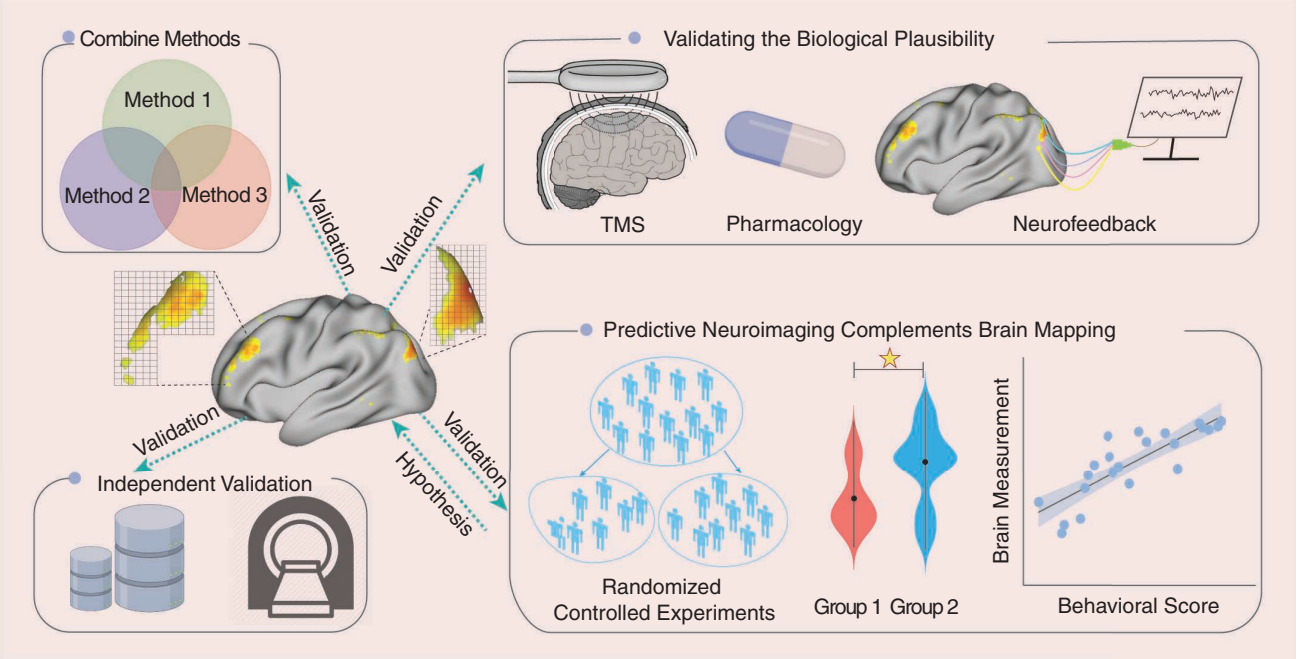

FIGURE 4. The approaches to validate neuroimaging signatures. First, to obtain more interpretable biomarkers, researchers can use many techniques to validate and examine the convergence between them. Second, external heterogeneous data sets can be leveraged to test whether and to what extent models based on identified interpretable neuromarkers can generalize. Moreover, noninvasive techniques, such as real-time neurofeedback and neuropharmacology, can be employed to validate the biological plausibility of identified brain signatures. Furthermore, predictive neuroimaging and brain mapping are not mutually exclusive but complementary in biomarker discovery. TMS: transcranial magnetic stimulation.

define a specific set of brain voxels or connections that can serve as the benchmark to be tested against. In this regard, validation techniques that can determine the validity of model-identified brain features are urgently needed. On the one hand, external heterogeneous data sets can be leveraged to test whether and to what extent models based on identified interpretable neuromarkers can generalize across contexts (scanners, laboratories, populations, and disease characteristics) [7], [20]. On the other hand, real-time noninvasive techniques, such as neurofeedback and neuropharmacology, can be applied to identified brain signatures in clinical trials to validate their intervention effects (Figure 4) [28]. Imaging biomarkers confirmed by these noninvasive validation effects usually suggest more translational implications.

Following best practices to build robust prediction models

Model interpretability relies heavily on the reliability and efficacy of the prediction model itself, which necessitates a protocol for establishing robust and powerful prediction models. Indeed, predictive biomarkers derived from different quantification approaches are more likely to be consistent with one another when the model is reliable enough to be generalizable across different contexts where more confidence can be placed. Accordingly, the predictive features derived from the three interpretation approaches in our experimental analysis demonstrate high overlap among one another, which may be due to the relatively large sample size and adoption of repeated cross-validation strategies (Figure 2). In this regard, the optimal practices that have been established in predictive neuroimaging should always be followed and pursued whenever possible. For example, researchers should carefully control for covariates in model building to ensure that their models are not influenced by confounds. Other feasible practices involve increasing the sample size, reducing model complexity [63], integrating multimodal data [64], extending fMRI scan durations [42], and defining individual-specific functional space [65].

Not until a brain signature is externally validated across different contexts can it become a usable biomarker.

Combining univariate inferences and multivariate predictions

Although predictive neuroimaging and brain mapping differ in multiple aspects when establishing brain–behavior relationships, they are not mutually exclusive but, rather, complementary [1]. We encourage their combined use to gain comprehensive insights into the neurobiological substrates of human cognition and disease pathology. On the one hand, candidate brain biomarkers derived from predictive neuroimaging can serve as prior hypotheses and clinical targets, while well-designed and randomized controlled experiments can be leveraged to confirm their biological plausibility to facilitate interpretability [25]. On the other hand, brain regions surviving rigorous statistical testing can serve as prior knowledge, and machine learning approaches can work with these low-dimensional features to test their predictability and relate their interpretability to prediction performance. The combination of these two approaches can prospectively catalyze biomarker discovery on the path to translational neuroscience.

Beyond neuroimaging

While the current review primarily focuses on neuroimaging applications from connectome-based predictive modeling, the points raised here can be extended to research problems such as decoding task activation maps from functional connectivity [21], diagnosing psychiatric diseases through classification [66], and delineating disease biotypes via clustering [67]. Going beyond the neuroimaging context, these interpretation strategies can be easily adapted to other research fields using machine learning because they generally follow similar workflows and since a majority of the available machine learning methods are not specifically developed for neuroimaging. Therefore, the interpretation approaches are generalizable and transferable across different areas.

Indeed, many of the interpretation methods discussed in this review have been leveraged in other fields. For example, Wei et al. employed a relative importance analysis method (the CAR score) to determine the relative contribution of social, economic, and physical variables affecting domestic energy use in London [68]. We encourage the future introduction and adoption of interpretation approaches from other fields to neuroimaging investigations. Furthermore, the neuroimaging community is witnessing increasing interest in interpreting deep learning models. A detailed discussion of the interpretability of deep learning is beyond the scope of the current review, and we point interested readers to a series of recent work in [12]–[15].

Conclusions

The burgeoning field of predictive neuroimaging is rapidly evolving, aiming at quantitatively predicting phenotypic outcomes on a continuum. This review dug into details about how to interrogate the contribution of brain features in the context of regression-based predictive neuroimaging. Despite a specific focus on neuroimaging applications from connectome-based predictive modeling, the ideas raised here can be extended to studies using other imaging modalities and, more broadly, to research practices such as classification and biotype clustering. Collectively, interpreting neuroimaging results through appropriate approaches can help better unveil the underlying mechanisms of human cognitive ability and disease progress and even facilitate clinical intervention, thereby accelerating the pace of biomarker discovery.

Acknowledgments

We would like to thank Dr. Vince D. Calhoun for long and fruitful collaborations and his insightful comments. This work is supported, in part, by the China Natural Science Foundation (grants 82022035 and 61773380) and the National Institutes of Health (grants 1R01MH117107 and 1R01MH094524). This work involved human subjects or animals in its research. Use of

HCP data for these analyses was deemed exempt from IRB review by the Yale Human Investigation Committee. This article has supplementary downloadable material available at https://doi.org/10.1109/MSP.2022.3155951, provided by the authors. The corresponding authors are Jing Sui and Rongtao Jiang.

Authors

Rongtao Jiang (rongtao.jiang@yale.edu) received his Ph.D. degree in pattern recognition and intelligence systems from the Institute of Automation, Chinese Academy of Sciences, in 2020. He is a postdoctoral research associate in the Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, Connecticut, 06520, USA. His research interests include multimodal brain imaging analysis, individualized prediction, connectome-based predictive modeling, and the application of machine learning in medical image analysis.

Choong-Wan Woo (waniwoo@g.skku.edu) received his dual Ph.D. degree in the Department of Psychology and Neuroscience and the Institute of Cognitive Sciences, University of Colorado Boulder. He is the director of the Computational Cognitive Affective Neuroscience Laboratory, Sungkyunkwan University, Suwon, 16419, South Korea. His research interests include understanding how the human brain represents, processes, and regulates pain and emotions by using machine learning and computational modeling.

Shile Qi (shile.qi@nuaa.edu.cn) received her Ph.D. degree in pattern recognition and intelligence systems from the Institute of Automation, Chinese Academy of Sciences. She is a full professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China. Her research interests include multimodal fusion method development and its application in brain diseases.

Jing Wu (jingwu1209@gmail.com) received her M.D. degree in internal medicine from Capital Medical University in 2019. She is a Ph.D. candidate in the Department of Critical Care Medicine of Liver Disease, Capital Medical University, Beijing, 100069, China. Her research interests include liver diseases and clinical and basic research into acute chronic liver failure.

Jing Sui (jsui@bnu.edu.cn) received her Ph.D. degree from the Beijing Institute of Technology in 2007. She is a full professor at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, 100875, China. Her research interests include multimodal neuroimaging data fusion and its applications in mental illnesses. She is a Senior Member of IEEE.

References

[1] M. D. Rosenberg, B. J. Casey, and A. J. Holmes, "Prediction complements explanation in understanding the developing brain," *Nature Commun.*, vol. 9, no. 1, p. 589, 2018, doi: 10.1038/s41467-018-02887-9.

[2] D. Bzdok and J. P. A. Ioannidis, "Exploration, inference, and prediction in neuroscience and biomedicine," *Trends Neurosci.*, vol. 42, no. 4, pp. 251–262, 2019, doi: 10.1016/j.tins.2019.02.001.

[3] J. Dubois and R. Adolphs, "Building a science of individual differences from fMRI," *Trends Cognitive Sci.*, vol. 20, no. 6, pp. 425–443, 2016, doi: 10.1016/j.tics.2016.03.014.

[4] M. D. Rosenberg, E. S. Finn, D. Scheinost, X. Papademetris, X. Shen, R. T. Constable, and M. M. Chun, "A neuromarker of sustained attention from whole-brain functional connectivity," *Nature Neurosci.*, vol. 19, no. 1, pp. 165–171, 2016, doi: 10.1038/nn.4179.

[5] J. Sui, R. Jiang, J. Bustillo, and V. Calhoun, "Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: Methods and promises," *Biol. Psychiatry*, vol. 88, no. 11, pp. 818–828, 2020, doi: 10.1016/j.biopsych.2020.02.016.

[6] S. Qi *et al.*, "Common and unique multimodal covarying patterns in autism spectrum disorder subtypes," *Mol. Autism*, vol. 11, no. 1, p. 90, 2020, doi: 10.1186/s13229-020-00397-4.

[7] J. J. Lee *et al.*, "A neuroimaging biomarker for sustained experimental and clinical pain," *Nature Med.*, vol. 27, no. 1, pp. 174–182, 2021, doi: 10.1038/s41591-020-1142-7.

[8] R. A. Poldrack, G. Huckins, and G. Varoquaux, "Establishment of best practices for evidence for prediction: A review," *JAMA Psychiatry*, vol. 77, no. 5, pp. 534–540, 2020, doi: 10.1001/jamapsychiatry.2019.3671.

[9] L. Kohoutova, J. Heo, S. Cha, S. Lee, T. Moon, T. D. Wager, and C. W. Woo, "Toward a unified framework for interpreting machine-learning models in neuroimaging," *Nature Protocols*, vol. 15, no. 4, pp. 1399–1435, 2020, doi: 10.1038/s41596-019-0289-5.

[10] S. Haufe, F. Meinecke, K. Gorgen, S. Dähne, J. D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *Neuroimage*, vol. 87, pp. 96–110, Feb. 2014, doi: 10.1016/j.neuroimage.2013.10.067.

[11] N. Kriegeskorte and P. K. Douglas, "Interpreting encoding and decoding models," *Current Opin. Neurobiol.*, vol. 55, pp. 167–179, Apr. 2019, doi: 10.1016/j.conb.2019.04.002.

[12] F. Eitel, M. A. Schulz, M. Seiler, H. Walter, and K. Ritter, "Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research," *Exp. Neurol.*, vol. 339, p. 113,608, May 2021, doi: 10.1016/j.expneurol.2021.113608.

[13] A. W. Thomas, H. R. Heekeren, K.-R. Müller, and W. Samek, "Analyzing neuroimaging data through recurrent deep learning models," *Frontiers Neurosci.*, vol. 13, p. 1321, Dec. 2019, doi: 10.3389/fnins.2019.01321.

[14] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," *Frontiers Aging Neurosci.*, vol. 11, p. 194, Jul. 2019, doi: 10.3389/fnagi.2019.00194.

[15] S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neurosci. Biobehav. Rev.*, vol. 74, pp. 58–75, Mar. 2017, doi: 10.1016/j.neubiorev.2017.01.002.

[16] A. W. Thomas, C. Ré, and R. A. Poldrack, "Challenges for cognitive decoding using deep learning methods," 2021, *arXiv:2108.06896*.

[17] A. S. Greene, S. Gao, D. Scheinost, and R. T. Constable, "Task-induced brain state manipulation improves prediction of individual traits," *Nature Commun.*, vol. 9, no. 1, p. 2807, 2018, doi: 10.1038/s41467-018-04920-3.

[18] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable, "Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity," *Nature Neurosci.*, vol. 18, no. 11, pp. 1664–1671, 2015, doi: 10.1038/nn.4135.

[19] D. Wu, X. Li, and T. Jiang, "Reconstruction of behavior-relevant individual brain activity: An individualized fMRI study," *Sci. China Life Sci.*, vol. 63, no. 3, pp. 410–418, 2020, doi: 10.1007/s11427-019-9556-4.

[20] C. W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager, "Building better biomarkers: Brain models in translational neuroimaging," *Nature Neurosci.*, vol. 20, no. 3, pp. 365–377, 2017, doi: 10.1038/nn.4478.

[21] I. Tavor, O. P. Jones, R. B. Mars, S. M. Smith, T. E. Behrens, and S. Jbabdi, "Task-free MRI predicts individual differences in brain activity during task performance," *Science*, vol. 352, no. 6282, pp. 216–220, 2016, doi: 10.1126/science.aad8127.

[22] N. Luo *et al.*, "Structural brain architectures match intrinsic functional networks and vary across domains: A study from 15 000+ individuals," *Cerebral Cortex*, vol. 30, no. 10, pp. 5460–5470, 2020, doi: 10.1093/cercor/bhaa127.

[23] R. Whelan and H. Garavan, "When optimism hurts: Inflated predictions in psychiatric neuroimaging," *Biol. Psychiatry*, vol. 75, no. 9, pp. 746–748, 2014, doi: 10.1016/j.biopsych.2013.05.014.

[24] S. W. Yip, B. Kiluk, and D. Scheinost, "Toward addiction prediction: An overview of cross-validated predictive modeling findings and considerations for future neuroimaging research," *Biol. Psychiatry Cognitive Neurosci. Neuroimag.*, vol. 5, no. 8, pp. 748–758, 2020, doi: 10.1016/j.bpsc.2019.11.001.

[25] T. Yarkoni and J. Westfall, "Choosing prediction over explanation in psychology: Lessons from machine learning," *Perspectives Psychol. Sci.*, vol. 12, no. 6, pp. 1100–1122, 2017, doi: 10.1177/1745691617693393.

[26] E. Krapohl *et al*., "Multi-polygenic score approach to trait prediction," *Mol. Psychiatry*, vol. 23, no. 5, pp. 1368–1374, 2018, doi: 10.1038/mp.2017.163.

[27] X. Shen, E. S. Finn, D. Scheinost, M. D. Rosenberg, M. M. Chun, X. Papademetris, and R. T. Constable, "Using connectome-based predictive modeling to predict individual behavior from brain connectivity," *Nature Protocols*, vol. 12, no. 3, pp. 506–518, 2017, doi: 10.1038/nprot.2016.178.

[28] P. A. Kragel, L. Koban, L. F. Barrett, and T. D. Wager, "Representation, pattern information, and brain signatures: From neurons to neuroimaging," *Neuron*, vol. 99, no. 2, pp. 257–273, 2018, doi: 10.1016/j.neuron.2018.06.009.

[29] M. D. Rosenberg *et al*., "Functional connectivity predicts changes in attention observed across minutes, days, and months," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 7, pp. 3797–3807, 2020, doi: 10.1073/pnas.1912226117.

[30] R. Jiang *et al*., "SMRI biomarkers predict electroconvulsive treatment outcomes: Accuracy with independent data sets," *Neuropsychopharmacology*, vol. 43, no. 5, pp. 1078–1087, 2018, doi: 10.1038/npp.2017.165.

[31] A. N. Nielsen, D. M. Barch, S. E. Petersen, B. L. Schlaggar, and D. J. Greene, "Machine learning with neuroimaging: Evaluating its applications in psychiatry," *Biol. Psychiatry Cognitive Neurosci. Neuroimag.*, vol. 5, no. 8, pp. 791–798, 2020, doi: 10.1016/j.bpsc.2019.11.007.

[32] K. F. Nimon and F. L. Oswald, "Understanding the results of multiple linear regression," *Org. Res. Methods*, vol. 16, no. 4, pp. 650–674, 2013, doi: 10.1177/1094428113493929.

[33] J. S. Siegel *et al*., "Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 30, pp. E4367–E4376, 2016, doi: 10.1073/pnas.1521083113.

[34] Z. Cui *et al*., "Individual variation in functional topography of association networks in youth," *Neuron*, vol. 106, no. 2, pp. 340–353.e8, 2020, doi: 10.1016/j.neuron.2020.01.029.

[35] A. Kraha, H. Turner, K. Nimon, L. Zientek, and R. Henson, "Tools to support interpreting multiple regression in the face of multicollinearity," *Frontiers Psychol.*, vol. 3, p. 44, Mar. 2012, doi: 10.3389/fpsyg.2012.00044.

[36] U. Gromping, "Relative importance for linear regression in R: The package relaimpo," *J. Statist. Softw.*, vol. 17, no. 1, p. 27, 2006, doi: 10.18637/jss.v017.i01.

[37] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.1080/00401706.1970.10488634.

[38] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B-Methodol.*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[39] Z. Cui, M. Su, L. Li, H. Shu, and G. Gong, "Individualized prediction of reading comprehension ability using gray matter volume," *Cerebral Cortex*, vol. 28, no. 5, pp. 1656–1672, 2018, doi: 10.1093/cercor/bhx061.

[40] C. Feng *et al*., "Prediction of trust propensity from intrinsic brain morphology and functional connectome," *Human Brain Mapping*, vol. 42, no. 1, pp. 175–191, 2021, doi: 10.1002/hbm.25215.

[41] C. Sripada, S. Rutherford, M. Angstadt, W. K. Thompson, M. Luciana, A. Weigard, L. H. Hyde, and M. Heitzeg, "Prediction of neurocognition in youth from resting state fMRI," *Mol. Psychiatry*, vol. 25, no. 12, pp. 3413–3421, 2020, doi: 10.1038/s41380-019-0481-6.

[42] R. Jiang *et al*., "Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships," *Neuroimage*, vol. 207, p. 116,370, Feb. 2020, doi: 10.1016/j.neuroimage.2019.116370.

[43] K. Yoo, M. D. Rosenberg, W.-T. Hsu, S. Zhang, C.-S. R. Li, D. Scheinost, R. T. Constable, and M. M. Chun, "Connectome-based predictive modeling of attention: Comparing different functional connectivity features and prediction methods across datasets," *Neuroimage*, vol. 167, pp. 11–22, Feb. 2018, doi: 10.1016/j.neuroimage.2017.11.010.

[44] U. Grömping, "Variable importance in regression models," *Wiley Interdisciplinary Rev., Comput. Statist.*, vol. 7, no. 2, pp. 137–152, 2015, doi: 10.1002/wics.1346.

[45] L. A. Maglanoc *et al*., "Brain connectome mapping of complex human traits and their polygenic architecture using machine learning," *Biol. Psychiatry*, vol. 87, no. 8, pp. 717–726, 2020, doi: 10.1016/j.biopsych.2019.10.011.

[46] V. Zuber and K. Strimmer, "High-dimensional regression and variable selection using CAR scores," *Statist. Appl. Genetics Mol. Biol.*, vol. 10, no. 1, p. 34, 2011, doi: 10.2202/1544-6115.1730.

[47] C. Walsh and R. M. Nally, "The hier.part package," Hierarchical Partitioning. R Project for Statistical Computing, Version 0.5–1, 2003. [Online]. Available: http://cran.r-project.org/

[48] J. Liu, X. Liao, M. Xia, and Y. He, "Chronnectome fingerprinting: Identifying individuals and predicting higher cognitive functions using dynamic brain connectivity patterns," *Human Brain Mapping*, vol. 39, no. 2, pp. 902–915, 2018, doi: 10.1002/hbm.23890.

[49] L. Wei, B. Jing, and H. Li, "Bootstrapping promotes the RSFC-behavior associations: An application of individual cognitive traits prediction," *Human Brain Mapping*, vol. 41, no. 9, pp. 2302–2316, 2020, doi: 10.1002/hbm.24947.

[50] N. U. Dosenbach *et al*., "Prediction of individual brain maturity using fMRI," *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010, doi: 10.1126/science.1194144.

[51] S. D. Lichenstein, D. Scheinost, M. N. Potenza, K. M. Carroll, and S. W. Yip, "Dissociable neural substrates of opioid and cocaine use identified via connectome-based modelling," *Mol. Psychiatry*, vol. 26, pp. 4383–4393, Nov. 2019, doi: 10.1038/s41380-019-0586-y.

[52] M. D. Rosenberg *et al*., "Methylphenidate modulates functional network connectivity to enhance attention," *J. Neurosci.*, vol. 36, no. 37, pp. 9547–9557, 2016, doi: 10.1523/JNEUROSCI.1746-16.2016.

[53] J. Yu, I. Rawtaer, J. Fam, L. Feng, E.-H. Kua, and R. Mahendran, "The individualized prediction of cognitive test scores in mild cognitive impairment using structural and functional connectivity features," *Neuroimage*, vol. 223, p. 117,310, Dec. 2020, doi: 10.1016/j.neuroimage.2020.117310.

[54] S. Gao, A. S. Greene, R. T. Constable, and D. Scheinost, "Combining multiple connectomes improves predictive modeling of phenotypic measures," *Neuroimage*, vol. 201, p. 116,038, Nov. 2019, doi: 10.1016/j.neuroimage.2019.116038.

[55] C. Feng, Z. Cui, D. Cheng, R. Xu, and R. Gu, "Individualized prediction of dispositional worry using white matter connectivity," *Psychol. Med.*, vol. 49, no. 12, pp. 1999–2008, 2019, doi: 10.1017/S0033291718002763.

[56] R. E. Beaty *et al*., "Robust prediction of individual creative ability from brain functional connectivity," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 5, pp. 1087–1092, 2018, doi: 10.1073/pnas.1713532115.

[57] R. Jiang *et al*., "Multimodal data revealed different neurobiological correlates of intelligence between males and females," *Brain Imag. Behav.*, vol. 14, no. 5, pp. 1979–1993, 2020, doi: 10.1007/s11682-019-00146-z.

[58] N. Reggente, T. D. Moody, F. Morfini, C. Sheen, J. Rissman, J. O'Neill, and J. D. Feusner, "Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive-compulsive disorder," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 9, pp. 2222–2227, 2018, doi: 10.1073/pnas.1716686115.

[59] J. Dubois, P. Galdi, L. K. Paul, and R. Adolphs, "A distributed brain network predicts general intelligence from resting-state human neuroimaging data," *Philos. Trans. R Soc. London B Biol. Sci.*, vol. 373, no. 1756, p. 20,170,284, 2018, doi: 10.1098/rstb.2017.0284.

[60] H. J. V. Rutherford, M. N. Potenza, L. C. Mayes, and D. Scheinost, "The application of connectome-based predictive modeling to the maternal brain: Implications for mother-infant bonding," *Cerebral Cortex*, vol. 30, no. 3, pp. 1538–1547, 2020, doi: 10.1093/cercor/bhz185.

[61] A. N. Nielsen, D. J. Greene, C. Gratton, N. U. F. Dosenbach, S. E. Petersen, and B. L. Schlaggar, "Evaluating the prediction of brain maturity from functional connectivity after motion artifact denoising," *Cerebral Cortex*, vol. 29, no. 6, pp. 2455–2469, 2019, doi: 10.1093/cercor/bhy117.

[62] A. Abi-Dargham and G. Horga, "The search for imaging biomarkers in psychiatric disorders," *Nature Med.*, vol. 22, no. 11, pp. 1248–1255, 2016, doi: 10.1038/nm.4190.

[63] R. Jiang *et al*., "Gender differences in connectome-based predictions of individualized intelligence quotient and sub-domain scores," *Cerebral Cortex*, vol. 30, no. 3, pp. 888–900, 2020, doi: 10.1093/cercor/bhz134.

[64] J. Sui *et al*., "Multimodal neuromarkers in schizophrenia via cognition-guided MRI fusion," *Nature Commun.*, vol. 9, no. 1, p. 3028, 2018, doi: 10.1038/s41467-018-05432-w.

[65] D. Wang *et al*., "Individual-specific functional connectivity markers track dimensional and categorical features of psychotic illness," *Mol. Psychiatry*, vol. 25, no. 9, pp. 2119–2129, 2020, doi: 10.1038/s41380-018-0276-1.

[66] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *Neuroimage*, vol. 145, no. Pt B, pp. 137–165, 2017, doi: 10.1016/j.neuroimage.2016.02.079.

[67] A. T. Drysdale *et al*., "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nature Med.*, vol. 23, no. 1, pp. 28–38, 2017, doi: 10.1038/nm.4246.

[68] W. Tian, Y. Liu, Y. Heo, D. Yan, Z. Li, J. An, and S. Yang, "Relative importance of factors influencing building energy in urban environment," *Energy*, vol. 111, pp. 237–250, Sep. 2016, doi: 10.1016/j.energy.2016.05.106.

[69] Z. Cui *et al*., "Optimization of energy state transition trajectory supports the development of executive function during youth," *Elife*, vol. 9, p. e53060, Mar. 2020, doi: 10.7554/eLife.53060.

[70] Y. Ju *et al*., "Connectome-based models can predict early symptom improvement in major depressive disorder," *J. Affect. Disorders*, vol. 273, pp. 442–452, Aug. 2020, doi: 10.1016/j.jad.2020.04.028.

SP