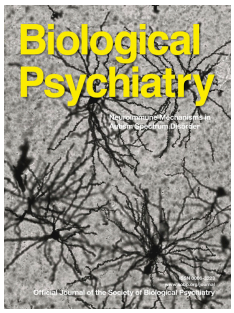
Journal Pre-proof



Multimodal Representation Learning for Parsing Biological Heterogeneity in Psychiatric Neuroimaging

Logan Grosenick, Conor Liston

PII: S0006-3223(26)00086-7

DOI: https://doi.org/10.1016/j.biopsych.2026.01.023

Reference: BPS 16045

To appear in: *Biological Psychiatry*

Received Date: 9 July 2025

Revised Date: 4 January 2026

Accepted Date: 17 January 2026

Please cite this article as: Grosenick L. & Liston C., Multimodal Representation Learning for Parsing Biological Heterogeneity in Psychiatric Neuroimaging, *Biological Psychiatry* (2026), doi: https://doi.org/10.1016/j.biopsych.2026.01.023.

Multimodal Representation Learning for Parsing Biological Heterogeneity in Psychiatric Neuroimaging

Logan Grosenick[1] and Conor Liston[1]

1. Department of Psychiatry and
Brain and Mind Research Institute
Weill Cornell Medicine, New York, NY

Correspondence to: log4002@med.cornell.edu
or col2004@med.cornell.edu

Short Title:
Parsing Heterogeneity in Psychiatric Neuroimaging

Abstract

For decades, psychiatric neuroimaging has searched for biomarkers of depression and other disorders, but they remain elusive in clinical practice. While the last five years have seen rapid progress, other large-scale correlative studies have found only small, unreliable links between brain measures and clinical symptoms. Growing evidence suggests that such limitations are not just about sample size but depend critically on how models represent data. This review traces a recent shift away from univariate methods to multivariate/multiview approaches that learn more effective representations of biological and symptom measures by flexibly learning multimodal latent representations. We first review how linear multiview embedding methods have revealed reproducible biological depression subtypes, but do not perform well in small samples or samples enriched for mild symptoms. We then consider newer work exploring more sophisticated representations for neuroimaging data, including deep-learning and graph-based representations, and multimodal extensions that uncover complex latent patterns that single-modality studies miss. We then review recent developments in "foundation" models which, once trained on large corpora, can "transfer learn" readily to small clinical cohorts, potentially bringing the advantages of large-scale learning to small, privacy-limited data. Finally, we highlight emerging representation tools that treat the brain as a dynamic, stateful multivariate process. Taken together, these advances point to a future in which the value of neuroimaging will be determined not just by ever-larger sample sizes but also by data quality and by how well our algorithms capture the distributed, multimodal, and evolving nature of psychiatric disorders.

Linking psychiatric symptoms and behaviors to their underlying neurobiological mechanisms, and ultimately discovering biomarkers for diagnosis and treatment, is a primary goal of psychiatric neuroscience. Despite rapid progress in the past decade, clinically useful biomarkers remain elusive for two related reasons: psychiatric diagnoses are heterogeneous, complicating efforts to model them as unitary conditions; and univariate effect sizes in psychiatric neuroimaging are typically small and difficult to replicate.

These challenges are especially apparent in depression, a primary focus of this review. Major depressive disorder (MDD) is not a unitary disease but a clinical label for a constellation of symptoms that can combine in >200 distinct ways under current diagnostic criteria. Distinct mechanisms underlie divergent presentations, and conversely, different mechanistic pathways can yield superficially similar symptom profiles. The same pathophysiological mechanism may also manifest differently across individuals, shaped by environmental factors, neurodevelopment, and other moderators.

We view this heterogeneity not only as a challenge but also as a scientific opportunity. Parsing individual differences in symptoms, neurobiology, and treatment response has the potential to sharpen effect sizes and reveal mechanistically distinct subtypes. Rather than seeking one-size-fits-all models, this approach aims to define biologically coherent subgroups and brain-behavior dimensions that explain individual differences, and ultimately support mechanistic, personalized treatments and improved outcomes.

Perhaps unsurprisingly given their heterogeneity, large-scale, multi-site psychiatric neuroimaging studies often report small effect sizes. Analyses relying on mass-univariate statistics may return small correlations (|r|~=0.05-0.1) that may not replicate unless sample sizes are >1,000 (1). It has thus been suggested that analyses of large-scale consortia datasets may, in a rough statistical sense, resemble genome-wide association studies in that univariate effect sizes are small and require many observations to reproduce (1). This facet of large cross-sectional correlations relates to broader challenges in neuroimaging research, limitations of reverse inference, and severe multiple comparisons burdens that may not capture the distributed nature of brain function.

Importantly, small univariate correlations may not signal any analytic failure but instead reflect how psychiatrically relevant patterns are represented in brain-wide neuroimaging data. Distributed network dysconnectivity models posit that symptoms emerge from subtle, spatially dispersed, multi-locus connectivity perturbations that only become behaviorally salient in aggregate (2–4). Looking across time, "state space models" (5,6) conceptualize mental illness as deviations from healthy dynamic trajectories observed in high-dimensional measurements but generated by underlying stateful brain-based manifolds (7,8). Such state shifts may leave only faint shadows in any one region-of-interest contrast when averaged together (2,9,10), and transient effects may be averaged away across time and dominated by static functional connectivity alterations (10–12). Finally, study inclusion criteria are critical: broad categories such as “psychopathology,” mixtures of diagnoses, or samples including mild, severe, and fully remitted patients can mask effects (13).

Large-scale reproducibility studies (1,14) and work in affective and neuro-developmental disorders (15–19) show that multiview models can recover stronger, more stable links by learning shared latent factors that span complex constellations of brain regions and symptoms. Flexibly learning such latent factors, rather than hand-picking features one at a time, is the essence of *representation learning* (20), which seeks latent embeddings in which meaningful structures emerge (**Suppl. Box 1**). Dynamic representation learning variants learn which latent network states a patient's brain occupies at a given time and how it transitions between them. Other approaches have begun to identify rapid spatiotemporal changes that may not be detectable in traditional static analyses (8,10,12,21). These methods move us toward estimating not only "where" but also "when" aberrant processes emerge, laying groundwork for personalized, time-sensitive interventions. In practice, such interventions will require intensively sampled longitudinal datasets that link repeated symptom and context measures with periodic neuroimaging and/or electrophysiology. Finally, psychiatric models must capture overlapping syndromes and marked person-to-person variability to uncover robust, clinically meaningful subtypes and individual deviations that group-average contrasts miss.

To illustrate these principles, we begin with a focused review of data-driven approaches to parsing heterogeneity in depression, starting with linear multiview embeddings. We then review how nonlinear multimodal methods and related approaches discover robust multimodal and dynamic representations, fusing diverse data types and modeling longitudinal trajectories in heterogeneous populations. Throughout, we emphasize the complementary strengths and limitations of multiview modeling in large (often multi-site) datasets and in smaller, densely sampled cohorts. Although we emphasize depression as a compelling case study, we also draw on work like that in schizophrenia and autism that has benefited from large public datasets. The core principles, however, should apply broadly across psychiatric domains.

Recent Progress in Linear Multiview Modeling of Neurobiological Heterogeneity

In this section we explain how linear models can link patterns of brain connectivity with symptom profiles to parse heterogeneity, using depression as a case study. We first describe approaches that cluster patients into data-driven subgroups, then approaches that place individuals along continuous symptom-brain dimensions, and finally hybrid methods that combine both ideas. The key point is that instead of looking at one region or one symptom at a time, multiview methods learn shared patterns across whole-brain connectivity and clinical measures, yielding subtypes and dimensions based on covariance structures that often relate to treatment response. Importantly, by isolating shared covariation (e.g., between clinical symptoms and brain connectivity), multiview methods can surface disease-relevant signal that would otherwise be obscured by the larger marginal variance within each data type alone.

Three general strategies have gained traction for building models for parsing heterogeneity in depression using neuroimaging and other biological measures: categorical, dimensional, and hybrid approaches. Categorical models carve the population into discrete subgroups and then assess whether these clusters differ meaningfully in symptoms or clinical outcomes. This approach may be intuitively appealing for some clinicians, by offering labels that align with traditional diagnostic heuristics, which are typically categorical, not dimensional. In an early and

influential example of this approach, Price and colleagues used data-driven clustering of resting-state fMRI connectivity measures to define two subtypes of depression (22,23) with distinct symptom profiles and connectivity patterns involving the default mode network, ventral affective network, and cognitive control network. Interestingly, one subgroup was associated with *reduced* default mode network connectivity while the second was associated with *increased* connectivity in ventral affective areas. This pattern of an optimal two-cluster subtyping solution involving one cluster with predominantly hyperconnectivity and another with predominantly hypoconnectivity has also been observed in other studies (24–26).

In contrast, dimensional approaches eschew discrete boundaries, modeling variation continuously across individuals (17,18,27,28) by identifying brain-behavior axes that explain individual differences in symptoms. This strategy is particularly well suited to capturing the graded nature of psychiatric phenomena and aligns with frameworks like the Research Domain Criteria (RDoC) and the hierarchical taxonomy of psychopathology (HiTOP) models (29,30). For example, Xia and colleagues identified four brain-behavior dimensions explaining individual differences in mood-, psychosis-, fear-, and externalizing-related symptoms using sparse canonical correlation analysis (CCA; a linear multiview embedding) in a large transdiagnostic youth sample (17). More recently, a similar approach in a larger sample of >1,000 adults with MDD, alcohol use disorder, eating disorders, or no psychiatric diagnosis identified six generalizable brain-behavior dimensions associated with depressed mood, impulsivity, emotion dysregulation, stress, disordered eating, and social avoidance (18).

Finally, hybrid approaches combine these strengths by first modeling continuous associations between brain function and clinical symptoms and then clustering individuals based on their dimension scores (15,19,31,32), or jointly modeling dimensions and clusters (33). An advantage of this approach is that it allows for mechanistic interpretation of subtypes while preserving the dimensional richness of the underlying data (limitations are discussed below). In our prior work, CCA identified two brain-behavior dimensions explaining individual differences in anhedonia or anxiety and insomnia, respectively (15). Hierarchical clustering revealed four subtypes in this two-dimensional latent space, which were associated with distinct clinical symptom profiles (e.g. higher or lower levels of anhedonia or anxiety) and varying responses to repetitive transcranial magnetic stimulation (rTMS)—a key step toward clinically actionable subtyping models.

We went on to show that generalization could be improved by incorporating regularization and a stabilized bootstrapped feature selection procedure (13), and later work replicated key findings in a larger sample with deeper clinical phenotyping (31) (**Fig. 1A-F)**. Leveraging a richer dataset, the latter work identified three brain-behavior dimensions: one explaining individual differences in anhedonia as in (15) while the second and third tracked with anxiety and insomnia, respectively, features that were previously fused into a single construct in (15). These studies, in turn, gave rise to a large-scale, multi-site randomized controlled trial (NCT 04041479) that is currently testing their practical utility for improving antidepressant outcomes. Similarly, pioneering work from Dr. Leanne Williams and colleagues has shown how hybrid modeling that incorporates resting state and task-evoked fMRI signals can capture new forms of complexity (19,34). In one such example (19), hierarchical clustering on 41 dimensional circuit scores in six depression-related networks

identified six categorical biotypes of depression and anxiety that predicted differential responses to treatments such as venlafaxine and psychotherapy.

Together, these strategies are already yielding replicable subtype solutions that predict outcomes in depression. At the same time, this work has revealed several challenges. First, when applied to high-dimensional neuroimaging data, CCA and related approaches classically "overfit" and do not perform well with small samples and/or very small univariate correlations (**Fig. 1G-H**) (1,13,14). Second, effect sizes for canonical correlations will be influenced by factors including fMRI data quality, the reliability of clinical symptom scores, fMRI artifacts, and sample heterogeneity. Interestingly, we have consistently observed that effect sizes are smaller in samples enriched for participants with mild or absent symptoms (**Fig. 1I**). For early-intervention and youth cohorts, this implies that brain-symptom effect sizes could be modest if we rely solely on cross-sectional symptom scores. One way forward is to augment these designs with additional, more trait-like "views" of risk (e.g. genetics, early-life adversity, sleep, reward-behavior), which can provide complementary signal even when current symptoms are mild. Third, scanner effects are often much larger than clinically relevant effects. This is a major problem for multi-site datasets, especially when the samples from a given site are not carefully matched with respect to clinical symptoms. In such cases, removing scanner effects will also remove or degrade signals linking fMRI measures and clinical data. These challenges notwithstanding, a host of studies have now shown that with careful attention to data quality, preprocessing, sample heterogeneity, and model training, it is possible to generate robust and reproducible multivariate models linking fMRI data and clinical symptoms (13–19,31).

Promising new cluster-aware matrix factorization methods like P3CA (33) extend this virtue of interpretability while sidestepping two long-standing pain points: (i) the need to first embed and then cluster in a two-stage procedure (where the embedding knows nothing about the clustering and vice versa) and (ii) the brittle scaling of CCA and clustering to tens of thousands of network edges and many observations. By jointly discovering patient clusters and the latent cross-modal canonical axes that best separate them, such methods preserve the interpretability clinicians value, while providing a distributed and clustered representation. New work, reviewed in detail below, extends this work further in modeling brain-behavior associations in depression and other highly heterogeneous psychiatric conditions.

Cross-sectional Multiview Mapping

A first wave of multiview work in psychiatry extended simple linear CCA-style models to cope with higher dimensionality and more complex brain-symptom relationships. Rather than hand-picking a few regions or scales, these approaches learn sparse and/or low-dimensional latent factors that summarize distributed connectivity-symptom constellations. Here we give a brief conceptual overview; methodological details and specific examples are provided in **Suppl. Note 1**.

Sparse and structured-sparse linear multiview methods use penalties such as Lasso, Elastic-Net, or GraphNet to prune weak connections, yielding compact loading maps that improve interpretability and generalizability. These approaches have identified a small number of brain-symptom dimensions in disorders such as depression and schizophrenia, mapping onto clinically

familiar constructs (e.g., anhedonia, anxiety) traceable to specific networks (35–37). Kernel CCA and multiple-kernel learning relax linearity by mapping each modality into a higher-dimensional feature space before learning shared latent axes. In affective and psychosis-risk cohorts, such models can distinguish diagnostic groups, separate bipolar from unipolar depression, and predict psychosis conversion better than unimodal or purely linear baselines, albeit with heavier hyperparameter tuning and a greater risk of overfitting (36–39).

Deep multiview extensions, including deep CCA and related autoencoder architectures, push this logic further by letting deep neural networks learn nonlinear transformations before maximizing cross-modal correlation. These models can recover highly correlated shared embeddings and improve classification or scale harmonization compared with linear CCA, while beginning to reveal distributed "imaging-genetic connectomes" (40–42). However, they are data-hungry, sensitive to optimization choices, and more challenging to interpret than sparse linear models. Complementary manifold-learning tools such as diffusion maps, t-SNE, UMAP, and Temporal-PHATE provide exploratory low-dimensional visualizations of high-dimensional fMRI trajectories, making latent structure and state transitions easier to see, but are best treated as hypothesis-generating lenses rather than inferential models (43).

Taken together, these cross-sectional multiview methods show that relatively modest extensions of classical linear models can recover stronger brain-behavior links while preserving varying degrees of interpretability.

Multimodal Representation Learning and Emerging Models for Psychiatry

Here we focus on models that learn latent representations directly from data, including multimodal deep networks, graph neural networks, and emerging "foundation" models trained on thousands of scans. Our goal is to give an intuitive account of why concepts such as self-supervised and contrastive learning, or tokenization of brain signals matter for psychiatry (because they can improve generalization, enable transfer learning to small local data, and may reduce sample-size requirements). For more details, see **Suppl. Note 2**.

The Need for Nonlinear and Heterogeneous Data Integration. Linear multivariate pipelines that rely on hand-engineered feature sets plus classic regularization can still deliver strong brain-behavior links when inclusion criteria are narrow. For example, ridge-regularized CCA with feature selection applied to fronto-limbic connectivity edges in depressed patients captures depressive-symptom variance with fully held-out correlations of r~0.62 in single-site MDD cohorts (31) (**Fig. 2**). Transdiagnostic, population-scale, multi-site datasets expose the limits of this strategy: when the full resting-state FC matrix is retained and only ridge penalties are tuned across a large population, normative-youth correlations hover near r~0.30 (44) but fall toward ~0.20 when developmental heterogeneity is fully included (14). Thus, linear model performance may erode as phenotypic diversity widens (13).

At the same time, the range of available measurement modalities keeps expanding. Structural MRI, functional MRI, DTI, EEG, genetics, clinical assessments, and digital phenotypes are now routinely collected in large datasets, often spanning multiple time scales and sites. Emerging deep

learning methods are well suited for discovering complex, nonlinear relationships across such heterogeneous data, and can accommodate missing data by leveraging multiple “views” of the underlying latent processes. Interestingly, theory from nonlinear independent component analysis and multi-view representation learning have begun to show that access to multiple, conditionally independent “views” or auxiliary variables can render latent factors identifiable up to simple component-wise transformations (45). In principle, multimodal psychiatric datasets (e.g., fMRI+EEG+clinical scales) provide such rich auxiliary information, making it more plausible that deep multimodal models could recover stable and meaningful underlying latent processes rather than more arbitrary encodings.

High-Dimensional Geometry and Self-Supervision. In high-dimensional settings typical of brain data, distances between points tend to become indistinguishable, a “distance-concentration” effect that makes them less informative and can mislead classification or clustering (46,47). This is a manifestation of the infamous "curse of dimensionality" (48) and contributes to the fragility of univariate association and high-dimensional clustering studies. Classical pipelines combat these issues with explicit regularization (Ridge, Lasso, GraphNet) to constrain solutions and improve generalization.

By contrast, contemporary multimodal deep neural networks (DNNs) often use parameter counts that dwarf sample sizes, placing them in an "over-parameterized" regime (49). Counter-intuitively, once the model becomes flexible enough to interpolate the training data, test error can decrease as model size grows (a phenomenon called “double descent”), in part because over-parameterized models can still generalize when they converge to low-complexity solutions (50,51). A key enabler of such models is self-supervised pre-training, which allows them to learn structured representations from unlabeled data before any clinical labels are introduced.

Self-supervised methods for fMRI have begun to show efficacy (52), and contrastive learning methods (a subset of self-supervised techniques that maximize agreement between augmented views of the same data sample while minimizing agreement with other data samples) have shown particular promise in neuroimaging (53) and broader neuroscience (54) applications. Such approaches leverage large unlabeled datasets for representation learning and can improve downstream task performance with limited labeled data. However, while learned manifolds act as priors that reduce the number of cases needed to reach a target effect size, they also make pipelines sensitive to scanner harmonization, motion artifact, and label noise, which can degrade generalization (55,56). Consequently, state-of-the-art studies pair over-parameterized models with aggressive quality-control and sample-reweighting strategies (57,58).

Multimodal Fusion. Deep neural networks promise to advance multimodal integration in psychiatric neuroimaging by learning hierarchical representations that capture cross-modal interactions at multiple levels of abstraction. Early fusion approaches concatenate features from different modalities before processing, whereas late fusion combines modality-specific predictions after fitting separate encoders (59–61). For example, in psychosis-risk cohorts, multimodal machine-learning pipelines that fuse multiple neuroimaging modalities and integrate clinical, cognitive, and genetic information improve prediction of conversion compared with single-modality models (38,62,63). Fully deep multimodal fusion architectures that jointly embed resting-

state functional connectivity, structural MRI, and clinical features using intermediate strategies are also emerging in affective disorders (64,65). Within these architectures, attention mechanisms and other learned weighting schemes can highlight which combinations of connectivity patterns and clinical features contribute most strongly to predictions, offering candidate features for further mechanistic investigation (although we must remain interpretively cautious for complex architectures) (64,65).

Graph Neural Networks: "Connectome-Native" Representation Learning. Graph neural networks occupy a sweet spot between classical graph-theoretic metrics and fully connected deep nets, as their message-passing layers treat the connectome itself as the "coordinate system". This connectome-native inductive bias should be an improvement for representation learning on brain networks and can capture network similarity more faithfully (66,67). Compared with earlier pipelines, modern graph neural networks (GNNs) ingest heterogeneous node attributes (e.g., cortical thickness, gene-expression gradients) and model edge dynamics, making them well-suited for multimodal fusion (68). Applications to psychiatric and related brain-disorder datasets include temporal-adaptive graph convolutional networks for major depressive disorder, atlas-constrained graph-attention networks for autism, dynamic graph convolutional frameworks for ADHD, and local-to-global graph neural networks for ASD and Alzheimer's disease (69–71). GNNs with "learnable neighborhood quantization" further extend the successful inductive bias or convolutional neural network (CNN) architectures to irregular graphs like cortical or EEG scalp meshes (72). At the same time, the expressive power that allows GNNs to capture subtle patterns can also promote overfitting, especially when scanners, parcellations, or demographic strata shift (1,14,73). Attention weights, gradient-based saliency maps, and related attribution scores remain post-hoc proxies for interpretability; systematic reviews highlight that edge-importance rankings and putative "biomarker" subnetworks are often unstable across perturbations, datasets or training runs (73). Looking ahead, methods that relax the need for a single fixed atlas (for example, by supporting multi-resolution or atlas-free representations) could improve cross-site generalization, though they will still depend on well-characterized input graphs.

Foundation Models and Large-Scale Representation Learning. "Foundation Models" (FMs) are large, overparameterized neural networks pre-trained on broad datasets and then lightly adapted ("fine-tuned") to specific downstream tasks. In neuroimaging, FMs trained on fMRI, connectomes, and EEG/iEEG use self-supervised objectives such as masked-volume prediction and cross-modal contrastive losses to reshape the raw, "distance-concentrated" signal space into lower-dimensional manifolds where distances are more likely to track biology rather than noise (52,54,74,75). Unimodal fMRI FMs, cross-connectome models, and multimodal or brain-text/vision FMs support downstream prediction of age, symptoms, diagnosis, and treatment response (76,77) (**Fig. 2**). A prerequisite for applying transformer-style architectures to brain data is tokenization: converting continuous neural signals into discrete input units. Current approaches tokenize fMRI volumes as spatial parcels, temporal windows, spatiotemporal patches, or learned discrete codes, and this choice directly shapes what patterns the model can capture. It also determines the "effective token" counts used in scaling comparisons such as Fig. 3, making cross-modality comparisons necessarily approximate. With that caveat in mind, in **Fig. 3** parameter-token space, the largest brain foundation models currently sit approximately between BERT (2018) and GPT-2, farther from more modern text foundation models. They also cluster in the

lower-left region of the plot, lying well above the text-LLM optimality frontier (78,79) indicating that, for a given parameter count, they are typically trained on far fewer effective tokens than current language models (see Supplementary Note 3 for further details and caveats). Together, these patterns suggest that brain foundation models are still in an early, under-trained/data-starved "BERT/GPT-2 era", implying that major gains in psychiatric prediction and clinical utility may be on the horizon. Still, many brain foundation model results cited here are preprints, so reported gains should be interpreted with appropriate reservations pending peer review and independent replication.

Longitudinal, Dynamic, and Causal Representation Learning

Thus far we have focused on static, cross-sectional "snapshot views" of the brain. Here we briefly review longitudinal models and those that track how brain networks change over time and how those changes relate to symptoms, cognition, and treatment.

Precision Functional Mapping and Serial Imaging. Precision functional mapping (PFM) has shifted our understanding of individual brain organization through dense, within-person sampling. By recovering individualized network architectures, PFM also provides personalized coordinate systems that can serve as inputs for downstream multimodal representation learning. Early work (80,81) collected many hours of resting-state fMRI from single individuals, revealing highly reliable, idiosyncratic network architectures, and (82) showed these "connectome fingerprints" enable individual identification with >90% accuracy. More recent studies have made PFM more clinically feasible: (83) demonstrated that multi-echo acquisition can recover reliable individual network maps from tens of minutes rather than many hours of scanning, and (84) showed that these individual-specific networks are stable across years in depression cohorts, identifying a nearly twofold expansion of frontostriatal salience networks as a trait-like marker, largely unaffected by mood or neuromodulation and detectable in children before symptom onset.

Stateful Dynamic Models. Dynamic models extend representation learning from static snapshots to temporal sequences, learning when and how brain states shift (11). Hidden Markov models reveal transient brain states invisible to static analyses, with (12) demonstrating hierarchically organized temporal dynamics across rest and task states, and (10) identifying fast transient networks in spontaneous activity that static connectivity measures miss. These discrete state-space models have proven useful in cognitive and clinical neuroscience: (85) uncovered hidden brain state dynamics that regulate cognitive performance and decision-making, linking transitions between states to behavioral variability. K-means clustering has likewise been used to define discrete connectivity states for Network Control Theory analyses; for example, Singleton et al. (86) clustered dynamic connectivity patterns when relating LSD and psilocybin to changes in the brain's control energy landscape. Ajirak et al. (8) introduced discrete representation learning for multivariate fMRI time series, discretizing continuous connectivity dynamics into switching "states" and bridging classical state-space modeling with interpretable representation learning.

Transformer-based and new related state space models are now beginning to reshape spatiotemporal neuroimaging analysis. For example, Kim et al. (21) introduced SwiFT, a transformer that applies 4D attention across space and time (and over local windows in

feature/frequency dimensions), significantly outperforming recurrent neural network baselines when predicting age and cognitive scores from fMRI. Similarly, (87) developed BolT (fused window transformers), showing that transformer architectures can capture long-range temporal dependencies in fMRI time series. For psychiatric applications, such attention mechanisms provide a flexible way to learn which patterns of regional activity, connectivity, and (for electrophysiological data) frequency bands are most predictive of behavioral or symptom trajectories, patterns often averaged out in traditional static analyses.

Emerging Methods: From Causal Inference to Adaptive Interventions and Federated Learning. Recent advances in causal representation learning provide formal frameworks for disentangling latent variables and reasoning about how interventions might affect neural representations (88), though applications to psychiatry remain largely theoretical. Related emerging concepts such as "Digital Twins" (personalized computational models that could simulate individual treatment responses) will likely depend on multimodal representation learning as a statistical substrate but remain conceptual in psychiatry, which lacks the explicit physical models that make Digital Twins tractable in engineering (89). Dynamic causal modeling (6) and network control theory (86) offer complementary perspectives on individual brain dynamics but are still in the process of translation to clinical practice.

A promising multimodal direction is integration of these methods with digital phenotyping and ecological momentary assessment (EMA). By aligning intensive longitudinal mood and behavior measurements (often collected via smartphones or wearables) with neuroimaging and other biomarkers, researchers can build dynamic models of symptom trajectories (90). Just-in-time adaptive interventions (JITAIs) leverage these continuous data streams to deliver personalized interventions when most needed (91), and early micro-randomized JITAI pilots in individuals with elevated depressive symptoms demonstrate feasibility in mood-disorder populations (92). Reinforcement learning methods for estimating dynamic treatment regimes offer a complementary formal framework for optimizing such sequential treatment decisions from observational or experimental data (93). However, current methods require substantially more data than typical clinical trials provide, off-policy evaluation (estimating a new policy's value from previously collected data) remains unreliable, and real-world settings where clinicians observe only partial information about a patient's underlying state call for partially observable decision-making frameworks (suited to the multimodal, incomplete information clinicians actually observe) that are still under development (93). When combined with periodic neuroimaging, these approaches can in principle enable unprecedented temporal resolution in understanding brain-behavior relationships.

Finally, federated learning methods (94,95) promise solutions to the privacy and data-sharing challenges inherent in multi-site psychiatric neuroimaging, potentially allowing institutions to collaborate on model development without sharing raw patient data. These approaches, together with fine-tuning of pre-trained foundation models, may prove particularly crucial for rare psychiatric conditions and deeply phenotyped cohorts where no single site can realistically accumulate sufficient sample size (provided care is taken to not average away such rare cases).

Challenges and Limitations

In this final section we summarize key obstacles that limit multimodal psychiatric neuroimaging and highlight why prospective validation, rigorous statistics, and interpretable-by-design models are essential for real-world impact.

Functional connectivity reflects temporal covariance in BOLD rather than direct anatomical or causal interactions, especially for widely distributed patterns that may arise from global factors, physiological confounds, or common inputs. Multiview and representation-learning methods can isolate reproducible distributed patterns, but interpreting these as mechanistic “connectivity” requires caution. Group differences in depression are often assumed to be pathological; yet deviations from normative connectivity can also reflect adaptive or compensatory reorganization, and distinguishing maladaptive from resilient changes generally requires longitudinal or multimodal data.

Despite promising advances, several other obstacles remain. A “dark matter” problem persists: clinically relevant signals may lie outside current measurement capabilities, and poor or heterogeneous data can amplify noise rather than signal, degrading predictions despite theoretical gains from multimodal fusion (96). High dimensionality and analytic flexibility encourage p-hacking and overfitting (97); feature selection bias, hyperparameter overfitting, and cross-validation leakage can substantially inflate performance, particularly when model selection and evaluation are not properly nested (98). Robust multi-site generalization is difficult, as scanner effects can persist despite harmonization (99) and models may inadvertently learn site differences rather than disease-related variation.

Fairness and bias are growing concerns: data imbalance and deployment context can disadvantage specific demographic groups (100), and brain MRI models can exhibit race- and sex-related biases even under seemingly neutral pipelines (101). Large, demographically diverse samples from consortia such as ENIGMA (102) and developmental cohorts showing sex- and age-dependent brain–symptom relationships (17,18) underscore that a single “global” network-clinical mapping is unlikely to be sufficient. Moving forward will require independent validation in new cohorts, fairness-aware evaluation, and interpretable models that prioritize real clinical utility over marginal gains in predictive accuracy.

Acknowledgements and Disclosures. This work was supported by the Hope for Depression Research Foundation, the Pritzker Neuropsychiatric Disorders Research Consortium, and grants from the National Institutes of Health to LG and CL (R01 MH118388, R01 MH118451, UG3 MH137656, R01 MH131534, R01 OD039830).

CL and LG are listed as inventors on Cornell University patent applications on neuroimaging biomarkers for depression that are pending or in preparation. CL has served on the Scientific Advisory Board for Delix Therapeutics and Brainify.AI.

Supplement Description:

Supplement Notes 1-3 and Boxes S1-S2

References

1. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, *et al.* (2022): Reproducible brain-wide association studies require thousands of individuals. *Nature* 603: 654–660.
2. Kringelbach ML, Deco G (2020): Brain states and transitions: Insights from computational neuroscience. *Cell Rep* 32: 108128.
3. Fornito A, Zalesky A, Breakspear M (2015): The connectomics of brain disorders. *Nat Rev Neurosci* 16: 159–172.
4. Fornito A, Bullmore ET, Zalesky A (2017): Opportunities and challenges for psychiatry in the connectomic era. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2: 9–19.
5. Paninski L, Ahmadian Y, Ferreira DG, Koyama S, Rahnama Rad K, Vidne M, *et al.* (2010): A new look at state-space models for neural data. *J Comput Neurosci* 29: 107–126.
6. Friston KJ, Preller KH, Mathys C, Cagnan H, Heinzle J, Razi A, Zeidman P (2019): Dynamic causal modelling revisited. *Neuroimage* 199: 730–744.
7. Linderman S, Johnson M, Miller A, Adams R, Blei D, Paninski L (2017): Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In: Singh A, Zhu J, editors. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54. Fort Lauderdale, FL, USA: PMLR, pp 914–922.
8. Marzieh Ajirak, Immanuel Elbau, Nili Solomonov, Logan Grosenick (2024): Discrete representation learning for multivariate time series. *EUSIPCO 2024*.
9. Shine JM, Breakspear M, Bell PT, Ehgoetz Martens KA, Shine R, Koyejo O, *et al.* (2019): Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nat Neurosci* 22: 289–296.
10. Baker AP, Brookes MJ, Rezek IA, Smith SM, Behrens T, Probert Smith PJ, Woolrich M (2014): Fast transient networks in spontaneous human brain activity. *Elife* 3: e01867.
11. Hutchison RM, Womelsdorf T, Allen EA, Bandettini PA, Calhoun VD, Corbetta M, *et al.* (2013): Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage* 80: 360–378.
12. Vidaurre D, Abeysuriya R, Becker R, Quinn AJ, Alfaro-Almagro F, Smith SM, Woolrich MW (2018): Discovering dynamic brain networks from big data in rest and task. *Neuroimage* 180: 646–656.
13. Grosenick L, Shi TC, Gunning FM, Dubin MJ, Downar J, Liston C (2019): Functional and Optogenetic Approaches to Discovering Stable Subtype-Specific Circuit Mechanisms in Depression. *Biol Psychiatry Cogn Neurosci Neuroimaging* 4: 554–566.
14. Spisak T, Bingel U, Wager TD (2023): Multivariate BWAS can be replicable with moderate sample sizes. *Nature* 615: E4–E7.
15. Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, Meng Y, *et al.* (2017): Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine* 23(1): 28–38.
16. Buch AM, Vértes PE, Seidlitz J, Kim SH, Grosenick L, Liston C (2023): Molecular and network-level mechanisms explaining individual differences in autism spectrum disorder. *Nat Neurosci* 26(4): 650–663.
17. Xia CH, Ma Z, Ciric R, Gu S, Betzel RF, Kaczkurkin AN, *et al.* (2018): Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat Commun* 9: 3003.
18. Lett TA, Vaidya N, Jia T, Polemiti E, Banaschewski T, Bokde ALW, *et al.* (2025): Framework for brain-derived dimensions of psychopathology. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2025.1246
19. Tozzi L, Zhang X, Pines A, Olmsted AM, Zhai ES, Anene ET, *et al.* (2024): Personalized brain circuit scores identify clinically distinct biotypes in depression and anxiety. *Nat Med* 30: 2076–2087.

20. Bengio Y, Courville A, Vincent P (2013): Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 35: 1798–1828.
21. Kim PY, Kwon J, Joo S, Bae S, Lee D, Jung Y, *et al.* (2023, July 12): SwiFT: Swin 4D fMRI Transformer. *arXiv [cs.CV]*. Retrieved from http://arxiv.org/abs/2307.05916
22. Price RB, Lane S, Gates K, Kraynak TE, Horner MS, Thase ME, Siegle GJ (2017): Parsing Heterogeneity in the Brain Connectivity of Depressed and Healthy Adults During Positive Mood. *Biol Psychiatry* 81: 347–357.
23. Price RB, Gates K, Kraynak TE, Thase ME, Siegle GJ (2017): Data-Driven Subgroups in Depression Derived from Directed Functional Connectivity Paths at Rest. *Neuropsychopharmacology* 42: 2623–2632.
24. Feder S, Sundermann B, Wersching H, Teuber A, Kugel H, Teismann H, *et al.* (2017): Sample heterogeneity in unipolar depression as assessed by functional connectivity analyses is dominated by general disease effects. *J Affect Disord* 222: 79–87.
25. Cheng Y, Xu J, Yu H, Nie B, Li N, Luo C, *et al.* (2014): Delineation of early and later adult onset depression by diffusion tensor imaging. *PLoS One* 9: e112307.
26. Liang S, Deng W, Li X, Greenshaw AJ, Wang Q, Li M, *et al.* (2020): Biotypes of major depressive disorder: Neuroimaging evidence from resting-state default mode network patterns. *NeuroImage Clin* 28: 102514.
27. Wen J, Fu CHY, Tosun D, Veturi Y, Yang Z, Abdulkadir A, *et al.* (2022): Characterizing heterogeneity in neuroimaging, cognition, clinical symptoms, and genetics among patients with late-life depression. *JAMA Psychiatry* 79: 464–474.
28. Mihalik A, Adams RA, Huys Q (2020): Canonical correlation analysis for identifying biotypes of depression. *Biol Psychiatry Cogn Neurosci Neuroimaging* 5: 478–480.
29. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, *et al.* (2010): Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* 167: 748–751.
30. Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, *et al.* (2017): The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *J Abnorm Psychol* 126: 454–477.
31. Dunlop K, Grosenick L, Downar J, Vila-Rodriguez F, Gunning FM, Daskalakis ZJ, *et al.* (2024): Dimensional and Categorical Solutions to Parsing Depression Heterogeneity in a Large Single-Site Sample. *Biol Psychiatry*. https://doi.org/10.1016/j.biopsych.2024.01.012
32. Tokuda T, Yoshimoto J, Shimizu Y, Okada G, Takamura M, Okamoto Y, *et al.* (2018): Identification of depression subtypes and relevant brain regions using a data-driven approach. *Sci Rep* 8: 14082.
33. Buch AM, Liston C, Grosenick L (2024): Simple and Scalable Algorithms for Cluster-Aware Precision Medicine. *Proc Mach Learn Res* 238: 136–144.
34. Williams LM (2016): Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *Lancet Psychiatry* 3: 472–480.
35. Song X, Li R, Wang K, Bai Y, Xiao Y, Wang Y-P (2023): Joint Sparse Collaborative Regression on imaging genetics study of schizophrenia. *IEEE/ACM Trans Comput Biol Bioinform* 20: 1137–1146.
36. Mihalik A, Ferreira FS, Moutoussis M, Ziegler G, Adams RA, Rosa MJ, *et al.* (2020): Multiple holdouts with stability: Improving the generalizability of machine learning analyses of brain-behavior relationships. *Biol Psychiatry* 87: 368–376.
37. Vai B, Parenti L, Bollettini I, Cara C, Verga C, Melloni E, *et al.* (2020): Predicting differential diagnosis between bipolar and unipolar depression with multiple kernel learning on multimodal structural neuroimaging. *Eur Neuropsychopharmacol* 34: 28–38.
38. Reinen JM, Polosecki P, Castro E, Corcoran CM, Cecchi GA, Colibazzi T (2024): Multimodal fusion of brain signals for robust prediction of psychosis transition. *Schizophrenia (Heidelb)* 10: 54.

39. Castro E, Gómez-Verdejo V, Martínez-Ramón M, Kiehl KA, Calhoun VD (2014): A multiple kernel learning approach to perform classification of groups from complex-valued fMRI data analysis: application to schizophrenia. *Neuroimage* 87: 1–17.
40. Li G, Han D, Wang C, Hu W, Calhoun VD, Wang Y-P (2020): Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia. *Comput Methods Programs Biomed* 183: 105073.
41. Qi J, Tejedor J (2016): Deep multi-view representation learning for multi-modal features of the schizophrenia and schizo-affective disorder. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. https://doi.org/10.1109/icassp.2016.7471816
42. Wang M, Shao W, Hao X, Huang S, Zhang D (2022): Identify connectome between genotypes and brain network phenotypes via deep self-reconstruction sparse canonical correlation analysis. *Bioinformatics* 38: 2323–2332.
43. Busch EL, Huang J, Benz A, Wallenstein T, Lajoie G, Wolf G, *et al.* (2023): Multi-view manifold learning of human brain-state trajectories. *Nat Comput Sci* 3: 240–253.
44. Nicolaisen-Sobesky E, Mihalik A, Kharabian-Masouleh S, Ferreira FS, Hoffstaedter F, Schwender H, *et al.* (2022): A cross-cohort replicable and heritable latent dimension linking behaviour to multi-featured brain structure. *Commun Biol* 5: 1297.
45. Khemakhem I, Kingma DP, Hyvärinen A (2019): Variational Autoencoders and Nonlinear ICA: A Unifying Framework ((S. Chiappa & R. Calandra, editors)). *AISTATS* 108: 2207–2217.
46. Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999): When is “nearest neighbor” meaningful? *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp 217–235.
47. Aggarwal CC, Hinneburg A, Keim DA (2001): On the surprising behavior of distance metrics in high dimensional space. *Database Theory — ICDT 2001*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp 420–434.
48. Bellman RE (2021): *Adaptive Control Processes*. Hassell Street Press.
49. Belkin M, Hsu D, Ma S, Mandal S (2019): Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc Natl Acad Sci U S A* 116: 15849–15854.
50. Neyshabur B, Bhojanapalli S, McAllester D, Srebro N (2017): Exploring generalization in deep learning. *Neural Inf Process Syst* 5947–5956.
51. Bartlett PL, Long PM, Lugosi G, Tsigler A (2019, June 26): Benign Overfitting in Linear Regression. *arXiv [stat.ML]*. https://doi.org/10.1073/pnas.1907378117
52. Thomas AW, Ré C, Poldrack RA (2022, June 22): Self-supervised learning of brain dynamics from broad neuroimaging data. *arXiv [q-bio.NC]*. Retrieved from http://arxiv.org/abs/2206.11417
53. Dufumier B, Gori P, Victor J, Grigis A, Wessa M, Brambilla P, *et al.* (2021): Contrastive learning with continuous proxy meta-data for 3D MRI classification. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Cham: Springer International Publishing, pp 58–68.
54. Schneider S, Lee JH, Mathis MW (2023): Learnable latent embeddings for joint behavioural and neural analysis. *Nature* 617: 360–368.
55. D’Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, *et al.* (2020, November 6): Underspecification presents challenges for credibility in modern machine learning. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/2011.03395
56. Sagawa S, Koh PW, Hashimoto TB, Liang P (2019, November 20): Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv [cs.LG]*. Retrieved July 6, 2025, from http://arxiv.org/abs/1911.08731
57. Liu EZ, Haghgoo B, Chen AS, Raghunathan A, Koh PW, Sagawa S, *et al.* (2021, July 19):

Just train twice: Improving group robustness without training group information. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/2107.09044
58. Ren M, Zeng W, Yang B, Urtasun R (2018, March 23): Learning to reweight examples for robust deep learning. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/1803.09050
59. Baltrusaitis T, Ahuja C, Morency L-P (2019): Multimodal machine learning: A survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell* 41: 423–443.
60. Li G, Wang C, Han D-P, Zhang Y-P, Peng P, Calhoun VD, Wang Y-P (2020): Deep principal correlated auto-encoders with application to imaging and genomics data integration. *IEEE Access* 8: 20093–20107.
61. Kanyal A, Mazumder B, Calhoun VD, Preda A, Turner J, Ford J, Ye DH (2024): Multi-modal deep learning from imaging genomic data for schizophrenia classification. *Front Psychiatry* 15: 1384842.
62. Sanfelici R, Dwyer DB, Antonucci LA, Koutsouleris N (2020): Individualized diagnostic and prognostic models for patients with psychosis risk syndromes: A meta-analytic view on the state of the art. *Biol Psychiatry* 88: 349–360.
63. Koutsouleris N, Dwyer DB, Degenhardt F, Maj C, Urquijo-Castro MF, Sanfelici R, *et al.* (2021): Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry* 78: 195–209.
64. Liu S, Zhou J, Zhu X, Zhang Y, Zhou X, Zhang S, *et al.* (2024): An objective quantitative diagnosis of depression using a local-to-global multimodal fusion graph neural network. *Patterns (N Y)* 5: 101081.
65. Jiao Y, Zhao K, Wei X, Carlisle NB, Keller CJ, Oathes DJ, *et al.* (2025): Deep graph learning of multimodal brain networks defines treatment-predictive signatures in major depression. *Mol Psychiatry*. https://doi.org/10.1038/s41380-025-02974-6
66. Bessadok A, Mahjoub MA, Rekik I (2023): Graph neural networks in network neuroscience. *IEEE Trans Pattern Anal Mach Intell* 45: 5833–5848.
67. Luo X, Wu J, Yang J, Xue S, Beheshti A, Sheng QZ, *et al.* (2024): Graph neural networks for brain graph learning: A survey. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. presented at the Thirty-Third International Joint Conference on Artificial Intelligence {IJCAI-24}, California: International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2024/903
68. Mohammadi H, Karwowski W (2024): Graph Neural Networks in brain connectivity studies: Methods, challenges, and future directions. *Brain Sci* 15. https://doi.org/10.3390/brainsci15010017
69. Yao D, Sui J, Yang E, Yap P-T, Shen D, Liu M (2020): Temporal-adaptive graph convolutional network for automated identification of major depressive disorder using resting-state fMRI. *Mach Learn Med Imaging* 12436: 1–10.
70. Yang C, Wang P, Tan J, Liu Q, Li X (2021): Autism spectrum disorder diagnosis using graph attention network based on spatial-constrained sparse functional brain networks. *Comput Biol Med* 139: 104963.
71. Zhao K, Duka B, Xie H, Oathes DJ, Calhoun V, Zhang Y (2022): A dynamic graph convolutional neural network framework reveals new insights into connectome dysfunctions in ADHD. *Neuroimage* 246: 118774.
72. Nkansah IO, Gallagher N, Sandilya R, Liston C, Grosenick L (2024): Generalizing CNNs to graphs with learnable neighborhood quantization. *Advances in Neural Information Processing Systems* 37: 82318–82349.
73. Chan YH, Girish D, Gupta S, Xia J, Kasi C, He Y, *et al.* (2024, May 1): Discovering robust biomarkers of psychiatric disorders from resting-state functional MRI via graph neural networks: A systematic review. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/2405.00577

74. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, *et al.* (2021, February 26): Learning transferable visual models from natural language supervision. *arXiv [cs.CV]*. Retrieved from http://arxiv.org/abs/2103.00020
75. Wang T, Isola P (2020, May 20): Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/2005.10242
76. Wang Y-W, Holmes AJ, Yeo BTT, Yip SW (2025): From big to small: Emerging methods for enhancing precision psychiatry through transfer learning. *Biol Psychiatry*. https://doi.org/10.1016/j.biopsych.2025.10.022
77. Caro JO, Fonseca AH de O, Averill C, Rizvi SA, Rosati M, Cross JL, *et al.* (2023, September 13): BrainLM: A foundation model for brain activity recordings. *bioRxiv*. https://doi.org/10.1101/2023.09.12.557460
78. Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, *et al.* (2022, March 29): Training Compute-Optimal Large Language Models. *arXiv [cs.CL]*. Retrieved from http://arxiv.org/abs/2203.15556
79. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, *et al.* (2020, January 22): Scaling laws for neural language models. *arXiv [cs.LG]*. https://doi.org/10.48550/arXiv.2001.08361
80. Laumann TO, Gordon EM, Adeyemo B, Snyder AZ, Joo SJ, Chen M-Y, *et al.* (2015): Functional system and areal organization of a highly sampled individual human brain. *Neuron* 87: 657–670.
81. Poldrack RA, Laumann TO, Koyejo O, Gregory B, Hover A, Chen M-Y, *et al.* (2015): Long-term neural and physiological phenotyping of a single human. *Nat Commun* 6: 8885.
82. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, *et al.* (2015): Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci* 18: 1664–1671.
83. Lynch CJ, Power JD, Scult MA, Dubin M, Gunning FM, Liston C (2020): Rapid Precision Functional Mapping of Individuals Using Multi-Echo fMRI. *Cell Rep* 33: 108540.
84. Lynch CJ, Elbau IG, Ng T, Ayaz A, Zhu S, Wolk D, *et al.* (2024): Frontostriatal salience network expansion in individuals in depression. *Nature* 633: 624–633.
85. Taghia J, Cai W, Ryali S, Kochalka J, Nicholas J, Chen T, Menon V (2018): Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nat Commun* 9: 2505.
86. Singleton SP, Luppi AI, Carhart-Harris RL, Cruzat J, Roseman L, Nutt DJ, *et al.* (2022): Receptor-informed network control theory links LSD and psilocybin to a flattening of the brain's control energy landscape. *Nat Commun* 13: 5812.
87. Bedel HA, Sivgin I, Dalmaz O, Dar SUH, Çukur T (2023): BolT: Fused window transformers for fMRI time series analysis. *Med Image Anal* 88: 102841.
88. Scholkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y (2021): Toward causal representation learning. *Proc IEEE Inst Electr Electron Eng* 109: 612–634.
89. Spitzer M, Dattner I, Zilcha-Mano S (2023): Digital twins and the future of precision mental health. *Front Psychiatry* 14: 1082598.
90. Lydon-Staley DM, Bassett DS (2018): The promise and challenges of intensive longitudinal designs for imbalance models of adolescent substance use. *Front Psychol* 9: 1576.
91. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, Murphy SA (2018): Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Ann Behav Med* 52: 446–462.
92. Elmer T, Wolf M, Snippe E, Scholz U (2025): A social support just-in-time adaptive intervention for individuals with depressive symptoms: Feasibility study with a microrandomized trial design. *JMIR Ment Health* 12: e74103.
93. Hargrave M, Spaeth A, Grosenick L (2024): EpiCare: A reinforcement learning benchmark

for Dynamic Treatment Regimes. *Advances in Neural Information Processing Systems* 37: 130536–130568.
94. Li T, Sahu AK, Talwalkar A, Smith V (2019, August 21): Federated learning: Challenges, methods, and future directions. *arXiv [cs.LG]*. https://doi.org/10.1109/MSP.2020.2975749
95. Li W, Milletarì F, Xu D, Rieke N, Hancox J, Zhu W, *et al.* (2019, October 2): Privacy-preserving federated brain tumour segmentation. *arXiv [cs.CV]*. Retrieved July 7, 2025, from http://arxiv.org/abs/1910.00962
96. Thompson PM, Jahanshad N, Ching CRK, Salminen LE, Thomopoulos SI, Bright J, *et al.* (2020): ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl Psychiatry* 10: 100.
97. Poldrack RA, Huckins G, Varoquaux G (2020): Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry* 77: 534–540.
98. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B (2017): Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage* 145: 166–179.
99. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, *et al.* (2018): Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 167: 104–120.
100. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, *et al.* (2024): Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol* 42: 3–15.
101. Piçarra C, Glocker B (2023): Analysing race and sex bias in brain age prediction. *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*. Cham: Springer Nature Switzerland, pp 194–204.
102. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, *et al.* (2014): The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav* 8: 153–182.
103. Royer J, Kebets V, Piguet C, Chen J, Ooi LQR, Kirschner M, *et al.* (2024): Multimodal neural correlates of childhood psychopathology. *Elife* 13. https://doi.org/10.7554/eLife.87992
104. Li Q, Zhao Y, Hu Y, Liu Y, Wang Y, Zhang Q, *et al.* (2024): Linked patterns of symptoms and cognitive covariation with functional brain controllability in major depressive disorder. *EBioMedicine* 106: 105255.
105. Lee J-E, Byeon K, Kim S, Park B-Y, Park H (2025): Revealing the multivariate associations between autistic traits and principal functional connectome. *Neuroinformatics* 23: 27.
106. Zhang K, Klumpp H, Jimmy J, Phan KL, Milad MR, Wen Z (2025): Functional connectivity predicting transdiagnostic treatment outcomes in internalizing psychopathologies. *JAMA Netw Open* 8: e2530008.
107. Venkatapathy S, Votinov M, Wagels L, Kim S, Lee M, Habel U, *et al.* (2023): Ensemble graph neural network model for classification of major depressive disorder using whole-brain functional connectivity. *Front Psychiatry* 14: 1125339.
108. Zhu M, Quan Y, He X (2023): The classification of brain network for major depressive disorder patients based on deep graph convolutional neural network. *Front Hum Neurosci* 17: 1094592.
109. Wen G, Cao P, Liu L, Yang J, Zhang X, Wang F, Zaiane OR (2023): Graph self-supervised learning with application to brain networks analysis. *IEEE J Biomed Health Inform* 27: 4154–4165.
110. Peng L, Wang N, Xu J, Zhu X, Li X (2022, March 16): GATE: Graph CCA for temporal SElf-supervised learning for label-efficient fMRI analysis. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/2203.09034
111. Wang X, Chu Y, Wang Q, Cao L, Qiao L, Zhang L, Liu M (2023): Unsupervised contrastive graph learning for resting-state functional MRI analysis and brain disorder detection. *Hum*

Brain Mapp 44: 5672–5692.
112. Zhu H, Tong X, Carlisle NB, Xie H, Keller CJ, Oathes DJ, *et al.* (2025): Contrastive functional connectivity defines neurophysiology-informed symptom dimensions in major depression. *Cell Rep Med* 6: 102151.
113. Xu Z, Chen CLP, Zhang T (2025): TFAGL: A novel agent graph learning method using time-frequency EEG for major depressive disorder detection. *IEEE Trans Affect Comput* 16: 1592–1605.
114. Li H, Srinivasan D, Zhuo C, Cui Z, Gur RE, Gur RC, *et al.* (2023): Computing personalized brain functional networks from fMRI using self-supervised deep learning. *Med Image Anal* 85: 102756.
115. Lin A, Wang W, Han H, Zhu F, Ma Q, Zheng Z, Zhou B (2025): KnowMDD: Knowledge-guided cross contrastive learning for major Depressive Disorder diagnosis. *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence* 7536–7544.
116. Zhang S, Chen X, Shen X, Ren B, Yu Z, Yang H, *et al.* (2023): A-GCL: Adversarial graph contrastive learning for fMRI analysis to diagnose neurodevelopmental disorders. *Med Image Anal* 90: 102932.
117. Wang X, Fang Y, Wang Q, Yap P-T, Zhu H, Liu M (2025): Self-supervised graph contrastive learning with diffusion augmentation for functional MRI analysis and brain disorder detection. *Med Image Anal* 101: 103403.
118. Kabir MS, Kurkin S, Portnova G, Martynova O, Wang Z, Hramov A (2024): Contrastive machine learning reveals in EEG resting-state network salient features specific to autism spectrum disorder. *Chaos Solitons Fractals* 185: 115123.
119. Yang Y, Ye C, Su G, Zhang Z, Chang Z, Chen H, *et al.* (2024, March 3): BrainMass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *arXiv [cs.CE]*. Retrieved July 7, 2025, from http://arxiv.org/abs/2403.01433
120. Wang C, Jiang Y, Peng Z, Li C, Bang C, Zhao L, *et al.* (2025, June 11): Towards a general-purpose foundation model for fMRI analysis. *arXiv [cs.CV]*. Retrieved from http://arxiv.org/abs/2506.11167
121. Yuan Z, Shen F, Li M, Yu Y, Tan C, Yang Y (2024, February 15): BrainWave: A brain signal foundation model for clinical applications. *arXiv [q-bio.NC]*. Retrieved from http://arxiv.org/abs/2402.10251
122. Wei X, Zhao K, Jiao Y, He L, Zhang Y (2025, August 3): A Brain Graph Foundation Model: Pre-training and prompt-tuning for any atlas and disorder. *arXiv [q-bio.NC]*. Retrieved from http://arxiv.org/abs/2506.02044
123. Wei Z, Dan T, Wu G (2025, October 21): Large connectome model: An fMRI foundation model of brain connectomes empowered by brain-environment interaction in multitask learning landscape. *arXiv [cs.LG]*. https://doi.org/10.48550/arXiv.2510.18910
124. Wei X, Zhao K, Jiao Y, Carlisle NB, Xie H, Fonzo GA, Zhang Y (2025): Multi-modal cross-domain self-supervised pre-training for fMRI and EEG fusion. *Neural Netw* 184: 107066.
125. Bi Y, Abrol A, Fu Z, Calhoun VD (2024): A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. *Hum Brain Mapp* 45: e26783.
126. Liu R, Huang Z-A, Hu Y, Zhu Z, Wong K-C, Tan KC (2024): Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI. *IEEE Trans Neural Netw Learn Syst* 35: 7627–7641.
127. Rahaman MA, Chen J, Fu Z, Lewis N, Iraji A, Calhoun VD (2021): Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness. *Annu Int Conf IEEE Eng Med Biol Soc* 2021: 3267–3272.
128. Wang G, Shi S, An S, Fan F, Ge W, Wang Q, *et al.* (2024): A bi-pyramid multimodal fusion method for the diagnosis of bipolar disorders. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1746–1750.
129. Rahaman MA, Chen J, Fu Z, Lewis N, Iraji A, van Erp TGM, Calhoun VD (2023): Deep

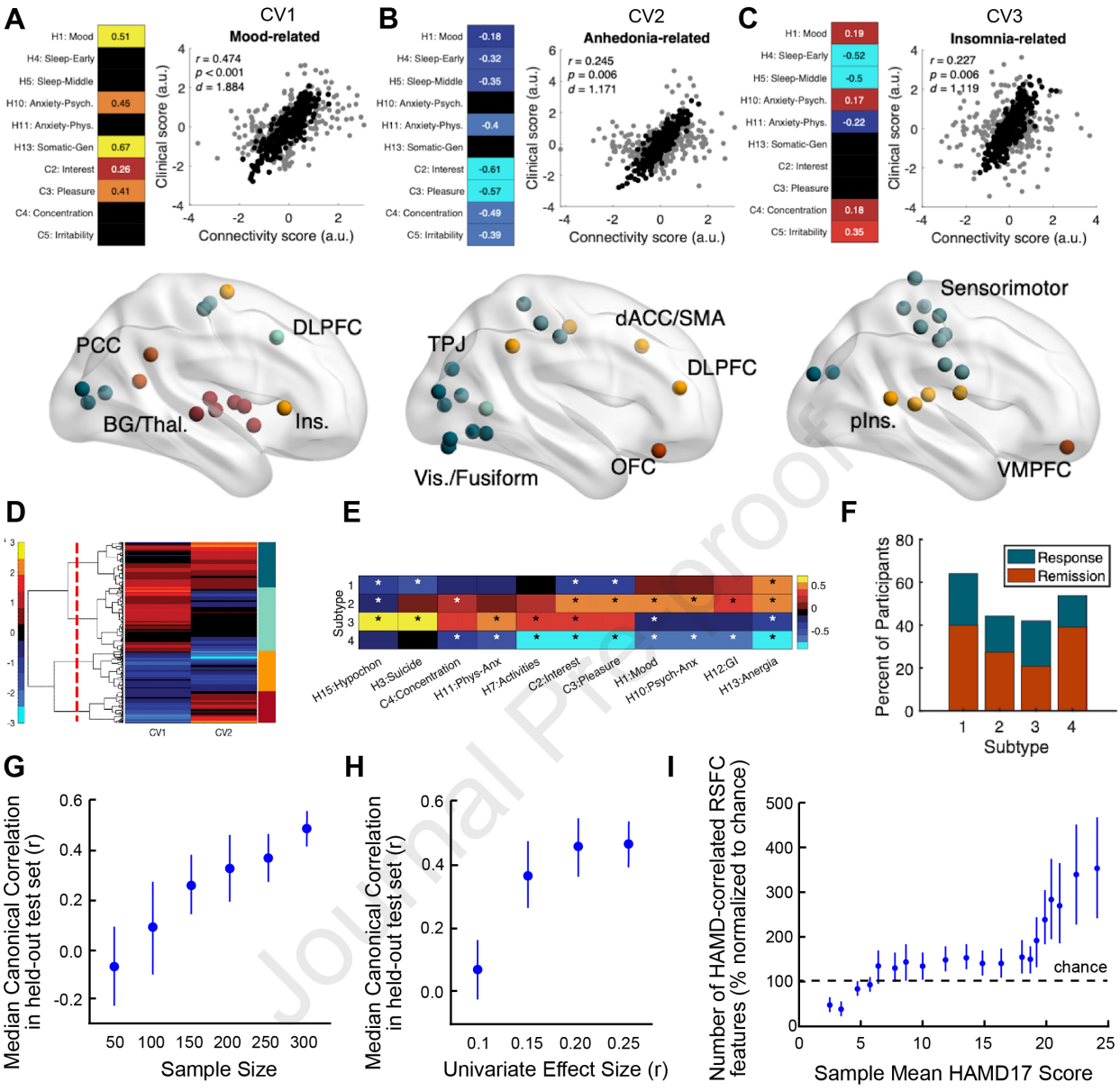
multimodal predictome for studying mental disorders. *Hum Brain Mapp* 44: 509–522.
130. Shaik NS, Cherukuri TK, Calhoun VD, Ye DH (2024, July 27): Multi-modal Imaging Genomics Transformer: Attentive integration of imaging with genomic biomarkers for schizophrenia classification. *arXiv [cs.CV]*. Retrieved from http://arxiv.org/abs/2407.19385
131. Poirot MG, Ruhe HG, Mutsaerts H-JMM, Maximov II, Groote IR, Bjørnerud A, *et al.* (2024): Treatment response prediction in major depressive disorder using multimodal MRI and clinical data: Secondary analysis of a randomized clinical trial. *Am J Psychiatry* 181: 223–233.
132. Pilmeyer J, Lamerichs R, Schielen S, Ramsaransing F, van Kranen-Mastenbroek V, Jansen JFA, *et al.* (2024): Multi-modal MRI for objective diagnosis and outcome prediction in depression. *NeuroImage Clin* 44: 103682.
133. Devlin J, Chang M-W, Lee K, Toutanova K (2018, October 11): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]*. Retrieved from http://arxiv.org/abs/1810.04805
134. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, *et al.* (2020, May 28): Language Models are Few-Shot Learners. *arXiv [cs.CL]*. Retrieved July 7, 2025, from http://arxiv.org/abs/2005.14165
135. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019): Language Models are Unsupervised Multitask Learners. Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
136. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, *et al.* (2022, April 5): PaLM: Scaling language modeling with Pathways. *arXiv [cs.CL]*. Retrieved July 7, 2025, from http://arxiv.org/abs/2204.02311
137. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, *et al.* (2023, May 17): PaLM 2 Technical Report. *arXiv [cs.CL]*. Retrieved July 7, 2025, from http://arxiv.org/abs/2305.10403
138. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, *et al.* (2023, March 15): GPT-4 Technical Report. *arXiv [cs.CL]*. Retrieved July 7, 2025, from http://arxiv.org/abs/2303.08774
139. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, *et al.* (2023, July 18): Llama 2: Open foundation and fine-tuned chat models. *arXiv [cs.CL]*. Retrieved from http://arxiv.org/abs/2307.09288
140. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, *et al.* (2022): A large language model for electronic health records. *NPJ Digit Med* 5: 194.
141. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, *et al.* (2023): Large language models encode clinical knowledge. *Nature* 620: 172–180.
142. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E (2021, October 29): MentalBERT: Publicly available pretrained language models for mental healthcare. *arXiv [cs.CL]*. Retrieved July 7, 2025, from http://arxiv.org/abs/2110.15621
143. Shi E, Zhao K, Yuan Q, Wang J, Hu H, Yu S, Zhang S (2024, September 19): FoME: A Foundation Model for EEG using adaptive temporal-lateral attention scaling. *arXiv [cs.LG]*. Retrieved from http://arxiv.org/abs/2409.12454
144. Cui W, Jeong W, Thölke P, Medani T, Jerbi K, Joshi AA, Leahy RM (2023, November 7): Neuro-GPT: Towards A foundation model for EEG. *arXiv [cs.LG]*. Retrieved July 6, 2025, from http://arxiv.org/abs/2311.03764
145. Dong Z, Li R, Wu Y, Nguyen TT, Chong JSX, Ji F, *et al.* (2024, September 28): Brain-JEPA: Brain dynamics foundation model with Gradient Positioning and Spatiotemporal Masking. *arXiv [q-bio.NC]*. Retrieved from http://arxiv.org/abs/2409.19407
146. Qian X, Wang Y, Huo J, Feng J, Fu Y (2023, November 1): fMRI-PTE: A Large-scale fMRI Pretrained Transformer Encoder for Multi-Subject Brain Activity Decoding. *arXiv [cs.CV]*. Retrieved July 7, 2025, from http://arxiv.org/abs/2311.00342

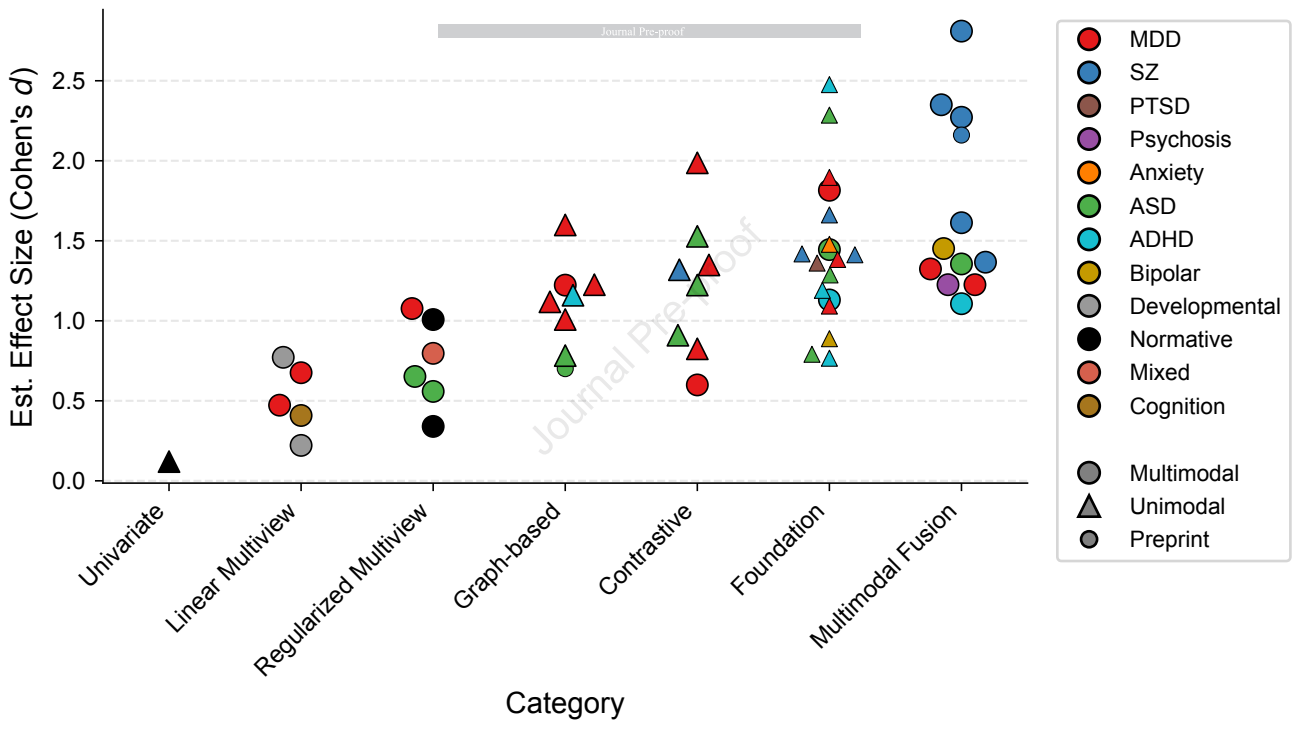
147. Wei F, Mo J, Zhang K, Shen H, Nagarajan S, Jiang F (2024, October 4): Nested Deep Learning model towards A foundation model for brain signal data. *arXiv [stat.ML]*. Retrieved July 7, 2025, from http://arxiv.org/abs/2410.03191
148. Wang EY, Fahey PG, Ding Z, Papadopoulos S, Ponder K, Weis MA, *et al.* (2025): Foundation model of neural activity predicts response to new stimulus types. *Nature* 640: 470–477.
149. Liu Y, Ma Y, Zhou W, Zhu G, Zheng N (2023, February 24): BrainCLIP: Bridging brain and visual-linguistic representation via CLIP for generic natural visual stimulus decoding. *arXiv [cs.CV]*. Retrieved from http://arxiv.org/abs/2302.12971
150. Zhang Y, Wang Y, Jimenez-Beneto D, Wang Z, Azabou M, Richards B, *et al.* (2024): Towards a “universal translator” for neural dynamics at single-cell, single-spike resolution ((A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang, editors)). *Neural Inf Process Syst* 37: 80495–80521.

Figure 1 | Parsing diagnostic heterogeneity in major depressive disorder. A-C. In our prior work (31)**,** regularized CCA revealed three generalizable brain-behavior dimensions explaining individual differences in **A)** mood & anxiety, **B)** anhedonia, and **C)** insomnia, respectively (N=328 subjects, r = 0.23–0.47 in test set). The scatterplots depict connectivity and clinical scores for each dimension (black dots = training data, gray dots = test data). Correlations are for held out test data. The clinical loadings of each dimension are depicted at the left of each scatterplot (black cells = not significant). The neuroanatomical distribution of connectivity loadings are depicted in the glass brains below each scatterplot. **D.** Hierarchical clustering on CV1 and CV2 revealed four MDD subtypes. The dendrogram visualizes the optimal k = 4 clustering solution. Heatmap values indicate participants' CV scores (arbitrary units). **E.** Clinical symptom profiles differed by subtype. The heatmap depicts Z-scored severity for each item. Asterisks represent significant differences (FDR-p < 0.05) relative to the severity of the entire sample. **F.** Differences in antidepressant response and remission to rTMS (response X^2=4.67, P=0.031; remission X^2=5.57, P=0.018). **G-I.** RCCA is sensitive to characteristics of the training sample. The median canonical correlation for CV1 increased with sample size (**G**) and with the strength of univariate correlations between connectivity features and clinical symptoms (**H**). RCCA also did not perform well in samples enriched for participants with mild or absent symptoms in (**I**). Here, the number of RSFC features that were nominally significantly correlated (P<0.001, uncorrected) with one or more HAMD items (normalized to the number expected by chance) increased with the mean HAMD17 score of the sample. For each data point, we selected a subset of subjects with a given mean HAMD17 from a dataset with N=485 subjects and repeatedly subsampled 80% of that subset without replacement to generate a mean +/– SD for the number of HAMD-correlated RSFC features. Adapted with permission from (31).

Figure 2 | Comparative effect sizes (estimated Cohen's d) for predictive neuroimaging models in psychiatry, grouped by analytic approach and modality. Effect sizes from published studies are shown as individual points, color-coded by clinical phenotype (e.g., major depressive disorder, schizophrenia, ASD, etc.; see legend) and grouped by analytic approach along the x-axis (univariate, linear multiview, regularized multiview, graph-based, contrastive, multimodal, and foundation models). Circles denote multimodal or multiview models; triangles indicate unimodal models. For each study, out-of-sample effect sizes are standardized to Cohen's *d*, regardless of whether the original metric was a correlation (r), accuracy (ACC), or area under the receiver-operating characteristic curve (AUC) (Note converting heterogeneous metrics to Cohen's d involves distributional assumptions that vary across study designs, so these standardized values are approximate; because studies differ in sample composition, prediction targets, evaluation protocol, etc., differences across method categories reflect these confounds as well as analytic approach and should be treated as hypothesis-generating not a benchmark; see Supplemental Methods). Marker size reflects sample size (not shown numerically). Citations by group: Univariate (1); Linear Multiview (1,14,15,103–106); Regularized Multivariate (16,19,31); Graph-based (64,69,71,107–110); Contrastive (111–118); Foundation (119–124); Multimodal Fusion (38,61,125–132). Key trend: Methods that integrate multiple data types, use nonlinear or deep architectures, or leverage large foundation models achieve the largest held-out effect sizes (especially in focused clinical samples) while classical, hand-engineered features yield lower predictive power. Abbreviations: ACC = accuracy; AUC = area under the receiver operating characteristic curve; d = Cohen's d; r = Pearson correlation coefficient; CCA = canonical correlation analysis; GNN = graph neural network; MDD = major depressive disorder; ASD = autism spectrum disorder; SZ = schizophrenia; PTSD = posttraumatic stress disorder.

Figure 3 | Scaling laws for text and brain foundation models: Model parameter count vs. effective training data ("tokens") for large-scale neural networks across language and neurobiological modalities. Each point represents a foundation or large neural network model, plotted by its number of trainable parameters (y-axis, billions) and normalized "effective tokens" (see Supplemental Methods for approach and caveats) or training examples (x-axis, billions, log–log scale). Color indicates the data modality (e.g., text, fMRI, EEG, iEEG, clinical text, calcium imaging, electrophysiology, or multimodal). Text-only models (BERT, GPT-3, PaLM, LLaMA, etc.;(133–142)) define the upper-right "frontier" of large language model scaling. The dashed "Chinchilla Optimal" line (computed following (78)) represents the empirical scaling law for text LLMs, balancing model size and data to maximize performance. Most brain foundation models (e.g., BrainLM, BrainWave, FoME, Neuro-GPT;(77,119,121,143–150)) cluster well below this line, indicating that, for a given size, they are likely under-trained relative to current language models. The largest multimodal brain models (BrainCLIP) remain smaller and trained on fewer examples than LLMs but are growing rapidly. Labels highlight key models across modalities. All brain models use the same axes for direct comparison. "Clinical text" models (Med-PaLM 2, GatorTron, MentalRoBERTa) are plotted after normalizing their effective token count to account for reduced text entropy in clinical notes. Models integrating EEG, fMRI, or calcium imaging (BrainLM, BrainWave, MICrONS) are shown in color. Brain foundation models are entering a scaling regime analogous to text LLMs but remain data- and compute-limited compared to language models. Closing the gap by increasing both parameter count and curated neurobiological training data may yield gains in psychiatric prediction and generalization, following empirical scaling laws. Abbreviations: BERT = Bidirectional Encoder Representations from Transformers; GPT = Generative Pretrained Transformer; fMRI = functional MRI; EEG = electroencephalography; iEEG = intracranial EEG; LLM = large language model; FoME = Foundation Model for EEG; MICrONS = Machine Intelligence from Cortical Networks.




Journal Pre-proof
Est. Effect Size (Cohen's d)
0.0
0.5
1.0
1.5
2.0
2.5
Univariate
Linear Multiview
Regularized Multiview
Graph-based
Contrastive
Foundation
Multimodal Fusion
Category
MDD
SZ
PTSD
Psychosis
Anxiety
ASD
ADHD
Bipolar
Developmental
Normative
Mixed
Cognition
Multimodal
Unimodal
Preprint
Journal Pre-proof