Reviewed Preprint
v1 • February 14, 2025
Not revised

Reviewed Preprint
v2 • June 10, 2026
Revised by authors

Brain-Cognitive Gaps in relation to Dopamine and Health-related Factors: Insights from AI-Driven Functional Connectome Predictions

Morteza Esmaeili[1,2] ✉, **Erin Bjørkeli**[2,3], **Robin Pedersen**[4,5,6], **Farshad Falahati**[4,5,6], **Jarkko Johansson**[4,8], **Kristin Nordin**[6], **Nina Karalija**[7,8], **Lars Bäckman**[6], **Lars Nyberg**[5,7,8], **Alireza Salami**[4,5,6,7,9]

[1]Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway • [2]Department of Diagnostic Imaging, Akershus University Hospital, Lørenskog, Norway • [3]Institute of Clinical Medicine, University of Oslo, Oslo, Norway • [4]Wallenberg Centre for Molecular Medicine (WCMM), Umeå University, Umeå, Sweden • [5]Department of Medical and Translational Biology, Umeå University, Umeå, Sweden • [6]Aging Research Center, Karolinska Institute and Stockholm University, Solna, Sweden • [7]Umeå Center for Functional Brain Imaging (UFBI), Umeå University, Umeå, Sweden • [8]Department of Diagnostics and Intervention, Diagnostic Radiology, Umeå University, Umeå, Sweden • [9]Department of Psychology, Florida State University, Tallahassee, United States

✉ **For correspondence:** Morteza.Esmaeili@uis.no

Competing interests: No competing interests declared

Funding: See page 24

Reviewing editor: Guido van Wingen, Amsterdam UMC Location University of Amsterdam, Netherlands

eLife Assessment

This multimodal neuroimaging study leverages fMRI, PET, and deep learning to predict memory performance. The authors introduce the brain-cognition gap to link these different imaging modalities to cognition and evaluate their results in two independent cohorts. The results are **solid** and provide an **important** contribution to the literature and will be of interest to neuroscientists working at the interface of cognition, neuroimaging.

https://doi.org/10.7554/eLife.104053.2.sa3

Abstract

A key question in human neuroscience is to understand how individual differences in brain function relate to cognitive differences. However, the optimal condition of brain function to study between-person differences in cognition remains unclear. While many studies have developed objective biomarkers to accurately predict intelligence and general cognition, consensus on domain-specific markers has not yet emerged. Brain age has been proposed as a potential candidate, but recent research suggests that brain age offers minimal additional information on cognitive decline beyond what chronological age provides, prompting a shift toward approaches focused directly on cognitive prediction. Using a deep learning approach, we evaluated the predictive power of the functional connectome during various states (resting state, movie-watching, and n-back) on episodic memory and working memory performance. Our findings show that while connectomes during task, especially during movie watching, better predict both episodic and working memory, resting state connectomes are equally effective in predicting episodic memory. Furthermore, individuals with a negative brain-cognition gap (where brain predictions underestimate actual performance) exhibited lower physical activity and higher cardiovascular risk compared to those with a positive gap. This shows that knowledge of the brain-cognition gap provides insights into factors contributing to cognitive resilience. Further lower PET-derived measures of dopamine binding were linked to a greater brain-cognition gap, mediated by regional functional variability. Together, our findings highlight the importance of brain state in connectome-based cognitive prediction and introduce the brain cognitive gap as a potentially informative, dopamine-modulated marker of vulnerability to compromise brain function.

1. Introduction

Resting state functional magnetic resonance imaging (fMRI) serves as a tool to map brain function (1, 2). Functional connectivity (FC) at rest, estimated with methods to gauge correlated spontaneous activity between two or more regions, serves as a measure of functional brain integrity (1, 3). Individual differences in FC at rest are associated with differences in various cognitive domains (4–6). Associations of this nature have been reported for several large-scale networks, including the default mode network (DMN) (7–9), the frontoparietal network (10–12), the dorsal attention network (13), and the salience network (14). Moreover, such associations have also been identified in more comprehensive investigations spanning the entire functional brain repertoire of the brain (15, 16). Most previous reports focused on associations, not predictions, utilizing the correlation between FC and behavioral phenotypes, which tend to overfit the data and therefore fail to generalize (17). Proper cross-validation, preferably in an independent sample, is therefore important to assert reliable population-level inferences (18). Recent machine learning-based predictive frameworks offer powerful tools for assessing the predictability of individual behavioral phenotypes based on brain connectivity (19–23). In particular, deep neural networks (DNN) methods have been successfully applied to behavioral and disease prediction (24–26), and were initially expected to outperform other machine learning approaches (27–29). However, this superiority remains debatable, as recent studies have reported comparable performance between DNNs and traditional methods (29, 30). Accordingly, the present study does not aim to benchmark deep learning against traditional machine learning approaches but instead uses a consistent predictive framework to examine how brain state influences the utility of FC for cognitive prediction.

The functional connectome has demonstrated predictive utility regarding trait-like cognitive phenotypes (31–34). The predictive-modeling framework of the functional connectome has been applied to various cognitive domains, including intelligence (35, 36), working memory (WM) (37), visuospatial ability (38), attention (39), creativity (40), as well as personality traits (41). Understanding the patterns contributing to predictions could offer insights into the functional organization underlying cognitive phenotypes, serving as biomarkers indicating current or prospective health conditions (42–44). Moreover, the whole-brain functional connectome acts as a fingerprint with accurate identification of subjects from a large population (45) within the same cognitive state (e.g., rest-to-rest) but also across different states (e.g., rest-to-task). Overall, past research suggests that the functional connectome is relatively robust within individuals, is unique across individuals, and can predict cognitive and personality phenotypes. However, less is known about how the predictive utility of the functional connectome depends on the brain state during which FC is measured.

Despite consensus on the value of resting state functional connectivity for mapping brain function, there is an ongoing debate about whether rest is the optimal brain state for investigating individual differences in neurocognitive function (46–49). A study using data from the Human Connectome Project has shown that resting state fMRI predicts differences in brain activity during various tasks, including social, language, relational, and motor tasks (50). This finding supports the notion that individual differences in neural activation can be predicted from resting state (48). However, results from the same dataset revealed that FC during task outperforms resting state FC in predicting individual differences in fluid intelligence, with FC during task explaining 20 % of the variance compared to just 6 % explained by rest (51). Consistent with this result, previous studies have investigated trait-as well as state-dependent FC, supporting the utility of an integrative approach (11, 49). More recent studies suggested that naturalistic viewing, such as movie-watching, may serve as a happy medium between unconstrained rest and overly-constrained tasks in predicting behavior differences (52, 53). Despite the presence of similar spatiotemporal activity patterns across individuals during movie-watching (54), notable individual differences in activity and functional connectivity (55) persist alongside these idiosyncratic features. This suggests that tasks which align individuals' functional connectome more closely to

an optimal level, neither completely unconstrained like rest nor overly synchronized like a task, also render them easiest to identify (46). Hence, it is plausible that FC during naturalistic paradigms improves sensitivity to predict behavioral differences (52, 56).

The primary objective of this study is to determine how brain state influences the predictive utility of the functional connectome for cognitive performance, using a deep learning framework. Specially, we test whether functional connectivity derived from different brain states differentially predicts WM and episodic memory (EM), two cores but functionally distinct cognitive domains. WM reflects the capacity to temporarily store and manipulate information and supports higher-order problem-solving, reasoning, and other key components of fluid intelligence (e.g., (57–59)), whereas, EM entails the recollection of specific experiences and events (60), which is regarded as an important element of mind-wandering (61) during resting state. We examine whether these domains are differentially predicted by connectomes derived from resting state, movie watching, and n-back task fMRI.

Past studies have used neuroimaging data, including resting state to predict brain age (62–64). These studies show that brain age, based on biological phenotypes, and their deviation from the chronological age (known as brain age gap prediction), could serve as a biomarker in characterizing disease risk (64–66). Importantly, an older-appearing brain has been shown to exacerbate physiological and cognitive aging, and risk of mortality (63). A recent study demonstrated that while brain age can predict chronological age with high accuracy from MRI, its utility for predicting cognition is limited (67). Specifically, Tetereva and colleagues (2024) (67) showed that brain age strongly tracks chronological age and that, to predict cognition, brain age largely overlapped with chronological age, such that controlling for chronological age eliminated the predictive contribution of brain age. This finding suggests that brain-age models may provide little unique explanatory power for cognitive decline beyond what is already captured by chronological age. Building on this observation and extending the concept of a brain age gap to a brain-cognition gap (BCG, defined as the discrepancy between predicted and observed cognitive performance), we propose that BCG may serve as an informative marker of individual differences. If the brain predicts lower performance than is observed (i.e., a negative BCG), it may be compensating for underlying issues not yet apparent through cognitive assessments. By this view, individuals with negative BCG should be less healthy than those whose brains predict higher cognitive function than their actual performance (i.e., a positive BCG). Our second aim was to extend the concept of brain-age prediction to cognition by introducing BCG. Considering the significance of lifestyle and cardiovascular risk for maintaining healthy brain function (68, 69), we assess whether BCG captures individual differences beyond chronological age and examine whether individuals with positive versus negative BCG differ in lifestyle factors and cardiovascular risk, which are known contributors to brain health.

The third aim of the current study is to investigate the neurobiological underpinnings of BCG by examining the role of dopaminergic (DA) integrity. We test the hypothesis that lower DA receptor availability is associated with increased blood-oxygen-level-dependent (BOLD) signal variability, reduced functional connectome uniqueness, and larger BCG, consistent with DA's role in modulating neural signal-to-noise ratio (SNR) and network coherence. DA is a vital neuromodulator with critical implications for motor function, reward-seeking behavior, and various higher-order cognitive functions (70–77). Insufficient DA modulation can affect neurocognitive functions detrimentally (71, 76, 78–80). (81, 82) (83, 84) (85). Pharmacological studies have shown that DA depletion increases the variability of the BOLD signal, subsequently leading to less synchronized connectivity within resting state networks (86). Consequently, we expect individuals with inadequate DA levels to exhibit increased regional signal variability, a less unique functional connectome, and greater BCG.

Using data from the DopamiNe, Age, connectoMe, and Cognition (DyNAMiC) study (n =180, 20-79 years, 50% female) (56), we evaluated the predictive power of the functional connectome during resting state, movie-watching, and n-back tasks for two different cognitive domains: episodic memory and working memory. Based on recent research indicating that movie-watching enhances predictability by highlighting key features of FC (52), we hypothesized that FC during movie-

watching would outperform FC during rest, and possibly during task, in predicting both cognitive measures. To achieve this objective, we employed a deep neural network approach, a specific subtype of artificial intelligence (AI), to predict cognitive scores from the functional connectome. Deep learning approaches offer a flexible modeling framework capable of capturing complex linear and non-linear associations in high-dimensional data (30) and have been shown to reliably predict intelligence (23, 87). Considering the importance of individual characteristics, such as age, in predicting behavior from FC (34), we conducted external validation of our model, initially derived from an age-heterogeneous sample, in an age-homogeneous sample (from the Cognition, Brain, and Aging (COBRA) study (88)). We subsequently investigated whether individuals with positive brain-cognition prediction gaps differ from those with negative gaps in terms of lifestyle and cardiovascular disease risk factors. Moreover, we tested the hypothesis of whether individuals with lower striatal dopamine D1-like receptor availability (D1DR), the brain's most abundant DA receptor subtype, have a less distinctive FC pattern (i.e., more regional variability) and, in turn, a larger BCG. Finally, we conducted an external validation of the link between DA and the prediction gap in an independent cohort with estimates of DA D2-like receptor (D2DR) availability (88).

2. Results

2.1. AI-Driven Predictive Modeling of Cognition Scores from the Functional Connectome

We used fMRI data from the DyNAMiC project, in which each subject underwent scanning during rest, movie-watching, and working memory n-back tasks. These data were parcellated into 273 nodes (264 with 9 additional subcortical nodes) using a previously published whole-brain functional atlas (89). The averaged time series of 273 regions were subsequently correlated to create the FC matrix for each participant and cognitive state (rest, movie-watching, and n-back).

We trained a convolutional neural network, DenselyAttention, derived from DenseNet (90) on FC matrices from each condition (resting state, movie-watching, and n-back) to generate cognition-specific prediction models of two memory domains (EM and WM). Model performance was quantified as the correlation between observed and predicted cognitive performance. Each model was then tested on all three conditions to examine the generalizability of each model across cognitive states. For example, the model trained on the resting state data (orange circles in Fig. 1) was used to predict EM scores using the test dataset derived from resting state (Fig. 1a), movie-watching, and n-back. Significance of our main predictions was assessed via linear correlations, and uncorrected p-values are presented in Tables 1-2.

2.2. Resting state and movie-watching models outperform n-back in episodic-memory prediction, with resting state offering the best generalizability

We first started with all cases in which congruent conditions were used for model building and prediction. Only models derived from the rest and movie-watching datasets yielded significant predictions of EM (Figs. 1a-b and Table 1), with resting state yielding the best-performing model (r =0.50, p <0.0001), followed by movie-watching (r =0.49, p <0.0001) (Table 1). While there was no significant difference between resting state and movie watching in predicting episodic memory (Δr =0.071, with a 95% confidence interval of [-0.097, 0.261], Fig. 1e), both models yielded a markedly better EM prediction than n-back (rest vs. n-back: Δr =0.333, with a 95% confidence interval of [0.054, 0.572]; movie vs. n-back: Δr =0.316, with a 95% confidence interval of [0.015, 0.619], Fig. 1e). Thus, the two models outperformed the n-back model (p <0.05, bootstrap test), indicating a significant improvement. To test the generalizability of these models, two types of validation analyses were performed: cross-condition and cross-data set. In the cross-condition analysis, models trained on one condition (e.g., rest) were tested on an incongruent condition (e.g., movie-watching, n-back; Table 1). Interestingly, the model trained on resting state significantly predicted EM when tested on movie-watching (r =0.44, p <0.001) and n-back (r =0.38, p =0.003)

Figure 1.

Model trained on functional connectivity maps acquired at rest predicts (**a**) episodic memory (EM) of the test dataset. The models trained on movie-watching (**b**) but not n-back (**c**) datasets predicted EM scores of the test dataset. Table 1 summarizes the p values, correlations, mean square error (MSE), and mean absolute error (MAE) for each model. Test datasets were obtained from the same cohort for rest, movie-watching, and n-back. The winning model trained on EM at rest was evaluated on the external COBRA cohort and yielded a significant prediction of EM scores (**d**). Bootstrap distributions of correlations between predicted and actual EM scores indicated no significant difference in the predictive power of EM between models trained on resting state and movie-watching data (**e**). Additionally, the bootstrap distribution revealed that models trained on resting state and movie-watching data yielded higher correlations than those trained on n-back data (**e**). Visualization of features contributing to the successful prediction of EM at rest (**f**). A grad-CAM-derived saliency map displays the features that contributed to the model's predictions. The hot spots overlaid on the FC map demonstrate noticeable cross-correlation contributions in "default mode" (DMN) regions. Another important feature visualized by Grad-CAM includes off-diagonal hot spots reflecting inter-connections of the DMN – "subcortical" node.

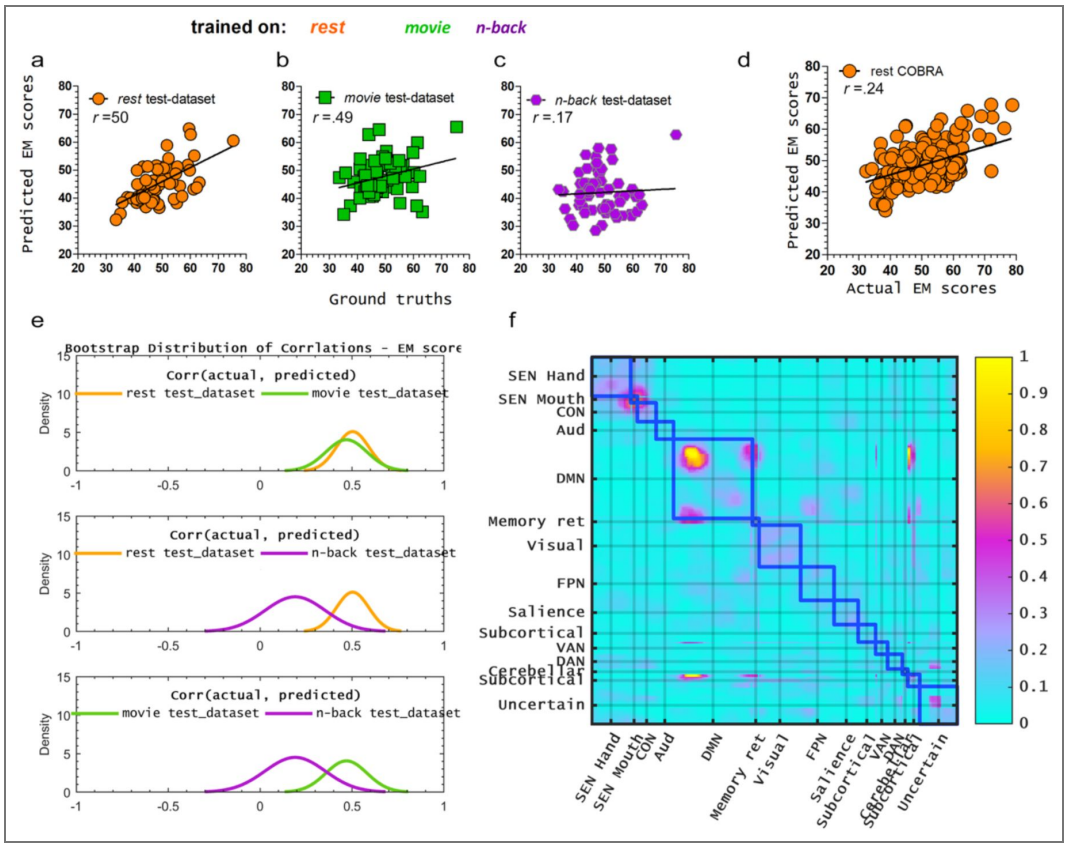


Table 1. Correlation results for episodic memory (EM) score predictions

Model trained on Model tested on	rest			movie			n-back		
	rest	movie	n-back	rest	movie	n-back	rest	movie	n-back
Correlation (r)	**0.50**	**0.44**	**0.38**	**0.28**	**0.49**	**0.09**	0.05	0.08	0.17
Correlation (r^2)	**0.128**	**0.054**	**0.043**	0.03	**0.118**	0.002	-0.01	0.001	0.015
Significance	**<.0001**	**0.0004**	**0.003**	**0.02**	**<.0001**	**0.49**	0.76	0.84	0.42
MSE	4.35	81.93	40.74	129.44	10.83	156.42	133.87	127.57	20.16
MAE	1.59	3.87	4.92	11.88	2.57	8.75	9.33	8.28	3.42
COBRA									
Correlation (r)	**.24**								
Correlation (r^2)	**0.053**								
Significance	**<.0001**								
MSE	37.62								
MAE	8.28								

Model trained on	rest			movie			n-back		
Model tested on	rest	movie	n-back	rest	movie	n-back	rest	movie	n-back
Correlation (r)	0.02	0.08	0.12	**0.42**	**0.57**	**0.46**	0.20	**0.46**	**0.47**
Correlation (r^2)	-0.09	-0.05	-0.03	**0.105**	**0.170**	**0.093**	0.0154	**0.046**	**0.141**
Significance	0.88	0.72	0.15	**<.0001**	**<.0001**	**0.004**	0.12	**0.0002**	**<0.0001**
MSE	85.67	214.81	169.02	70.57	49.19	98.36	127.72	116.15	61.23
MAE	7.47	9.74	9.37	6.67	5.70	8.10	8.89	8.71	6.31
COBRA									
Correlation (r)						**0.47**			
Correlation (r^2)						**0.102**			
Significance						**<.0001**			
						94.87			

Table 2. **Correlation results for working memory (WM) score predictions.**

conditions. In contrast, models trained on movie-watching or n-back could not be generalized to other conditions, unable to significantly predict EM (*p*'s >.1), except for significant generalizability from the movie to the rest condition (Table 1).

In a cross-dataset validation analysis, the best-performing model from the age-heterogeneous DyNAMiC dataset was tested on the corresponding condition in an age-homogeneous cohort from the COBRA dataset. By doing this, we found that the resting state model derived from DyNAMiC significantly predicted EM performance in the COBRA dataset (r =0.24, p <0.0001).

Next, we aimed to delineate the relative contributions of different brain regions for the best-performing model, the model trained on the "resting state data" in predicting episodic memory. Utilizing the Grad-CAM algorithm, saliency maps were generated for the 120 FC matrices used during training of the winning model. An averaged and interpolated description of all saliency maps is depicted in Figure 1f. The saliency map highlights specific edges, especially within the default-mode network, edges between the default-mode network and subcortical areas, and edges between the default-mode and the cerebellar network. These edges, indicated by a salience intensity of ≥0.5, exert a significant influence on the model (Fig. 1f).

2.3. Movie-watching and n-back models outperform resting state in working-memory predictions, with movie-watching offering the best generalizability

We next investigated whether the superiority of resting state in predicting EM was unique to this domain, considering that previous research demonstrated the advantages of task-based fMRI and naturalistic viewing in predicting fluid intelligence (91). To do so, we compared the predictive power of different states for WM, which is shown to be more directly associated with fluid intelligence compared to EM.

The model derived from the resting state failed to predict and generalize regarding WM (*p*'s >0.10; Fig. 2a and Table 2). By contrast, models trained on movie-watching and n-back yielded significant predictions of WM (Figs. 2b-c and Table 2), with the movie-watching model emerging as the best-performing model (r =0.57, p <0.0001) followed by n-back (r^2 =0.47, p <0.0001). While there was no significant difference between movie-watching and n-back in predicting WM (Δr =0.026, with a 95% confidence interval of [0.001, 0.052], Fig. 2e), these models yielded better WM prediction than resting state (Δr =0.517, with a 95% confidence interval [0.373, 0.662], Fig. 2e).

In cross-condition validation (Table 2), the movie-watching model yielded a significant WM prediction during both resting state (r =0.42, p <0.0001) and n-back (r =0.46, p =0.004). The model derived from n-back yielded a significant WM prediction during movie-watching (r =0.46, p =0.0002), but not during resting state (r =0.20, p =0.12).

For cross-dataset validation, the best-performing model in the DyNAMiC dataset (i.e., movie-watching) was applied to predict WM in the COBRA dataset. However, since COBRA does not include a movie-watching paradigm, we applied the model to the n-back task in COBRA. This approach revealed that the model from the DyNAMiC movie-watching condition yielded a significant WM prediction in the COBRA n-back task (r =0.47, p <0.0001).

Overall, our results suggest that movie-watching-based WM predictions generalize across different cognitive states and datasets. This finding could be further replicated using a different functional parcellation (Figs. S1-S2 and Tables S1-S2).

The Grad-CAM algorithm generated saliency maps of 120 FC maps from the DyNAMiC dataset employed for training. Figure 2f depicts an average of all Grad-CAM generated maps. The saliency map unveils that certain edges, specifically within network connectivity of task-positive regions such as the frontoparietal task control network, dorsal/ventral attention, visual, and subcortical networks, as well as between-network connectivity. FC between task-positive dorsal and ventral attention networks, and between the DMN and the fronto-parietal network,

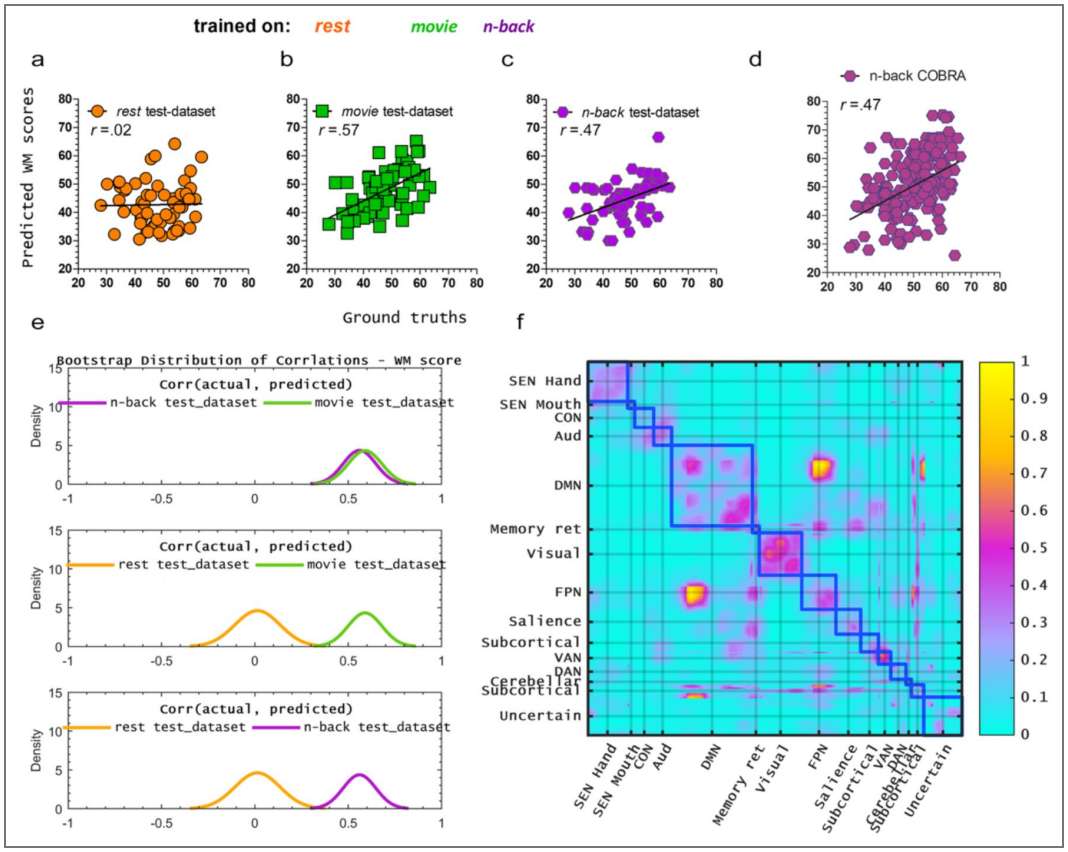


Figure 2.

Model trained on FC maps acquired at rest did not significantly predict (**a**) the working memory (WM) in the test dataset. The model trained on the movie-watching dataset yielded the best-performing model in predicting WM (**b**) while the model trained on the n-back dataset (**c**) was the second-best model. Table 2 summarizes the p-values, correlation power, MSE, and MAE for each model. (**d**) Results of cross-dataset validation, where the best-performing model in the DyNAMiC dataset (i.e., movie-watching) was applied to predict WM to the COBRA dataset. However, since COBRA does not include a movie-watching paradigm, we applied the model to the n-back task in COBRA (Table 2). Bootstrap distributions of correlations between predicted and actual WM scores showed no significant difference in predictive power between models trained on movie-watching and n-back data (**e**). The bootstrap distribution revealed that models trained on movie-watching and n-back data exhibited higher correlations than those trained on resting state data (**e**). The Grad-CAM-derived saliency map highlights dominant features in the FC maps that contributed to the model's predictions (**f**). The hot spots overlaid on an FC map demonstrate noticeable cross-correlation contributions in the "VAN", "visual", and to a lesser degree (<0.5) "DMN". Other important features visualized by Grad-CAM include off-diagonal hot spots reflecting inter-connections of the "DMN" – with "FPN, Fronto-parietal Task Control", "Subcortical", and "Cerebellar"; "Cerebellar" – "FPN" node.

contributed to the best-performing model derived from the movie-watching dataset. Applying two different parcellation methods (89, 92) to the DyNAMiC data indicated that parcellation resolution does not significantly impact model performance (see Figs. S1-S2 and Tables S1-S2).

2.4. The brain-cognition gap is related to physical activity levels and cardiovascular risk factors

Given the importance of lifestyle and cardiovascular health for maintaining healthy brain function (68, 69, 93), we examined whether individuals with positive versus negative prediction gaps differed in physical activity habits, education, and Framingham cardiovascular disease (CVD) risk score. Our primary focus was on EM predictions derived from resting state data, as this was the common condition across the DyNAMiC and COBRA datasets. We computed the difference between predicted and observed EM scores to generate BCG. A positive BCG indicates that an individual's brain predicted better-than-observed EM performance, whereas a negative BCG indicates more compromised brains relative to actual performance.

In the test sample of the DyNAMiC data (n =60) and the entire COBRA sample (n =177), we found that individuals with a negative BCG exhibited lower levels of physical activity and higher CVD risk scores compared to those with positive gaps (Fig. 3). Confirmatory analysis with continuous variables revealed positive relationships between GAP and physical activity (DyNAMiC: r (57) =0.40, p =0.001; COBRA: r(166) =0.17, p=0.03) and negative relationships between GAP and CVD risk score (DyNAMiC: r(57) =–0.27, p =0.03; COBRA: r(172)= –0.10, p =0.40). Moreover, individuals with negative BCG were less educated compared to those with positive BCG in the DyNAMiC dataset.

To test whether cognition on its own is related to physical activity and CVD score, we conducted a median split on EM and compared physical activity and cardiovascular risk score across the two groups. In contrast to the findings related to BCG, we found no significant difference in the level of physical activity (t(58) =-0.59, p =0.56) or cardiovascular risk score (t(58) =1.64, p =0.11) between high and low EM performers (Fig. 3). These results suggest that BCG may provide additional information, beyond cognitive measures, regarding factors that contribute to cognitive resilience.

2.5. Dopamine D1 and D2 receptor availability are associated with brain-cognition gaps

Given that BCG may partly reflect variability in neural signal, one plausible neurobiological factor contributing to BCG is dopaminergic integrity. We hypothesized that inadequate DA levels might be related to increased neural signal-to-noise ratio, thereby resulting in a less unique functional connectome, consequently leading to a greater prediction gap. We therefore initially investigated the relationship between DA receptor levels and predictive gaps across different types of DA receptors in the DyNAMiC and COBRA samples.

In the DyNAMiC sample, we found a significant correlation between striatal D1DR and prediction BCG (in those with a positive gap: r = –0.49, p =0.03; negative gap: r =0.40, p =0.01) (Fig. 4a), suggesting that lower D1DR is associated with greater BCG.

We replicated our finding in the COBRA sample using D2-like receptor availability (D2DR), revealing a significant relationship between striatal D2DR and prediction gap (in those with a positive gap: r = –0.49, p =0.001; negative gap: r =0.39, p =0.004) (Fig. 4b). Our findings provide support the view that lower D1DR/D2DR is associated with larger brain-cognition prediction gaps.

Both D1DR and D2DR availability in the striatum were associated with BCG, such that lower dopamine receptor availability was linked to a greater BCG. However, these associations varied by region. For D1DR, significant correlations with BCG were observed in the caudate (positive gap: r = –0.37, p =0.02; negative gap: r = 0.37, p =0.02) and putamen (positive gap: r = – 0.53, p =0.02; negative gap: r =0.34, p =0.03), but not in the nucleus accumbens (positive gap: r = –0.25, p = 0.31; negative gap: r =0.07, p =0.69) or the DLPFC (positive gap: r = – 0.30, p =0.21; negative gap: r =0.21, p

Figure 3.

The plots compare physical activity scores (total hours per week), CVD risk, between two groups with positive and negative BAG as well as high and low memory performance in the DyNAMiC (**top row**) and COBRA (**bottom row**) datasets. *, **, and *** denote p <0.05, p <0.01, and p <0.001 respectively.

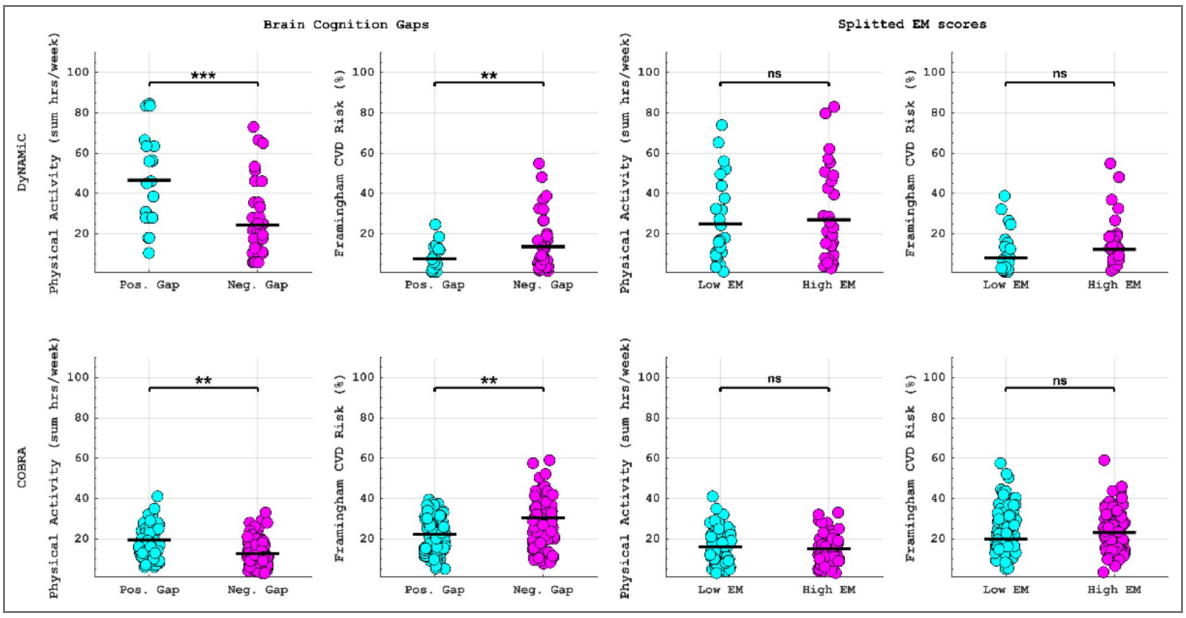
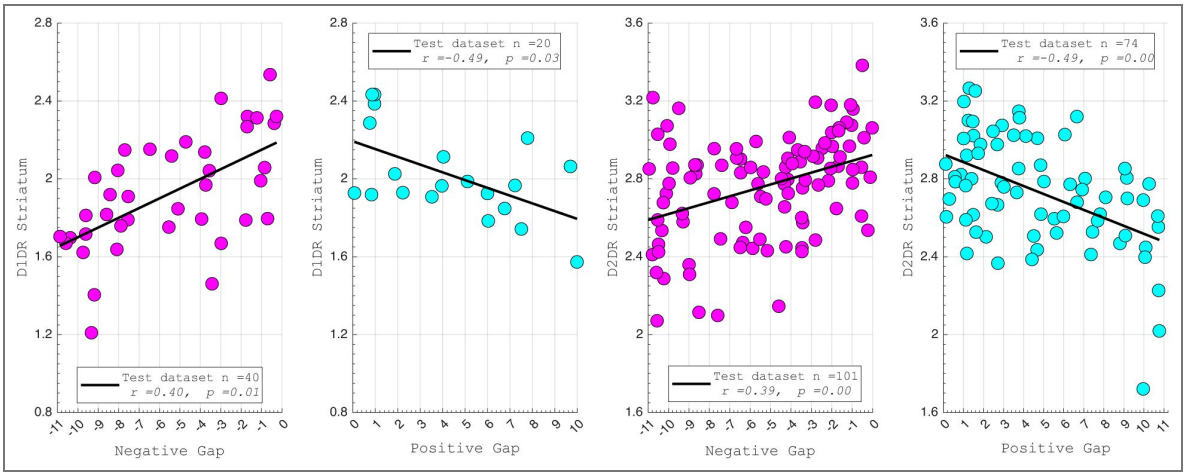


Figure 4.

Relationship between the gap measured from predicted and actual EM scores and dopamine D1 receptor (**a**). Negative gaps indicate that the predicted EM score was lower than actual EM scores, while the positive gaps indicate more higher predicted scores than actual EM scores. Partial correlation analysis showed a significant correlation between D1 receptor values and the measured negative and positive gaps. Relationship between the gap measured from predicted and actual EM scores and dopamine D2 receptor (**b**). Negative gaps indicate that the predicted EM score was lower than actual EM scores, while positive gaps indicate more higher predicted scores than actual EM scores. Partial correlation analysis showed a significant link of D2 receptor values to negative gaps and positive gaps.



=0.21). For D2DR, both caudate (positive gap: r = – 0.34, p =0.004; negative gap: r =0.36, p =0.0003) and putamen (positive gap: r = – 0.37, p =0.002; negative gap: r =0.22, p =0.03) showed significant associations with BCG.

2.6. Regional variability mediates the direct impact of dopamine on brain-cognition gaps

We showed that dopamine is associated with BCG. To evaluate whether functional variability mediated this relationship, we conducted additional mediation analyses. We computed BOLD signal entropy, which estimates within-region signal variability during resting state (86), that was averaged across striatal regions (left caudate: MNI coordinate ≈ [–12, 12, 6], right caudate: MNI coordinate ≈ [10, 14, 2], left putamen: MNI coordinates ≈ [–24, 6, 4], right putamen: MNI coordinates ≈ [29, 1, 4]), as we expected that reduced DA may primarily impact the local functional dynamics.

In the DyNAMiC sample, within the group exhibiting a negative BCG, we observed a negative association between striatal D1DR and entropy (r =–0.33, p =0.04) as well as a negative association between entropy and BCG (r =–0.36, p =0.03). Importantly, we observed a significant indirect effect of D1DR on the gap mediated by entropy, β =2.41, 95% CI [0.89, 4.51], p <0.0001, explaining 56.81% of the total effect of D1DR on the gap. The direct effect of D1DR, however, was not significant, β =1.83, 95% CI [–0.86, 3.79], p =0.19.

Similarly, in the group with a positive BCG, entropy was negatively associated with D1DR (r =–0.56, p =0.01) and positively associated with BCG (r =0.47, p =0.04). Importantly, an indirect effect of D1DR on BCG through entropy was observed, β =–6.78, 95% CI [–11.26, –3.05], p <.0001, accounting for 89.41% of the D1DR effect on the gap. The direct effect of D1DR was again non-significant, β =–0.803, 95% CI [–0.48, 2.27], p =0.26. These findings suggest that lower D1DR levels contribute to increased signal variability, which in turn leads to reduced specificity of FC and, consequently, a larger BCG.

In the COBRA sample, within the group with a negative BCG, we observed a negative association between striatal D2DR and entropy (r = –0.22, p =0.03) as well as a negative association between entropy and BCG (r = –0.27, p =0.007). In the group with the positive BCG, entropy was negatively associated with D2DR (r = –0.26, p =0.03) and positively associated with BCG (r =0.25, p =0.03). Moreover, we detected a significant indirect effect of D2DR on both the negative and positive gap groups through entropy. For the negative gap, β =2.18, 95% CI [0.01, 4.25], p =0.04, accounting for 63.43% of the D2DR effect on the gap; and for the positive gap, β =–2.18, 95% CI [–3.87, –0.04], p =0.04, accounting for 61.49% of the D2DR effect on the gap. Similar to the results reported for D1DR, these findings suggest that lower D2DR levels contribute to increased signal variability, which in turn may lead to reduced specificity of FC and, consequently, a larger BCG.

3. Discussion

Using deep learning models, we examined the predictive power of the functional connectome during various states (resting state, movie-watching, and n-back) on two different cognitive domains (EM and WM). Both rest and movie-watching states yielded significant predictions of EM, with the model derived from resting state generalized across states and datasets. Differences between the DyNAMiC and COBRA datasets make cross-dataset prediction a harder problem, as the age ranges of samples significantly vary, and prior studies highlight the importance of individual characteristics like age in predicting behavior from FC (34). In line with this, model performance decreased when predicting EM in the COBRA sample, whereas prediction of WM remained largely unchanged. Thus, validation outcomes suggest that the models, particularly those predicting WM, show robustness across datasets, whereas the reduced EM performance highlights potential data-specific influences that limit generalizability. The saliency map generated from the final layer of the deep learning model indicates that certain edges within DMN, and between DMN and the subcortical network contributed significantly to the prediction. Building on a recent finding by Kurkela and Ritchy (94), our finding reveals that a portion of the known-

memory subnetwork within the DMN, as well as a whole-brain multivariate pattern which notably encompasses interactions of the DMN with other networks, such as the subcortical network, made a more substantial contribution to prediction. Importantly, this prediction generalizes across conditions and datasets, suggesting that features derived from resting state FC serve as a relatively stable marker of individual differences in EM, though with reduced strength in COBRA. While such generalization is partly facilitated by the similarity of functional connectivity across states, it is not a trivial outcome. For instance, the model trained on movie-watching data generalized to EM prediction during rest but failed to do so for the n-back condition, even though movie-watching and n-back connectivity patterns are themselves highly correlated. This indicates that successful generalization depends not only on shared variance across states but also on the cognitive processes most relevant to the target behavior.

Our findings are in contrast to recent work suggesting that task paradigms, in general, and movie-watching, in particular, outperform resting state data in predicting cognitive performance (51, 52, 95). While previous studies have often demonstrated a superiority of task and naturalistic viewing over resting state in predicting fluid intelligence or WM (51, 52), there are fewer reports of FC predicting EM (e.g., (96, 97)), and, to our knowledge, no study has compared rest and movie-watching. While we acknowledge that the resting state represents a complex amalgamation of cognitive, emotional, and perceptual processes (98), the good prediction power of the resting state may arise from the presence of mind wandering during rest, which is strongly related to EM (99, 100). EM plays a crucial role in generating mental content during mind wandering, especially episodes characterized by distinct times and locations (61, 101).

In contrast to the EM prediction, both n-back and movie-watching connectomes yielded significant predictions of offline WM performance. Importantly, the models derived from movie-watching and n-back outperformed the resting state in WM prediction. These differences in model accuracy when predicting the same target behaviors (i.e., WM) suggest the presence of trait-state interactions. Specifically, movies and n-back enhance individual differences in WM-relevant connections. Indeed, we found that several WM-related networks (102–105), including the fronto-parietal, the salience, and the dorsal/ventral attention networks, contribute to prediction. Additionally, previous research showed that movie-watching alters the propagation of activity across cortical pathways (106), particularly within and between regions involved in audiovisual processing and attention. These alterations lead to a less segregated and more integrated network organization (107). Similarly, the n-back task has been associated with increased integration of task-positive cortico-cortical connectivity (105, 108) and striato-cortical connectivity (103). Our findings also suggest that certain task contexts strike an optimal balance between reducing neural variability and maintaining sufficient richness to capture individual differences. Prior work shows that task states quench neural variability, leading to a more reliable and predictable neural signal (109). In this context, movie watching may represent such a sweet spot constraining neural dynamics through shared audiovisual stimulation, while simultaneously engaging a broad range of cognitive processes that preserve individual differences. Taken together, our results confirm previous findings that movie watching is a suitable condition for studying individual differences across various cognitive domains. Nonetheless, if a movie-watching paradigm is not feasible/available, resting state still provides a robust means of studying individual differences, particularly in self-referential domains, such as EM.

Our study used a deep neural network architecture that features dense connections and incorporates an attentional mechanism. While our findings demonstrate that a deep learning framework can provide reasonable predictive accuracy, it is important to note that other machine learning approaches (e.g., tree-based models) may offer comparable predictive power, as suggested by prior benchmarking work (29, 30). Our study explicitly compares predictive power across different cognitive states (rest, movie watching, n-back) to identify the states that best capture individual differences across domains. The relative performance of deep learning and other non-linear approaches depends on multiple factors, including sample size, model architecture, feature representation, and domain-specific characteristics of the prediction target. In this context, deep learning was employed as a flexible framework capable of modeling high-

dimensional functional connectivity patterns across cognitive states, rather than as a claim of inherent methodological superiority. Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach.

We found a significant link between BCG, lifestyle, and risk factors for vascular disease, such that individuals with a negative BCG exhibited lower levels of physical activity and higher cardiovascular risk scores compared to those with a positive BCG. This finding was consistent across both age-heterogeneous and older age-homogeneous samples. BCG could serve as a potential biomarker for identifying individuals at risk (e.g., individuals with a negative gap). Previous studies suggest that the brain age prediction gap is associated with cognitive aging (64), some aspects of physiological aging (63), as well as aging in other organs (110) and even mortality in older age (63). However, a recent study revealed that brain age accounts for only a small portion of cognitive decline compared to chronological age (67), suggesting that cognitive prediction might be more informative. Our findings build upon this concept by extending the BCG to behavioral variables, demonstrating that the BCG could provide insights regarding physical activity status, education, and cardiovascular risk – key factors contributing to cognitive reserve (111–114). Note that the association with education was significant only in the DyNAMiC sample and did not reach significance in the COBRA dataset. An important caveat is that BCG can also be conceptualized as an error metric, like mean absolute error or mean square error, reflecting the extent to which models trained in one sample generalize to another. From this perspective, a larger gap may not only indicate individual differences related to resilience factors and dopaminergic function but also reduced model fit or generalizability across datasets. Thus, BCG likely reflects a combination of meaningful biological variability and methodological variance.

Critically, we found that D1DRs and D2DRs were strongly associated with the BCG, such that lower dopamine receptor levels were associated with greater gaps. More specifically, in two independent samples, we discovered greater correspondence (i.e., near zero in BCG) between brain function and cognition in individuals with higher D1DR/D2DR, whereas lower correspondence (i.e., significantly different BCG from zero) was found in individuals with lower D1DR/D2DR availability. Previous computational models proposed that DA modulates neuronal gain, which improves SNR in neural processing, contributing to more coherent activity across large-scale networks (e.g., balanced integration and segregation (85)). Past studies also showed that lower D1DR contributed to more BOLD variability in the subcortical area (83) and less functional segregation of the striatum (115) and the large-scale networks in aging (85), possibly due to increased noise (lower SNR). In support of this notion, we found for the first time that regional variability, estimated using entropy, mediated the impact of DA on BCG. Although the cross-sectional nature of our data warrants caution, this novel finding suggests that lower DA integrity relates to BOLD variability, which in turn is associated with a larger BCG.

An important caveat is that D1DR and D2DR availability do not provide a direct measure of dopamine signaling. Instead, it reflects receptor availability, which interacts with endogenous dopamine in a complex manner. PET measures of D1R and D2R availability reflect the density of unoccupied dopamine receptors and the degree to which endogenous dopamine competes with radioligand binding. D2R binding potential is sensitive to competition from synaptic dopamine, such that higher ambient dopamine generally reduces tracer binding; D1R binding, however, is less affected by endogenous dopamine under physiological conditions, reflecting more directly receptor expression levels. Previous studies demonstrated a significant association between D2R availability and dopamine synthesis capacity measured by FMT (116, 117), suggesting that postsynaptic receptor markers may, under certain conditions, serve as a proxy for dopaminergic signaling. Developmental factors, such as the number of dopamine-producing neurons innervating the striatum, may further influence the structural and functional relationship between pre- and post-synaptic markers. By contrast, smaller studies have reported non-significant (118, 119) or negative (120) associations, although these studies relied on [18F]FDOPA, which is considered a less precise index of dopamine synthesis than FMT. Taken together, these reports indicate that the relationship between pre- and post-synaptic markers is complex and not

necessarily linear. Accordingly, our observation that lower receptor availability is associated with greater neural variability should not be interpreted as direct evidence of weaker dopaminergic signaling, but rather as reflecting the interplay between receptor density and endogenous dopamine occupancy, particularly in the case of D2DR.

Finally, we did not directly compare BCG and brain-age gap (BAG). While our focus was to investigate whether the BCG provides information about factors contributing to cognitive resilience, we acknowledge that benchmarking BCG against the brain-age gap in predicting lifestyle and vascular risk factors would be valuable. However, addressing this question lies beyond the scope of the present study, and future work should systematically compare these approaches. Finally, we acknowledge that our main and validation samples are moderate in size for deep learning, which constrains statistical power and generalizability. Although external validation, early stopping, dropout, and regularization help mitigate overfitting, larger samples will be needed in future work to fully establish the robustness of these predictive models.

In summary, our findings reveal that while tasks like movie-watching predict both episodic and working memory, there are features during rest that can effectively predict internally oriented mind-wandering-type tasks, such as EM. Additionally, individuals whose brains predict poorer cognitive performance (i.e., negative gap) exhibit lower physical activity and higher cardiovascular risk compared to those whose brains predict higher cognitive function than their actual performance (i.e., positive gap). This finding suggests that our prediction model offers a potential marker to identify individuals at risk of compromised brain maintenance. Furthermore, individuals with lower DA showed less accurate cognitive prediction (larger BCG) due to increased BOLD variability and less unique and cohesive FC.

4. Materials and methods

4.1. Participants

All participants provided written informed consent, and studies were conducted in accordance with the Declaration of Helsinki and approved by the Regional Ethical Board and the local Radiation Safety Committee (reference numbers: 2012-57-31M; 2017-404-32M).

This study used data from DyNAMiC (56), which is a longitudinal study with a focus on changes in the brain connectome and the D1DR system. At baseline,180 participants (20-79 years, 50% female) across the adult lifespan underwent all tests between 2017 and 2020 (56) (Fig. 5). Rigorous exclusion criteria were used to recruit a sample without neurological conditions and medical treatments affecting brain functioning and cognition. Exclusion criteria included brain injury or neurological disorder, dementia, neurodevelopmental disorder, psychiatric diagnosis, psychopharmacological treatment, history of severe head trauma, substance abuse or dependence, and illicit drug use. Individuals with other chronic or severe medical conditions (e.g., cancer, diabetes, and Parkinson's disease) were also excluded. Here, we only use data from the baseline measurement.

We used a separate sample as a testing dataset from the COBRA study (88) (Fig. 5). COBRA is a longitudinal aging study in which 181 healthy individuals (64-68 years, 45% female) underwent baseline assessments of the brain, cognition, health, and lifestyle during 2012–2014 (88). Exclusion criteria at baseline included traumatic brain injury, stroke, dementia, intellectual disability, epilepsy, psychiatric and neurological disorders, diabetes, and cancer medications, severe visual or auditory impairment, claustrophobia, and poor Swedish language skills. In the current study, we used data from 177 subjects from COBRA who underwent both MRI and PET examinations at baseline (80, 88, 102, 103, 121).

4.2. Cognitive Measures

The same cognitive test battery was used in DyNAMiC and COBRA (56, 88) (Fig. S1) and assessed two cognitive domains: episodic memory (EM) and working memory (WM). Each domain was assessed using three separate tests, including letter-, number-, and figure-based material,

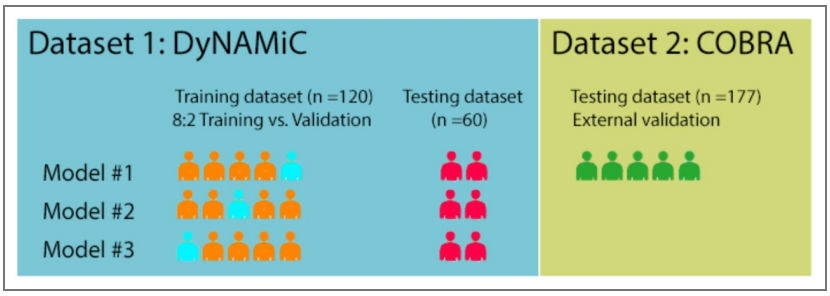


Figure 5. **Overview of the experimental procedure and the use of datasets.**

We used a 3-fold within-sample (DyNAMiC) cross-validation where we trained our model on 120 subjects (8:2; 80% training:20% validation during training) and tested it in a separate sample of 60 subjects. The winning within-sample model was used for between-sample (COBRA) external validation.

respectively. Participants completed all tasks on a computer and responded by either typing in words or numbers; using the computer mouse; or pressing keys marked by different colors. Each test included initial practice runs, after which testing followed across several trials. For each of the three tests within a domain, scores were summarized across trials for a measure of overall performance. A composite score of performances across the three tests was calculated and used as the measure of the cognitive domain in question (i.e., episodic memory, working memory). For each of the three tests, scores were summarized across the total number of trials. The three resulting sum scores were z-standardized and averaged to form one composite score for each domain. The standardization has been carried out independently for the training (DyNAMiC) and test (COBRA) samples.

4.2.1. Episodic Memory (EM)

Tests of EM included word recall, number-word recall, and object-location recall. In word recall, participants viewed 16 Swedish concrete nouns, successively appearing on the computer screen. Each word was presented for 6 s, with an inter-stimulus interval (ISI) of 1 s. Directly following encoding, participants reported as many words as they could using the keyboard. Two trials were completed, with a combined maximum score of 32. In number-word recall, participants encoded pairs of 2-digit numbers and concrete plural nouns (e.g., 46 dogs). During encoding, eight number-word pairs were presented, each for 6 s, with an ISI of 1 s. Directly after encoding, nouns were presented again, in re-arranged order, and participants had to report the 2-digit number associated with each presented noun (e.g., How many dogs?). This test included two trials with a total combined maximum score of 16. The third test assessed object-location memory. Participants viewed a 6 × 6 square grid in which 12 objects were consecutively shown at distinct locations. Each object-position pairing was displayed for 8 s, with an ISI of 1 s. Directly following encoding, all objects were simultaneously displayed next to the grid for the participant to place in their correct position within the grid. If unable to recall the correct position of an object, participants had to guess to the best of their ability. Two trials of this test were completed, making the total combined maximum score 24.

4.2.2. Working Memory (WM)

Working memory was examined with three tests: letter updating, number updating, and spatial updating. These tests differed from the working memory *n*-back task performed during fMRI scanning. In letter-updating, capital letters (A–D) were consecutively presented on the computer screen, with participants instructed to always keep the three final letters in memory. Each letter was presented for 1 s, with an ISI of 0.5 s. When prompted, at any given moment, participants provided their responses by typing in three letters using the keyboard and provided a guessing-based answer if they failed to remember the correct letters. Four practice trials were followed by 16 test trials consisting of either 7-, 9-, 11-, or 13-letter sequences. The combined maximum number of correct answers across trials was 48 (16 trials × 3 reported letters = 48). The number-updating test followed a columnized 3-back design. Three boxes were presented next to each other on the screen throughout the task, in which a single digit (1–9) was sequentially presented from left to right for 1.5 s with an ISI of 0.5 s. During an ongoing sequence, participants had to judge whether the number currently presented in a specific box matched the last number presented in the same box (i.e., appearing three numbers before). For each presented number, they responded yes/no by pressing assigned keys (“yes” = right index finger; “no” = left index finger). Two practice trials were followed by four test trials, each consisting of 30 numbers. The combined maximum number of correct answers across trials (after discarding responses to the first three numbers in every trial, as these were not preceded by any numbers to be matched with) was 108 (27 numbers × 4 trials). In spatial-updating, participants viewed three 3 × 3 square grids presented next to each other on the computer screen. At the start of each trial, a blue circular object was displayed at a random location within each grid. After 4 s, the circular objects disappeared, leaving the grids empty. An arrow then appeared below each grid, indicating that the circular object in the corresponding grid was to be mentally moved one step in the direction of the arrow. The arrows appeared stepwise from left to right in the grids, each presented for 2.5 s (ISI = 0.5 s). Prompted by three new arrows, participants mentally moved the circular objects one more time, resulting in

each circular object having moved two steps from its original location at the end of each trial. Participants indicated which square the circular object in each grid had ended up in using the computer mouse. If unsure, they were instructed to guess. The test was performed across 10 test trials, preceded by five practice trials. The combined maximum number of correct location indications was 30.

4.3. Measure of Physical Activity and Cardiovascular Disease Risk

4.3.1. Physical Activity

An extensive battery of self-rating questionnaires was administered in DyNAMiC and COBRA (56, 88). Participants were asked to indicate the frequency (number of hours during a typical summer week; options: 1-14 h with 1-h increments, or 15+ hrs) and the intensity (how physically demanding on a scale from 1 =“not at all” to 5 =“extremely”) by which they typically engage in a selection of activities relevant to life in northern Sweden. These included 15 specific activities. For the present study, we focused on physical activities and on those activities that are purely physical and that individuals are sufficiently engaged in (i.e., physically demanding ≥2.0). Each of these activities was performed by at least 20% of the participants at least once a week; e.g., walking, bicycling, jogging, strength training, household tasks, and daily work-related activities. We computed physical activity frequency (sum hrs/week) to generate physical activity scores accordingly.

4.3.2. Cardiovascular Disease risk

The risk of cardiovascular disease was determined via a multivariable score, according to the algorithm developed in the Framingham Heart Study (122). Variables include age, sex, hypertension diagnosis, systolic blood pressure, body mass index, smoking, and diabetes mellitus. The risk estimates were derived using an algorithm proposed by D’Agostino et al. (122), which employs Cox proportional-hazard regression models to predict the probability of developing any form of cardiovascular disease within 10 years:

$$\hat{p} = 1 - S_0(t) \exp\left(\sum_{i=1}^{m} \beta_i \left(X_i - \bar{X}_i\right)\right) \quad (1)$$

With $S_0(t)$ being the baseline survival at follow-up t (t =10 years), βi the estimated regression coefficient, X_i the log-transformed value of the i^{th} risk factor, X_i the corresponding mean, and m the number of risk factors included. Baseline survival, means, and regression coefficients were taken from the original algorithm, with DyNAMiC and COBRA participants’ risk variables inserted to compute the final scores. Risk score calculators can be found at *framinghamheartstudy.org*.

4.4. Image Acquisitions

Structural, functional, and neurochemical brain measures were acquired using MRI and PET at Umeå University Hospital in northern Sweden. For both DyNAMiC and COBRA, all MRI data were collected using a 3T Discovery MR750 MRI system (General Electric, Healthcare, Illinois, USA) equipped with a 32-channel phased-array head coil. PET was conducted in 3D mode with a Discovery PET/CT 690 (General Electric, WI, United States) to assess whole-brain D1DR with [^{11}C]SCH23390 and D2DR with [^{11}C]Raclopride at rest in DyNAMiC and COBRA, respectively. Comprehensive descriptions of MRI, PET, and cognitive testing protocols are given elsewhere (56, 88). In this study, we primarily focus on those data directly pertinent to the current investigation.

4.4.1. Functional MRI

For the DyNAMiC dataset, high-resolution anatomical T1-weighted images were collected using a 3

-dimensional (3D) fast spoiled gradient-echo sequence with acquisition parameters of 176 sagittal slices, thickness =1 mm, TR =8.2 ms, TE =3.2 ms, flip angle =12°, and a field of view (FOV) =250×250 mm. Whole-brain functional images were acquired during resting state, naturalistic viewing, and an n-back WM task. Functional images were acquired using a T2*-weighted single-shot echo-planar imaging (EPI) sequence, with 330 volumes collected over 12 min. The sequence provided 37 axial slices, slice thickness =3.4 mm, 0.5 mm spacing, TR =2,000 ms, TE =30 ms, flip angle =80°, and FOV =250×250 mm. Ten dummy scans were collected at the start of the sequence. During the resting state, participants were instructed to keep their eyes open and focus on a white fixation cross on a black background displayed on a computer screen through a tilted mirror attached to the head coil. WM was assessed in the scanner (12 min) with a numerical *n*-back task, which consisted of blocks of 1-back, 2-back, and 3-back (102, 103). During movie-watching, the participants viewed and listened to a 12-minute video consisting of selected and chronologically ordered sections from the Swedish movie *Cockpit* (123). Participants were instructed to view the movie attentively and answer a short multiple-choice questionnaire about the movie after the scanning session. We did not monitor participants' responses to the movie, and the chosen clips were selected to be relatively neutral in emotional content. The storyline follows Valle, a recently fired pilot whose marriage has ended, as he struggles to find new employment. In a desperate attempt to secure a job at an airline specifically recruiting a female pilot, he presents himself as a woman.

In COBRA, data were collected using identical scans for resting state and n-back WM. However, the resting state scan was shorter, lasting only 6 minutes (75, 103).

4.5. Functional Connectivity Analysis

Functional data from all conditions (i.e., rest, movie, n-back) were pre-processed using the Statistical Parametric Mapping software package (SPM12). The functional images were first corrected for slice-timing differences and in-scanner motion, followed by registration to anatomical T1 images. Distortion correction was performed using subject-specific and T1 co-registered field maps. The functional time series were subsequently demeaned and detrended, followed by simultaneous nuisance regression and temporal high-pass filtering (threshold at 0.009 Hz) to not re-introduce nuisance signals (124). The nuisance regression model included mean cerebrospinal and white-matter signal and their derivatives, 24-motion parameters (125), a binary vector flagging motion-contaminated volumes exceeding framewise displacement (FD) of 0.2 mm (126), in addition to an 18-parameter RETRICOR model (127, 128) of cardiac pulsation (up to third-order harmonics), respiration (up to fourth-order harmonics), and first-order cardio-respiratory interactions estimated using the PhysIO Toolbox v.5.0 (129). Regression models for n-back included an additional set of finite impulse response (FIR) task regressors (130) to avoid false positive connectivity due to task-evoked activations (131). The FIR regression approach involved fitting mean cross-block responses for each time point within a time-locked window of equal duration to each task block (27 blocks of 20 s), extended by an additional 18 s following each block to account for the duration of the hemodynamic response function (HRF). Given that this approach linearly fits a set of binary task regressors for each time point, it is nearly identical to subtracting the mean task response, with the difference being that FIR regression is better able to handle overlapping task responses and differences in the shape of the HRF (131). The nuisance-regressed images were subsequently normalized to sample-specific group templates (DARTEL) (132) for each dataset, respectively, followed by spatial smoothing using a 6-mm FWHM Gaussian kernel to mitigate DARTEL-induced aliasing and affine-transformed to stereotactic MNI space (ICBM152NLin2009) (133, 134). Functional images from both DyNAMiC and COBRA were preprocessed according to the steps described above, with the only exception that the RETRICOR parameters were excluded from the regression models in COBRA due to technical issues related to the respiration and cardiac traces.

4.5.1. Graph Construction

For all fMRI conditions, functional time series were averaged from 273 cortical and subcortical regions, represented by 5-mm radius spheres, based on a widely employed FC parcellation (89). In addition to 264 regions reported in the Power parcellation, we added 9 additional regions, including some subcortical regions, such as putamen, caudate, and anterior and posterior hippocampus, identified using independent component analysis (8, 135). These regions were categorized into 14 resting state networks according to a consensus partition (89). To mitigate sampling from non-gray matter voxels, each parcel underwent erosion by a permissive gray matter mask (eroding voxels <.1% threshold). The averaged time series were then subjected to Pearson's correlations, followed by Fisher's r-to-z transformation, resulting in the creation of a 273×273 adjacency matrix for each participant, with coefficients along the main diagonals set to zero.

To further investigate the impact of network parcellation, we replicated our prediction analysis (Supplementary Material, Figs S1-S2, and Table S1-S2) using Schaefer parcellation (92), which entails 300 cortical and subcortical regions.

4.6. Deep Neural Network Model

Based on convolutional neural networks, deep learning is an advanced form of artificial intelligence that uses multiple layers of "hidden" neural networks. Deep learning methodologies are capable of automatically identifying complex patterns and representations directly from raw data using these multi-layered networks, thus eliminating the need for explicit feature engineering or manual intervention (136). The success of the training and learning phases depends on the model's ability to process high-dimensional input data, extracting meaningful features from complex data. This is done while managing the number of trainable parameters, which are crucial for automated feature learning during the construction of the model.

In this study, the inputs to our deep learning models were subject-specific FC maps with a matrix size of 273×273 (e.g., Fig. 6a). We generated different versions of each FC map by replacing portions of the network. For example, we relocated the DMN network toward the last nodes, which were assigned as DAN, cerebellar, subcortical, and uncertain in Figure 6a. This approach allowed us to create a diverse set of FC matrices for each individual, each reflecting a different composition of edges and neighbors while maintaining the linear relationships exonerated in the original data. Consequently, we augmented the dataset by producing a total of 3,600 FC maps from the initial set of FC maps.

Following the DenseNet framework (90), we incorporated the Enhanced Residual Block (ERB) and High-Frequency Attention Block (HFAB) into the dense layers (Fig. 6b), termed DenselyAttention, to facilitate feature reuse in each layer. DenseNet (90) diverges from traditional methods like deepening layers or widening network structure by focusing on feature reuse and bypass settings. This results in fewer parameters than similar dense networks such as ResNet (137), enhances feature reuse, improves feature propagation, makes training easier, and reduces issues of gradient vanishing and model degradation.

The architecture of DenseNet is characterized by its dense connectivity pattern, which entails direct connections from each layer to all subsequent layers within its dense block (as illustrated in Fig. 6b). This design ensures that every layer has access to the feature maps generated by preceding layers, thereby facilitating a seamless and efficient gradient flow throughout the network. In essence, the knowledge acquired at each layer is propagated forward, enabling the model to effectively capture intricate patterns and dependencies within the data, ultimately enhancing its ability to learn and generalize (90). Additionally, dense blocks provide a regularizing effect, reducing overfitting, particularly on tasks with smaller training datasets (90). This is suitable for this study, which includes relatively small samples. Each sequence combines operations of batch normalization (BN), rectified linear unit (ReLU), and a 3×3 convolution (Conv). Batch normalization can effectively prevent overfitting, as described by equation 2:

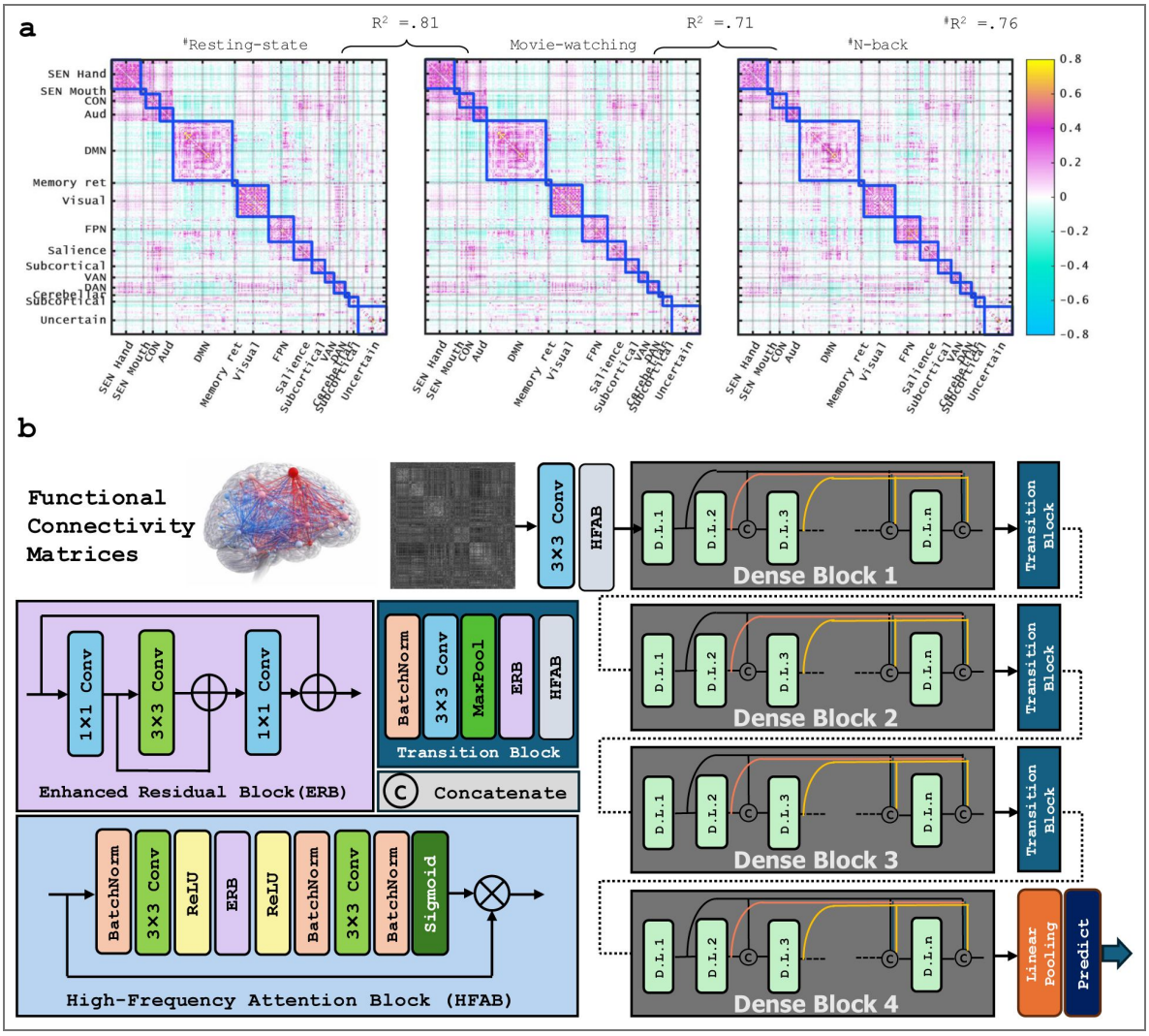


Figure 6.

(a) Example of a functional connectivity map across three different cognitive states. SEN hand: SENsory hand; SEN Mouth: SENsory Mouth; CON: Cingulo-Operculum control Network; Aud: Auditory; DMN: Default Mode Network; Memory Ret: Memory Retrieval network; Visual: Visual; FPN: Fronto-Parietal Network; Salience: Salience control network; Subcortical (upper row): subcortical network included in original Power parcellation; VAN: Ventral Attention Network; DAN: Dorsal Attention Network; Cerebellar: Cerebellar network; Subcortical (lower row): additional Subcortical regions, including hippocampus and caudate, added to the original Power Parcellation; Uncertain: Regions with less known network assignment. **(b)** DenselyAttention architecture. Enhanced Residual Block (ERB) and High-Frequency Attention Block (HFAB) into the Transition Block. Note that each "D.L." layer in the table corresponds to the sequence BatchNormalization-ReLU-Conv3×3.

$$BN\left(x_n\right)=\alpha\left(\frac{x_n-\mu_B}{\sqrt{\sigma_\beta^2+\epsilon}}\right)+\beta, \tag{2}$$

Where μ_β, mean; σ_β, standard deviation; ε, random noise; α and β are adaptable variables in training. We utilized the Rectified Linear Unit (ReLU) activation function (138), which activates neurons by directly outputting the input if it is positive, or zero otherwise, as outlined in equation 2.

$$\mathrm{ReLu}(x)=\max(0,x), \tag{3}$$

Where *x* is the input to a neuron. ReLU benefits the performance of networks with dense layers by decreasing the computation and selectively optimizing parameters. Four dense blocks facilitate a stepwise down-sampling in the network. These blocks are connected with transition layers, which consist of a 1×1 convolutional layer followed by a 2×2 average pooling layer.

Additionally, ERB and HFAB pairs are introduced for targeted high-frequency features and residual block enhancement. The ERB-HFAB pairs are stacked sequentially at the beginning of each dense block loop (Fig. 6b), with ERB and HFAB having 16 feature maps.

High-Frequency Attention Block (HFAB)

Our approach to attentional mechanisms, specifically the HFAB, introduces a sequential attention branch inspired by edge detection. This branch rescales each position based on its neighboring pixels, efficiently focusing on high-frequency areas. The HFAB employs a 3×3 convolution to enhance both the receptive field and computational efficiency. Batch normalization is seamlessly integrated into the attention branch, introducing global statistics during inference without additional computational cost.

Enhanced Residual Block (ERB)

We present ERB as an alternative to the traditional residual block. As illustrated in Figure 6b, ERB comprises a re-parameterization block (RepBlock) and a ReLU. In the training stage, the RepBlock utilizes a 1×1 convolution to either expand or contract the number of feature maps, employing a 3×3 convolution to extract features in a higher-dimensional space. Furthermore, two skip connections are integrated to mitigate training complexities. During inference, all linear transformations can be unified, facilitating the conversion of each RepBlock into a singular 3×3 convolution. Essentially, ERB capitalizes on the advantages of residual learning.

We implemented model training using Tensorflow 2.11.0 (139) and Keras 2.11.0 (140) as programming interfaces and trained on a fifth-generation MacBook Pro (Apple M1 MAX silicon chips, 10-core CPU, 24-core GPU, 16-core Neural Engine, 64 GB memory). For regression tasks, selecting an appropriate loss function is crucial for guiding the optimization process and ensuring accurate predictions. In this study, we opted for the mean squared error (MSE) loss function due to its suitability for regression problems. To minimize the loss function, we trained the network using the stochastic gradient descent (SGD) optimizer with a learning rate of $8e^{-5}$ and a Nesterov momentum of 0.9 (141). The number of epochs was set to 100, and the batch size was set to 74. We also added dropout (142) after each convolutional layer, except the first one, with a rate of 0.15.

We conducted training on distinct, identical datasets extracted from DyNAMiC, each comprising FC maps generated from 120 subjects. To prepare each training dataset, we randomly shuffled data for training and validation patches, allocating 80% for training and 20% for validation. For testing, we utilized all FC maps of 60 subjects from the same sample (the age-heterogeneous DyNAMiC study), enabling us to assess model performance through three-fold cross-validation (Fig. 5). Each cross-validation fold was a new training with an unseen validation set for the model. Based on its performance on the testing dataset, we selected and employed the winning model for all

subsequent analyses. This model, which demonstrated the best performance on the testing dataset, underwent external validation in an independent sample from the age-homogeneous COBRA study.

To explore and visually represent the crucial features of the deep learning models contributing to the prediction of cognitive scores, we used the Grad-CAM (Gradient-weighted Class Activation Mapping) technique (143). This method interprets the model's decisions by highlighting the regions of the input image with the most significant impact on the model's output. Grad-CAM conducts a backward pass to compute the gradients of the target class score with respect to the feature maps of the final convolutional layer. We present the average heatmaps calculated for all input data to the model (143). Grad-CAM saliency maps were interpreted qualitatively, with a heuristic threshold (≥ 0.5) applied to highlight regions with relatively higher contribution to the model's predictions. These values do not reflect statistical significance and should therefore be interpreted descriptively. A further limitation of this study is the absence of ground truth for validating whether highlighted regions truly correspond to the features used by the model during prediction. As such, Grad-CAM provides an approximation of model attention rather than a definitive measure of feature importance. Nevertheless, Grad-CAM remains one of the most widely used and empirically validated interpretability techniques in deep learning, particularly in medical imaging applications. Its integration with established frameworks such as Keras and TensorFlow, together with its ability to generate spatial attributions that align with domain knowledge, makes it a suitable choice for the present study. Future work may incorporate complementary interpretability approaches, including adaptations of the Haufe transformation where applicable to deep learning architectures.

4.7. Positron Emission Tomography (PET)

The scanning sessions started with a 5-minute low-dose helical CT scan (20 mA, 120 kV, 0.8 s per revolution), obtained for attenuation correction. During scanning, a thermoplastic mask was attached to the bed surface to minimize head movement.

In DyNAMiC, a 60-minute scan was performed following 350 MBq (337 ± 27 MBq) in list-mode format. Offline re-binning of list-mode data was conducted to achieve a total of 49 frames with increasing length. In COBRA, a 55-min, 18-frame dynamic PET scan was acquired during rest following intravenous bolus injection of approximately 250 MBq 11C-raclopride (264 ± 19 MBq). For both studies, attenuation- and decay-corrected images (47 slices, field of view = 25 cm, 256×256-pixel transaxial images, voxel size = 0.977× 0.97×3.27 mm^3) were reconstructed with the iterative VUE Point HD-SharpIR algorithm (GE; 6 iterations, 24 subsets, 3.0 mm post filtering; full-width-at-half-maximum (FWHM): 3.2 mm). The estimation of receptor availability or binding potential relative to non-displaceable binding (BP_{ND}) was carried out following previously described procedures with the cerebellum as a reference region (76). PET images were motion-corrected and co-registered with the structural T1-weighted images from the corresponding session using Statistical Parametric Mapping software (SPM12, Wellcome Department of Imaging Science, Functional Imaging Laboratory, London, UK). Motion-corrected PET data were resliced to match the spatial dimensions of MR data (1 mm^3 isotropic, 256×256×256). The mean of the first five frames was used as a source for co-registration. In DyNAMiC, frame-to-frame head motion correction, with translations ranging from 0.23 to 4.22 mm (mean ± sd =0.95 ± 0.54 mm), revealed a trend-level difference across age-groups (age <40 and age ≥40 years), as determined by Student's t-test (t =2.0, p =.047; mean ± sd for younger individuals = 1.07 ± 0.52, mean ± sd for older individuals =0.90 ± 0.55). Partial-volume-effect (PVE) correction was achieved using the symmetric geometric transfer matrix (SGTM) method for regional correction, implemented in FreeSurfer (144). An estimated point-spread-function of 2.5 mm FWHM was utilized. Regional estimates of BP_{ND} were calculated within the automated FreeSurfer segmentations employing the simplified reference tissue model (SRTM (145)). In the current study, we focused on the striatal BP_{ND}, calculated as an average of BP_{ND} across the caudate and putamen.

4.8. The direct impact of dopamine on BCG through mediation analysis

To evaluate whether functional variability mediates the relationship between D1DR and prediction gap connectivity, we conducted additional mediation analyses. We first computed entropy, which estimates within-region signal variability (86), and then averaged this measure across all striatal regions (left caudate: MNI coordinate ≈ [–12, 12, 6], right caudate: MNI coordinate ≈ [10, 14, 2], left putamen: MNI coordinates ≈ [–24, 6, 4], right putamen: MNI coordinates ≈ [29, 1, 4].). Following the mediation analysis framework proposed by Baron and Kenny (146), our goal was to determine whether the association between D1/D2 receptors and BCG is mediated by regional variability (entropy) or if the indirect effect exceeds the direct association between D1/D2 receptors and BCG. To assess the statistical significance of this mediation effect, we employed the bootstrapping method as outlined by Preacher and Hayes (147), and age has been controlled for in all statistical analyses.

4.9. Statistical significance analysis

Statistical analyses were carried out using SPSS (IBM Corp., V24.0.0, Armonk, NY, USA), MATLAB (The MathWorks Inc., V9.13.0 (R2022b), Natick, MA, USA), and GraphPad Prism (GraphPad Software, Inc., V5.01, CA, USA). We performed partial correlations between predicted and actual scores, as well as linear regression analyses. To investigate the relationship between generated gap variables and DA receptor availability, we controlled for age (in DyNAMiC) using partial correlation. The Mann-Whitney *U* test was used to calculate the mean differences in prediction accuracy. The level of statistical significance was set at *p-value* ≤0.05. For the bootstrap-based comparison of model performance (bootstrap resampling with 1000 iterations), no test statistic with an associated degree of freedom is reported. Instead, statistical inference is based on the bootstrap distribution of the difference in correlation coefficients (Δr) and its 95% confidence interval. As bootstrap confidence-interval-based inference does not rely on an analytic sampling distribution, degrees of freedom are not defined for this procedure.

Out-of-sample predictive performance was quantified using the coefficient of determination (r^2) computed via a sum-of-squares formulation (148). Unlike squared correlation coefficients, which capture only linear association, this metric evaluates how well model predictions approximate observed values relative to a baseline model. Specifically, out-of-sample r^2 was defined as

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where y_i denotes the observed outcome in the test set, $\hat{y}_i$ the corresponding model prediction, and $\bar{y}$ train denotes the mean of the outcome variable in the training set. Using the training-set mean as the baseline ensures a strictly out-of-sample evaluation and avoids information leakage. Under this formulation, positive r^2 values indicate performance exceeding the null model (predicting the training mean), whereas negative values indicate worse-than-baseline performance. Because this formulation directly compares prediction error to baseline variance, it provides a more appropriate measure of predictive accuracy than correlation-based metrics, particularly in the presence of scale or offset differences between predicted and observed values (148).

Data availability

The scripts used for developing the model are available at *https://github.com/MorEsm/AI-based-Prediction-of-Cognitive-Function*.

Acknowledgements

This work was funded by the Swedish Research Council (grant number 2021-02558), Knut and Alice Wallenberg Foundation (Wallenberg Fellow grant to A.S.), Bank of Sweden (RJ, grant number P20-0515 to A.S.), StratNeuro grant at Karolinska Institute (A.S.). Morteza Esmaeili and Erin

Bjørkeli were supported by the Southern Eastern Norway Regional Health Authority (Helse Sør-Øst RHF, HSØ, grant numbers 2018047 and 2021023, respectively).

Additional information

Author Contributions

M.E. and A.S. (together with the COBRA PIs N.K. and L.N.) conceptualized the study, formulated the research questions, and developed the methodology. M.E., R.P., J.J., E.B.B., and K.N. performed data processing and formal analyses. M.E., A.S., R.B., J.J., K.N., N.K., L.B., and L.N. contributed to the interpretation of the findings. M.E. and A.S. supervised the study and drafted the original manuscript. All authors participated in manuscript writing and editing and approved the final version.

Funding

Funder	Grant reference number	Author
HOD | Helse Sør-Øst RHF (sorost)	2018047	Morteza Esmaeili
HOD | Helse Sør-Øst RHF (sorost)	2021023	Erin Beate Bjørkeli
University of Gothenburg | Wallenberg Centre for Molecular and Translational Medicine (WCMTM)	P20-0515	Alireza Salami
Karolinska Institute	StratNeuro grant	Alireza Salami
Swedish Research Council	2021-02558	Alireza Salami

Author ORCID iDs

Morteza Esmaeili: https://orcid.org/0000-0003-0686-3571
Jarkko Johansson: https://orcid.org/0000-0002-4501-4735
Kristin Nordin: https://orcid.org/0000-0003-4157-1638
Lars Nyberg: https://orcid.org/0000-0002-3367-1746
Alireza Salami: https://orcid.org/0000-0002-4675-8437

Additional files

Supplementary materials

References

1. **Biswal B,** Yetkin FZ, Haughton VM, Hyde JS (1995) Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn Reson Med* **34**:537-41 https://doi.org/10.1002/mrm.1910340409 | PubMed
2. **Buckner RL,** Krienen FM (2013) The evolution of distributed association networks in the human brain. *Trends Cogn Sci* **17**:648-65 https://doi.org/10.1016/j.tics.2013.09.017 | PubMed
3. **van den Heuvel MP,** Hulshoff Pol HE (2010) Exploring the brain network: a review on resting-state fMRI functional connectivity. *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology* **20**:519-34 https://doi.org/10.1016/j.euroneuro.2010.03.008 | PubMed
4. **Ferreira LK,** Busatto GF (2013) Resting-state functional connectivity in normal brain aging. *Neurosci Biobehav Rev* **37**:384-400 https://doi.org/10.1016/j.neubiorev.2013.01.017 | PubMed
5. **Damoiseaux JS** (2017) Effects of aging on functional and structural brain connectivity. *Neuroimage* **160**:32-40 https://doi.org/10.1016/j.neuroimage.2017.01.077 | PubMed
6. **Fox MD,** Greicius M (2010) Clinical applications of resting state functional connectivity. *Front Syst Neurosci* **4** https://doi.org/10.3389/fnsys.2010.00019 | PubMed

7. **Andrews-Hanna JR,** Snyder AZ, Vincent JL, Lustig C, Head D, Raichle ME, et al. (2007) Disruption of large-scale brain systems in advanced aging. *Neuron* **56**:924-35 https://doi.org/10.1016/j.neuron.2007.10.038 | PubMed
8. **Salami A,** Wahlin A, Kaboodvand N, Lundquist A, Nyberg L (2016) Longitudinal Evidence for Dissociation of Anterior and Posterior MTL Resting-State Connectivity in Aging: Links to Perfusion and Memory. *Cereb Cortex* **26**:3953-63 https://doi.org/10.1093/cercor/bhw233 | PubMed
9. **Kaboodvand N,** Backman L, Nyberg L, Salami A (2018) The retrosplenial cortex: A memory gateway between the cortical default mode network and the medial temporal lobe. *Hum Brain Mapp* **39**:2020-34 https://doi.org/10.1002/hbm.23983 | PubMed
10. **Seeley WW,** Menon V, Schatzberg AF, Keller J, Glover GH, Kenna H, et al. (2007) Dissociable intrinsic connectivity networks for salience processing and executive control. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **27**:2349-56 https://doi.org/10.1523/jneurosci.5587-06.2007 | PubMed
11. **Avelar-Pereira B,** Backman L, Wahlin A, Nyberg L, Salami A (2017) Age-Related Differences in Dynamic Interactions Among Default Mode, Frontoparietal Control, and Dorsal Attention Networks during Resting-State and Interference Resolution. *Front Aging Neurosci* **9** https://doi.org/10.3389/fnagi.2017.00152 | PubMed
12. **Avelar-Pereira B,** Backman L, Wahlin A, Nyberg L, Salami A (2020) Increased functional homotopy of the prefrontal cortex is associated with corpus callosum degeneration and working memory decline. *Neurobiol Aging* **96**:68-78 https://doi.org/10.1016/j.neurobiolaging.2020.08.008 | PubMed
13. **Machner B,** Braun L, Imholz J, Koch PJ, Munte TF, Helmchen C, et al. (2021) Resting-State Functional Connectivity in the Dorsal Attention Network Relates to Behavioral Performance in Spatial Attention Tasks and May Show Task-Related Adaptation. *Front Hum Neurosci* **15** https://doi.org/10.3389/fnhum.2021.757128 | PubMed
14. **Schimmelpfennig J,** Topczewski J, Zajkowski W, Jankowiak-Siuda K (2023) The role of the salience network in cognitive and affective deficits. *Front Hum Neurosci* **17** https://doi.org/10.3389/fnhum.2023.1133367 | PubMed
15. **Ferreira LK,** Regina AC, Kovacevic N, Martin Mda G, Santos PP, Carneiro Cde G, et al. (2016) Aging Effects on Whole-Brain Functional Connectivity in Adults Free of Cognitive and Psychiatric Disorders. *Cereb Cortex* **26**:3851-65 https://doi.org/10.1093/cercor/bhv190 | PubMed
16. **Smith SM,** Elliott LT, Alfaro-Almagro F, McCarthy P, Nichols TE, Douaud G, et al. (2020) Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *eLife* **9** https://doi.org/10.7554/elife.52677 | PubMed
17. **Shen X,** Finn ES, Scheinost D, Rosenberg MD, Chun MM, Papademetris X, et al. (2017) Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat Protoc* **12**:506-18 https://doi.org/10.1038/nprot.2016.178 | PubMed
18. **Vul E,** Harris C, Winkielman P, Pashler H (2009) Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspect Psychol Sci* **4**:274-90 https://doi.org/10.1111/j.1745-6924.2009.01125.x | PubMed
19. **Bzdok D,** Eickenberg M, Varoquaux G, Thirion B (2017) Hierarchical Region-Network Sparsity for High-Dimensional Inference in Brain Imaging. *Inf Process Med Imaging* **10265**:323-35 https://doi.org/10.1007/978-3-319-59050-9_26 | PubMed
20. **Bzdok D,** Meyer-Lindenberg A (2018) Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging* **3**:223-30 https://doi.org/10.1016/j.bpsc.2017.11.007 | PubMed
21. **Bzdok D,** Krzywinski M, Altman N (2018) Machine learning: supervised methods. *Nat Methods* **15**:5-6 https://doi.org/10.1038/nmeth.4551 | PubMed
22. **Sui J,** Jiang R, Bustillo J, Calhoun V (2020) Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biol Psychiatry* **88**:818-28 https://doi.org/10.1016/j.biopsych.2020.02.016 | PubMed

23. **Vieira S,** Liang X, Guiomar R, Mechelli A (2022) Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clin Psychol Rev* **97**:102193 https://doi.org/10.1016/j.cpr.2022.102193 | PubMed

24. **Plis SM,** Hjelm DR, Salakhutdinov R, Allen EA, Bockholt HJ, Long JD, et al. (2014) Deep learning for neuroimaging: a validation study. *Front Neurosci* **8** https://doi.org/10.3389/fnins.2014.00229 | PubMed

25. **van der Burgh HK,** Schmidt R, Westeneng HJ, de Reus MA, van den Berg LH, van den Heuvel MP (2017) Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *Neuroimage Clin* **13**:361-9 https://doi.org/10.1016/j.nicl.2016.10.008 | PubMed

26. **Nguyen H,** Nguyen V, Nguyen T, Larsen ME, O'Dea B, Nguyen DT, et al. (2018) Jointly Predicting Affective and Mental Health Scores Using Deep Neural Networks of Visual Cues on the Web. In: Web Information Systems Engineering – WISE 2018.

27. **Li H,** Jiang G, Zhang J, Wang R, Wang Z, Zheng WS, et al. (2018) Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *Neuroimage* **183**:650-65 https://doi.org/10.1016/j.neuroimage.2018.07.005 | PubMed

28. **Choi Y,** Kwon Y, Lee H, Kim BJ, Paik MC, Won J-H (2016) Ensemble of Deep Convolutional Neural Networks for Prognosis of Ischemic Stroke. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-55524-9_22

29. **He T,** Kong R, Holmes AJ, Nguyen M, Sabuncu MR, Eickhoff SB, et al. (2020) Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage* **206** https://doi.org/10.1016/j.neuroimage.2019.116276 | PubMed

30. **Vieira BH,** Schöttner M, Calhoun VD, Salmon CEG (2024) Beyond functional connectivity: deep learning applied to resting-state fMRI time series in the prediction of 58 human traits in the HCP. *bioRxiv* https://doi.org/10.1101/2024.03.07.583858

31. **Kong J,** Wolcott E, Wang Z, Jorgenson K, Harvey WF, Tao J, et al. (2019) Altered resting state functional connectivity of the cognitive control network in fibromyalgia and the modulation effect of mind-body intervention. *Brain Imaging Behav* **13**:482-92 https://doi.org/10.1007/s11682-018-9875-3 | PubMed

32. **Sripada C,** Angstadt M, Taxali A, Clark DA, Greathouse T, Rutherford S, et al. (2021) Brain-wide functional connectivity patterns support general cognitive ability and mediate effects of socioeconomic status in youth. *Transl Psychiatry* **11**:571 https://doi.org/10.1038/s41398-021-01704-0 | PubMed

33. **Chen J,** Tam A, Kebets V, Orban C, Ooi LQR, Asplund CL, et al. (2022) Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nat Commun* **13**:2217 https://doi.org/10.1038/s41467-022-29766-8 | PubMed

34. **Omidvarnia A,** Sasse L, Larabi DI, Raimondo F, Hoffstaedter F, Kasper J, et al. (2023) Is resting state fMRI better than individual characteristics at predicting cognition?. *Research Square* https://doi.org/10.21203/rs.3.rs-2631029/v1

35. **Jiang L,** Qiao K, Li C (2021) Distance-based functional criticality in the human brain: intelligence and emotional intelligence. *BMC Bioinformatics* **22**:32 https://doi.org/10.1186/s12859-021-03973-4 | PubMed

36. **Vieira BH,** Pamplona GSP, Fachinello K, Silva AK, Foss MP, Salmon CEG (2022) On the prediction of human intelligence from neuroimaging: A systematic review of methods and reporting. *Intelligence* **93** https://doi.org/10.1016/j.intell.2022.101654

37. **Jangraw DC,** Gonzalez-Castillo J, Handwerker DA, Ghane M, Rosenberg MD, Panwar P, et al. (2018) A functional connectivity-based neuromarker of sustained attention generalizes to predict recall in a reading task. *Neuroimage* **166**:99-109 https://doi.org/10.1016/j.neuroimage.2017.10.019 | PubMed

38. **Chen Q,** Beaty RE, Cui Z, Sun J, He H, Zhuang K, et al. (2019) Brain hemispheric involvement in visuospatial and verbal divergent thinking. *Neuroimage* **202** https://doi.org/10.1016/j.neuroimage.2019.116065 | PubMed

39. **Rosenberg MD,** Finn ES, Scheinost D, Papademetris X, Shen X, Constable RT, et al. (2016) A neuromarker of sustained attention from whole-brain functional connectivity. *Nat Neurosci* **19**:165-71 https://doi.org/10.1038/nn.4179 | PubMed

40. **Beaty RE,** Kenett YN, Christensen AP, Rosenberg MD, Benedek M, Chen Q, et al. (2018) Robust prediction of individual creative ability from brain functional connectivity. *Proc Natl Acad Sci U S A* **115**:1087-92 https://doi.org/10.1073/pnas.1713532115 | PubMed

41. **Hsu WT,** Rosenberg MD, Scheinost D, Constable RT, Chun MM (2018) Resting-state functional connectivity predicts neuroticism and extraversion in novel individuals. *Soc Cogn Affect Neurosci* **13**:224-32 https://doi.org/10.1093/scan/nsy002 | PubMed

42. **Gabrieli JDE,** Ghosh SS, Whitfield-Gabrieli S (2015) Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* **85**:11-26 https://doi.org/10.1016/j.neuron.2014.10.047 | PubMed

43. **Finn ES,** Todd Constable R (2016) Individual variation in functional brain connectivity: implications for personalized approaches to psychiatric disease. *Dialogues Clin Neurosci* **18**:277-87 https://doi.org/10.31887/dcns.2016.18.3/efinn | PubMed

44. **Woo CW,** Chang LJ, Lindquist MA, Wager TD (2017) Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci* **20**:365-77 https://doi.org/10.1038/nn.4478 | PubMed

45. **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**:1664-71 https://doi.org/10.1038/nn.4135 | PubMed

46. **Finn ES,** Scheinost D, Finn DM, Shen X, Papademetris X, Constable RT (2017) Can brain state be manipulated to emphasize individual differences in functional connectivity?. *Neuroimage* **160**:140-51 https://doi.org/10.1016/j.neuroimage.2017.03.064 | PubMed

47. **Grady CL,** Ryan JD (2017) Age-Related Differences in the Human Hippocampus: Behavioral, Structural and Functional Measures. In: Hannula DE, Duff MC (Eds). *The Hippocampus from Cells to Systems: Structure, Connectivity, and Functional Contributions to Memory and Flexible Cognition* Cham: Springer International Publishing. pp. 167-208 https://doi.org/10.1007/978-3-319-50406-3_7

48. **Campbell KL,** Schacter DL (2017) Aging and the Resting State: Cognition is not Obsolete. *Lang Cogn Neurosci* **32**:692-4 https://doi.org/10.1080/23273798.2016.1265658 | PubMed

49. **Geerligs L,** Tsvetanov KA (2017) The use of resting state data in an integrative approach to studying neurocognitive ageing - Commentary on Campbell and Schacter (2016). *Lang Cogn Neurosci* **32**:684-91 https://doi.org/10.1080/23273798.2016.1251600 | PubMed

50. **Tavor I,** Parker Jones O, Mars RB, Smith SM, Behrens TE, Jbabdi S (2016) Task-free MRI predicts individual differences in brain activity during task performance. *Science* **352**:216-20 https://doi.org/10.1126/science.aad8127 | PubMed

51. **Greene AS,** Gao S, Scheinost D, Constable RT (2018) Task-induced brain state manipulation improves prediction of individual traits. *Nat Commun* **9**:2807 https://doi.org/10.1038/s41467-018-04920-3 | PubMed

52. **Finn ES,** Bandettini PA (2021) Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *Neuroimage* **235** https://doi.org/10.1016/j.neuroimage.2021.117963 | PubMed

53. **Finn ES,** Glerean E, Hasson U, Vanderwal T (2022) Naturalistic imaging: The use of ecologically valid conditions to study brain function. *Neuroimage* **247** https://doi.org/10.1016/j.neuroimage.2021.118776 | PubMed

54. **Hasson U,** Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. *Science* **303**:1634-40 https://doi.org/10.1126/science.1089506 | PubMed

55. **Geerligs L,** Rubinov M, Cam C, Henson RN (2015) State and Trait Components of Functional Connectivity: Individual Differences Vary with Mental State. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **35**:13949-61 https://doi.org/10.1523/jneurosci.1324-15.2015 | PubMed

56. **Nordin K,** Gorbach T, Pedersen R, Panes Lundmark V, Johansson J, Andersson M, et al. (2022) DyNAMiC: A prospective longitudinal study of dopamine and brain connectomes: A new window into cognitive aging. *J Neurosci Res* **100**:1296-320 https://doi.org/10.1002/jnr.25039 | PubMed

57. **Cattell RB** (1971) *Abilities: Their structure, growth, and action*

58. **Kyllonen PC** (2013) Is working memory capacity Spearman's g?. In: *Human abilities* Psychology Press. pp. 49-75

59. **Unsworth N,** Fukuda K, Awh E, Vogel EK (2014) Working memory and fluid intelligence: capacity, attention control, and secondary memory retrieval. *Cogn Psychol* **71**:1-26 https://doi.org/10.1016/j.cogpsych.2014.01.003 | PubMed

60. **Tulving E,** Schacter DL (1990) Priming and human memory systems. *Science* **247**:301-6 https://doi.org/10.1126/science.2296719 | PubMed

61. **Smallwood J,** Schooler JW (2015) The science of mind wandering: empirically navigating the stream of consciousness. *Annu Rev Psychol* **66**:487-518 https://doi.org/10.1146/annurev-psych-010814-015331 | PubMed

62. **Dosenbach NU,** Nardos B, Cohen AL, Fair DA, Power JD, Church JA, et al. (2010) Prediction of individual brain maturity using fMRI. *Science* **329**:1358-61 https://doi.org/10.1126/science.1194144 | PubMed

63. **Cole JH,** Franke K (2017) Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. *Trends Neurosci* **40**:681-90 https://doi.org/10.1016/j.tins.2017.10.001 | PubMed

64. **Dunas T,** Wahlin A, Nyberg L, Boraxbekk CJ (2021) Multimodal Image Analysis of Apparent Brain Age Identifies Physical Fitness as Predictor of Brain Maintenance. *Cereb Cortex* **31**:3393-407 https://doi.org/10.1093/cercor/bhab019 | PubMed

65. **Farnsworth von Cederwald B,** Josefsson M, Wahlin A, Nyberg L, Karalija N (2022) Association of Cardiovascular Risk Trajectory With Cognitive Decline and Incident Dementia. *Neurology* **98**:e2013-e22 https://doi.org/10.1212/wnl.0000000000200255 | PubMed

66. **Jonasson LS,** Nyberg L, Kramer AF, Lundquist A, Riklund K, Boraxbekk CJ (2016) Aerobic Exercise Intervention, Cognitive Performance, and Brain Structure: Results from the Physical Influences on Brain in Aging (PHIBRA) Study. *Front Aging Neurosci* **8** https://doi.org/10.3389/fnagi.2016.00336 | PubMed

67. **Tetereva A,** Pat N (2024) Brain age has limited utility as a biomarker for capturing fluid cognition in older individuals. *eLife* **12**:RP87297 https://doi.org/10.7554/eLife.87297 | PubMed

68. **Nyberg L,** Lovden M, Riklund K, Lindenberger U, Backman L (2012) Memory aging and brain maintenance. *Trends Cogn Sci* **16**:292-305 https://doi.org/10.1016/j.tics.2012.04.005 | PubMed

69. **Karalija N,** Wahlin A, Ek J, Rieckmann A, Papenberg G, Salami A, et al. (2019) Cardiovascular factors are related to dopamine integrity and cognition in aging. *Ann Clin Transl Neurol* **6**:2291-303 https://doi.org/10.1002/acn3.50927 | PubMed

70. **Backman L,** Lindenberger U, Li SC, Nyberg L (2010) Linking cognitive aging to alterations in dopamine neurotransmitter functioning: recent data and future avenues. *Neurosci Biobehav Rev* **34**:670-7 https://doi.org/10.1016/j.neubiorev.2009.12.008 | PubMed

71. **Bäckman L,** Nyberg L, Soveri A, Johansson J, Andersson M, Dahlin E, et al. (2011) Effects of working-memory training on striatal dopamine release. *Science* **333**:718 https://doi.org/10.1126/science.1204978 | PubMed

72. **Volkow ND,** Gur RC, Wang GJ, Fowler JS, Moberg PJ, Ding YS, et al. (1998) Association between decline in brain dopamine activity with age and cognitive and motor impairment in healthy individuals. *Am J Psychiatry* **155**:344-9 https://doi.org/10.1176/ajp.155.3.344 | PubMed

73. **Juarez EJ,** Samanez-Larkin GR. (2019) Exercise, Dopamine, and Cognition in Older Age. *Trends Cogn Sci* **23**:986-8 https://doi.org/10.1016/j.tics.2019.10.006 | PubMed

74. **de Boer L,** Axelsson J, Riklund K, Nyberg L, Dayan P, Backman L, et al. (2017) Attenuation of dopamine-modulated prefrontal value signals underlies probabilistic reward learning deficits in old age. *eLife* **6** https://doi.org/10.7554/elife.26424 | PubMed

75. **Nyberg L,** Karalija N, Salami A, Andersson M, Wahlin A, Kaboovand N, et al. (2016) Dopamine D2 receptor availability is linked to hippocampal-caudate functional connectivity and episodic memory. *Proc Natl Acad Sci U S A* **113**:7918-23 https://doi.org/10.1073/pnas.1606309113 | PubMed

76. **Nordin K,** Nyberg L, Andersson M, Karalija N, Riklund K, Backman L, et al. (2021) Distinct and Common Large-Scale Networks of the Hippocampal Long Axis in Older Age: Links to Episodic Memory and Dopamine D2 Receptor Availability. *Cereb Cortex* **31**:3435-50 https://doi.org/10.1093/cercor/bhab023 | PubMed

77. **Papenberg G,** Karalija N, Salami A, Rieckmann A, Andersson M, Axelsson J, et al. (2020) Balance between Transmitter Availability and Dopamine D2 Receptors in Prefrontal Cortex Influences Memory Functioning. *Cereb Cortex* **30**:989-1000 https://doi.org/10.1093/cercor/bhz142 | PubMed

78. **Cools R,** D'Esposito M (2011) Inverted-U-shaped dopamine actions on human working memory and cognitive control. *Biol Psychiatry* **69**:e113-25 https://doi.org/10.1016/j.biopsych.2011.03.028 | PubMed

79. **Zahrt J,** Taylor JR, Mathew RG, Arnsten AF (1997) Supranormal stimulation of D1 dopamine receptors in the rodent prefrontal cortex impairs spatial working memory performance. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **17**:8528-35 https://doi.org/10.1523/jneurosci.17-21-08528.1997 | PubMed

80. **Karalija N,** Papenberg G, Johansson J, Wahlin A, Salami A, Andersson M, et al. (2024) Longitudinal support for the correlative triad among aging, dopamine D2-like receptor loss, and memory decline. *Neurobiol Aging* **136**:125-32 https://doi.org/10.1016/j.neurobiolaging.2024.02.001 | PubMed

81. **Servan-Schreiber D,** Printz H, Cohen JD (1990) A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science* **249**:892-5 https://doi.org/10.1126/science.2392679 | PubMed

82. **Seamans JK,** Yang CR (2004) The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Prog Neurobiol* **74**:1-58 https://doi.org/10.1016/j.pneurobio.2004.05.006 | PubMed

83. **Guitart-Masip M,** Salami A, Garrett D, Rieckmann A, Lindenberger U, Backman L (2016) BOLD Variability is Related to Dopaminergic Neurotransmission and Cognitive Aging. *Cereb Cortex* **26**:2074-83 https://doi.org/10.1093/cercor/bhv029 | PubMed

84. **Li SC,** Rieckmann A (2014) Neuromodulation and aging: implications of aging neuronal gain control on cognition. *Curr Opin Neurobiol* **29**:148-58 https://doi.org/10.1016/j.conb.2014.07.009 | PubMed

85. **Pedersen R,** Johansson J, Salami A (2023) Dopamine D1-signaling modulates maintenance of functional network segregation in aging. *Aging Brain* **3** https://doi.org/10.1016/j.nbas.2023.100079 | PubMed

86. **Shafiei G,** Zeighami Y, Clark CA, Coull JT, Nagano-Saito A, Leyton M, et al. (2019) Dopamine Signaling Modulates the Stability and Integration of Intrinsic Brain Networks. *Cereb Cortex* **29**:397-409 https://doi.org/10.1093/cercor/bhy264 | PubMed

87. **Fan L,** Su J, Qin J, Hu D, Shen H (2020) A Deep Network Model on Dynamic Functional Connectivity With Applications to Gender Classification and Intelligence Prediction. *Front Neurosci* **14** https://doi.org/10.3389/fnins.2020.00881 | PubMed

88. **Nevalainen N,** Riklund K, Andersson M, Axelsson J, Ogren M, Lovden M, et al. (2015) COBRA: A prospective multimodal imaging study of dopamine, brain structure and function, and cognition. *Brain Res* **1612**:83-103 https://doi.org/10.1016/j.brainres.2014.09.010 | PubMed

89. **Power JD,** Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, et al. (2011) Functional network organization of the human brain. *Neuron* **72**:665-78 https://doi.org/10.1016/j.neuron.2011.09.006 | PubMed

90. **Huang G,** Liu Z, van der Maaten L, Weinberger KQ (2016) Densely Connected Convolutional Networks. *arXiv* https://doi.org/10.48550/arxiv.1608.06993

91. **Zhao W,** Makowski C, Hagler DJ, Garavan HP, Thompson WK, Greene DJ, et al. (2023) Task fMRI paradigms may capture more behaviorally relevant information than resting-state functional connectivity. *Neuroimage* **270** https://doi.org/10.1016/j.neuroimage.2023.119946 | PubMed

92. **Schaefer A,** Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, et al. (2018) Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex* **28**:3095-114 https://doi.org/10.1093/cercor/bhx179 | PubMed

93. **Kohncke Y,** Papenberg G, Jonasson L, Karalija N, Wahlin A, Salami A, et al. (2018) Self-rated intensity of habitual physical activities is positively associated with dopamine D(2/3) receptor availability and cognition. *Neuroimage* **181**:605-16 https://doi.org/10.1016/j.neuroimage.2018.07.036 | PubMed

94. **Kurkela K,** Ritchey M (2024) Intrinsic functional connectivity among memory networks does not predict individual differences in narrative recall. *Imaging Neuroscience* **2**:1-17 https://doi.org/10.1162/imag_a_00169 | PubMed

95. **Finn ES** (2021) Is it time to put rest to rest?. *Trends in Cognitive Sciences* **25**:1021-32 https://doi.org/10.1016/j.tics.2021.09.005 | PubMed

96. **Zhu Y,** Zang F, Wang Q, Zhang Q, Tan C, Zhang S, et al. (2021) Connectome-based model predicts episodic memory performance in individuals with subjective cognitive decline and amnestic mild cognitive impairment. *Behav Brain Res* **411** https://doi.org/10.1016/j.bbr.2021.113387 | PubMed

97. **Wahlheim CN,** Christensen AP, Reagh ZM, Cassidy BS (2022) Intrinsic functional connectivity in the default mode network predicts mnemonic discrimination: A connectome-based modeling approach. *Hippocampus* **32**:21-37 https://doi.org/10.1002/hipo.23393 | PubMed

98. **Gonzalez-Castillo J,** Kam JWY, Hoy CW, Bandettini PA (2021) How to Interpret Resting-State fMRI: Ask Your Participants. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **41**:1130-41 https://doi.org/10.1523/jneurosci.1786-20.2020 | PubMed

99. **Stawarczyk D,** D'Argembeau A (2015) Neural correlates of personal goal processing during episodic future thinking and mind-wandering: An ALE meta-analysis. *Hum Brain Mapp* **36**:2928-47 https://doi.org/10.1002/hbm.22818 | PubMed

100. **Blonde P,** Sperduti M, Makowski D, Piolino P (2022) Bored, distracted, and forgetful: The impact of mind wandering and boredom on memory encoding. *Q J Exp Psychol* **75**:53-69 https://doi.org/10.1177/17470218211026301 | PubMed

101. **Karapanagiotidis T,** Bernhardt BC, Jefferies E, Smallwood J (2017) Tracking thoughts: Exploring the neural architecture of mental time travel during mind-wandering. *Neuroimage* **147**:272-81 https://doi.org/10.1016/j.neuroimage.2016.12.031 | PubMed

102. **Salami A,** Garrett DD, Wahlin A, Rieckmann A, Papenberg G, Karalija N, et al. (2019) Dopamine D(2/3) Binding Potential Modulates Neural Signatures of Working Memory in a Load-Dependent Fashion. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **39**:537-47 https://doi.org/10.1523/jneurosci.1493-18.2018 | PubMed

103. **Salami A,** Rieckmann A, Karalija N, Avelar-Pereira B, Andersson M, Wahlin A, et al. (2018) Neurocognitive Profiles of Older Adults with Working-Memory Dysfunction. *Cereb Cortex* **28**:2525-39 https://doi.org/10.1093/cercor/bhy062 | PubMed

104. **Eriksson J,** Vogel EK, Lansner A, Bergstrom F, Nyberg L (2015) Neurocognitive Architecture of Working Memory. *Neuron* **88**:33-46 https://doi.org/10.1016/j.neuron.2015.09.020 | PubMed

105. **Liang X,** Zou Q, He Y, Yang Y (2016) Topologically Reorganized Connectivity Architecture of Default-Mode, Executive-Control, and Salience Networks across Working Memory Task Loads. *Cereb Cortex* **26**:1501-11 https://doi.org/10.1093/cercor/bhu316 | PubMed

106. **Gilson M,** Deco G, Friston KJ, Hagmann P, Mantini D, Betti V, et al. (2018) Effective connectivity inferred from fMRI transition dynamics during movie viewing points to a balanced reconfiguration of cortical interactions. *Neuroimage* **180**:534-46 https://doi.org/10.1016/j.neuroimage.2017.09.061 | PubMed

107. **Betzel RF,** Byrge L, Esfahlani FZ, Kennedy DP (2020) Temporal fluctuations in the brain's modular architecture during movie-watching. *Neuroimage* **213** https://doi.org/10.1016/j.neuroimage.2020.116687 | PubMed

108. **Zhang H,** Zhao R, Hu X, Guan S, Margulies DS, Meng C, et al. (2022) Cortical connectivity gradients and local timescales during cognitive states are modulated by cognitive loads. *Brain Struct Funct* **227**:2701-12 https://doi.org/10.1007/s00429-022-02564-0 | PubMed

109. **Ito T,** Brincat SL, Siegel M, Mill RD, He BJ, Miller EK, et al. (2020) Task-evoked activity quenches neural correlations and variability across cortical areas. *PLoS Comput Biol* **16**:e1007983 https://doi.org/10.1371/journal.pcbi.1007983 | PubMed

110. **Tian YE,** Cropley V, Maier AB, Lautenschlager NT, Breakspear M, Zalesky A (2023) Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nat Med* **29**:1221-31 https://doi.org/10.1038/s41591-023-02296-6 | PubMed

111. **Arida RM,** Teixeira-Machado L (2020) The Contribution of Physical Exercise to Brain Resilience. *Front Behav Neurosci* **14** https://doi.org/10.3389/fnbeh.2020.626769 | PubMed

112. **Klil-Drori S,** Cinalioglu K, Rej S (2022) Brain Health and the Role of Exercise in Maintaining Late-Life Cognitive Reserve: A Narrative Review Providing the Neuroprotective Mechanisms of Exercise. *The American Journal of Geriatric Psychiatry* **30**:S72 https://doi.org/10.1016/j.jagp.2022.01.067

113. **Nithianantharajah J,** Hannan AJ (2009) The neurobiology of brain and cognitive reserve: Mental and physical activity as modulators of brain disorders. *Progress in Neurobiology* **89**:369-82 https://doi.org/10.1016/j.pneurobio.2009.10.001 | PubMed

114. **Cabeza R,** Albert M, Belleville S, Craik FIM, Duarte A, Grady CL, et al. (2018) Maintenance, reserve and compensation: the cognitive neuroscience of healthy ageing. *Nat Rev Neurosci* **19**:701-10 https://doi.org/10.1038/s41583-018-0068-2 | PubMed

115. **Korkki SM,** Johansson J, Nordin K, Pedersen R, Bäckman L, Rieckmann A, et al. (2024) Dedifferentiation of caudate functional organization is linked to reduced D1 dopamine receptor availability and poorer memory function in aging. *bioRxiv* https://doi.org/10.1101/2024.03.18.585623

116. **Berry AS,** Shah VD, Furman DJ, White RL, Baker SL, O'Neil JP, et al. (2018) Dopamine Synthesis Capacity is Associated with D2/3 Receptor Binding but Not Dopamine Release. *Neuropsychopharmacology* **43**:1201-11 https://doi.org/10.1038/npp.2017.180 | PubMed

117. **Volkow ND,** Wang GJ, Fowler JS, Ding YS, Gur RC, Gatley J, et al. (1998) Parallel loss of presynaptic and postsynaptic dopamine markers in normal aging. *Ann Neurol* **44**:143-7 https://doi.org/10.1002/ana.410440125 | PubMed

118. **Kienast T,** Siessmeier T, Wrase J, Braus DF, Smolka MN, Buchholz HG, et al. (2008) Ratio of dopamine synthesis capacity to D2 receptor availability in ventral striatum correlates with central processing of affective stimuli. *Eur J Nucl Med Mol Imaging* **35**:1147-58 https://doi.org/10.1007/s00259-007-0683-z | PubMed

119. **Heinz A,** Siessmeier T, Wrase J, Buchholz HG, Grunder G, Kumakura Y, et al. (2005) Correlation of alcohol craving with striatal dopamine synthesis capacity and D2/3 receptor availability: a combined [18F]DOPA and [18F]DMFP PET study in detoxified alcoholic patients. *Am J Psychiatry* **162**:1515-20 https://doi.org/10.1176/appi.ajp.162.8.1515 | PubMed

120. **Ito H,** Kodaka F, Takahashi H, Takano H, Arakawa R, Shimada H, et al. (2011) Relation between presynaptic and postsynaptic dopaminergic functions measured by positron emission tomography: implication of dopaminergic tone. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **31**:7886-90 https://doi.org/10.1523/jneurosci.6024-10.2011 | PubMed

121. **Nyberg L,** Andersson M, Lundquist A, Baare WFC, Bartres-Faz D, Bertram L, et al. (2023) Individual differences in brain aging: heterogeneity in cortico-hippocampal but not caudate atrophy rates. *Cereb Cortex* **33**:5075-81 https://doi.org/10.1093/cercor/bhac400 | PubMed

122. **D'Agostino RB Sr,** Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. (2008) General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**:743-53 https://doi.org/10.1161/circulationaha.107.699579 | PubMed

123. **Johansson J,** Nordin K, Pedersen R, Karalija N, Papenberg G, Andersson M, et al. (2023) Biphasic patterns of age-related differences in dopamine D1 receptors across the adult lifespan. *Cell Rep* **42**:113107 https://doi.org/10.1016/j.celrep.2023.113107 | PubMed

124. **Hallquist MN,** Hwang K, Luna B (2013) The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity. *Neuroimage* **82**:208-25 https://doi.org/10.1016/j.neuroimage.2013.05.116 | PubMed

125. **Friston KJ,** Williams S, Howard R, Frackowiak RS, Turner R (1996) Movement-related effects in fMRI time-series. *Magn Reson Med* **35**:346-55 https://doi.org/10.1002/mrm.1910350312 | PubMed

126. **Power JD,** Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE (2012) Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**:2142-54 https://doi.org/10.1016/j.neuroimage.2011.10.018 | PubMed

127. **Glover GH,** Li TQ, Ress D (2000) Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magn Reson Med* **44**:162-7 https://doi.org/10.1002/1522-2594(200007)44:1<162::aid-mrm23>3.0.co;2-e | PubMed

128. **Hutton C,** Josephs O, Stadler J, Featherstone E, Reid A, Speck O, et al. (2011) The impact of physiological noise correction on fMRI at 7 T. *Neuroimage* **57**:101-12 https://doi.org/10.1016/j.neuroimage.2011.04.018 | PubMed

129. **Kasper L,** Bollmann S, Diaconescu AO, Hutton C, Heinzle J, Iglesias S, et al. (2017) The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *J Neurosci Methods* **276**:56-72 https://doi.org/10.1016/j.jneumeth.2016.10.019 | PubMed

130. **Fair DA,** Schlaggar BL, Cohen AL, Miezin FM, Dosenbach NU, Wenger KK, et al. (2007) A method for using blocked and event-related fMRI data to study "resting state" functional connectivity. *Neuroimage* **35**:396-405 https://doi.org/10.1016/j.neuroimage.2006.11.051 | PubMed

131. **Cole MW,** Ito T, Schultz D, Mill R, Chen R, Cocuzza C (2019) Task activations produce spurious but systematic inflation of task functional connectivity estimates. *Neuroimage* **189**:1-18 https://doi.org/10.1016/j.neuroimage.2018.12.054 | PubMed

132. **Ashburner J** (2007) A fast diffeomorphic image registration algorithm. *Neuroimage* **38**:95-113 https://doi.org/10.1016/j.neuroimage.2007.07.007 | PubMed

133. **Fonov V,** Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL, et al. (2011) Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage* **54**:313-27 https://doi.org/10.1016/j.neuroimage.2010.07.033 | PubMed

134. **Fonov VS,** Evans AC, McKinstry RC, Almli CR, Collins DL (2009) Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47** https://doi.org/10.1016/s1053-8119(09)70884-5

135. **Salami A,** Adolfsson R, Andersson M, Blennow K, Lundquist A, Adolfsson AN, et al. (2022) Association of APOE varepsilon4 and Plasma p-tau181 with Preclinical Alzheimer's Disease and Longitudinal Change in Hippocampus Function. *J Alzheimers Dis* **85**:1309-20 https://doi.org/10.3233/jad-210673 | PubMed

136. **Janiesch C,** Zschech P, Heinrich K (2021) Machine learning and deep learning. *Electronic Markets* **31**:685-95 https://doi.org/10.1007/s12525-021-00475-2

137. **He K,** Zhang X, Ren S, Sun J (2015) Deep Residual Learning for Image Recognition. *arXiv* https://doi.org/10.48550/arxiv.1512.03385

138 **Nair V,** Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. pp. 807-14

139. **Developers T** (no date) TensorFlow. v2.5.0 edition. https://www.tensorflow.org/

140. **Chollet F** (no date) Keras. 2.5.0 edition. https://keras.io

141. **Sutskever I,** Martens J, Dahl G, Hinton G, Sanjoy D, David M (2013) On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th International Conference on Machine Learning. pp. 1139-47

142. **Srivastava N,** Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**:1929-58

143. **Selvaraju RR,** Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* **128**:336-59 https://doi.org/10.1007/s11263-019-01228-7

144. **Greve DN,** Salat DH, Bowen SL, Izquierdo-Garcia D, Schultz AP, Catana C, et al. (2016) Different partial volume correction methods lead to different conclusions: An (18)F-FDG-PET study of aging. *Neuroimage* **132**:334-43 https://doi.org/10.1016/j.neuroimage.2016.02.042 | PubMed

145. **Lammertsma AA,** Hume SP (1996) Simplified reference tissue model for PET receptor studies. *Neuroimage* **4**:153-8 https://doi.org/10.1006/nimg.1996.0066 | PubMed

146. **Baron RM,** Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* **51**:1173-82 https://doi.org/10.1037//0022-3514.51.6.1173 | PubMed

147. **Preacher KJ,** Hayes AF (2004) SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav Res Methods Instrum Comput* **36**:717-31 https://doi.org/10.3758/bf03206553 | PubMed

148. **Poldrack RA,** Huckins G, Varoquaux G (2020) Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry* **77**:534-40 https://doi.org/10.1001/jamapsychiatry.2019.3671 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors attempted to identify if a new deep learning model could be applied to both resting and task state fMRI data to predict cognition and dopaminergic signaling. They found that resting state and moving watching conditions best predict episodic memory, but only movie watching predicts both episodic and working memory. A negative 'brain gap' (where the model trained on brain connectivity predicts worse performance than what is actually observed) was associated with less physical activity, poorer cardiovascular function, and lower D1R availability.

Strengths:

The paper should be of broad interest to the journal's readership, with implications for cognitive neuroscience, psychiatry, and psychology fields. The paper is very well-written and clear. The authors use two independent datasets to validate their findings, including two of the largest databases of dopamine receptor availability to link brain functional connectivity/activity with neurochemical signaling.

Weaknesses:

The deep learning findings represent a relatively small extension/enhancement of knowledge in a very crowded field.

It's unclear from these results how much utility the brain gaps provide above and beyond observed performance. It would be helpful to take a median split the dataset on observed performance, and plot aside the current Fig 3 results to see how the cardiovascular and physical activity measures differ based on actual performance. Could the authors perform additional analyses describing how much additional variance is explained in these measures by including brain gaps?

Some of the imaging findings require deeper analysis. For figure 1f - Which default mode regions have high salience? DMN is a huge network with subregions having differing functions.

Along the same lines, were the striatal D1R findings regionally specific at all? It would be informative to test whether the three nuclei (Accumbens, Caudate, Putamen) and/or voxelwise models would show something above and beyond what is achieved from averaging D1R across the striatum. What about cortical D1R, which are highly abundant, strongly associated with cognitive (especially WM) performance, and have much unique variance beyond striatal D1R? *https://www.science.org/doi/full/10.1126/sciadv.1501672*. The PET findings are one of the unique strengths of this paper and are underexplored. It's also unclear if the measure of brain entropy should simply be averaged across all regions.

It is not clear from the text that the authors met the preconditions for mediation analysis (that is, demonstrating significant correlations between D1R and entropy, in addition to the correlation with brain gap. Could they please report this as well?

Was age controlled for in the mediation analysis? I would not consider this result valid unless that is the case.

The discussion is long, but the authors would do better to replace some less helpful sections (e.g., the paragraph on methodological tweaks to parcellations and model alignment) with a couple of other important points, including:

(1) Discuss the 'sweet-spot' of movie watching for behavior prediction in the context of studies showing that task states 'quench' neural variability: *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007983*. This may not be mutually exclusive of the discussion on dopamine and signal-to-noise ratio, but it would be helpful for the authors to discuss their potential overlap vs. unique contributions to the observed findings.

(2) The argument that dopamine signaling increases signal-to-noise ratio is based on some preclinical data as well as correlational data using fMRI with pharmacological challenges. It is less clear how PET-derived estimates of D1R and D2R availability equate to 'dopamine signaling' as it is thought of in this context. Presumably, based on these data, higher D1R or D2R availability would be related to greater levels of tonic dopaminergic signaling. However, in the case of the COBRA dataset with D2R estimates, those are based on raclopride -- which competes with endogenous dopamine for the D2 receptor. Therefore, someone with higher levels of endogenous dopamine signaling should theoretically have lower raclopride binding and lower D2R estimates. I'm not arguing that the authors logic is flawed or that D1R and D2R are not good measures of dopamine signaling, but I'd ask the authors to dig into the literature and describe more direct potential links for how greater receptor availability might be associated with greater dopamine signaling (and hence lower entropy). Adding this to the discussion would be very valuable for PET research.

Comments on revised version:

I thank the authors for their extensive efforts to revise the manuscript. I have no further concerns.

https://doi.org/10.7554/eLife.104053.2.sa2

Reviewer #2 (Public review):

The authors have made several corrections to the original manuscript. For example, they revised the bootstrapping analysis to avoid arbitrarily inflating the degrees of freedom. However, most substantive concerns remain inadequately addressed.

(1) The primary issue is still the lack of baseline models against which to benchmark the predictive performance of the proposed DenseNet model. This concern was raised independently by two reviewers. Without such benchmarks, it is difficult to interpret the reported results in the context of prior work on MRI-based cognition prediction.

Notably, the authors state: "While we compared our model with the connectome predictive modeling (CPM) approach and observed better performance with our deep learning framework, we did not conduct a comprehensive benchmark across all available machine learning methods, nor was this the aim of the present study."

However, I could NOT find any discussion or results related to the CPM model in the manuscript. It is therefore unclear whether the DenseNet model was actually statistically compared with CPM, and, if so, how the comparison was conducted.

Note that the statement, "While Vieira et al. show that the majority (76%) of prior studies used linear modeling approaches, including CPM and penalized regressions, these models are often vulnerable to overfitting, especially when applied to high-dimensional fMRI data," is not entirely accurate. Linear models typically have far fewer parameters than deep-learning models and are therefore often less prone to overfitting. In fact, it is well established that deep-learning models are particularly susceptible to overfitting and usually require substantially larger sample sizes to achieve stable and reliable performance. Although deep-learning models may outperform shallower models once sufficient data are available and training is well controlled, this does not justify the authors' claim as stated. I therefore disagree with the argument put forward by the authors.

The authors further justify the absence of benchmarking by stating: "In this context, deep learning was employed as a flexible framework capable of modelling high-dimensional functional connectivity patterns across cognitive states, rather than as a claim of inherent methodological superiority. Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach." However, most shallow models can likewise be applied across different brain states and cognitive targets. This rationale does not establish deep learning as a uniquely appropriate or necessary choice. If deep learning is indeed a better approach in this context, the authors should demonstrate this empirically through appropriate benchmarking against established baseline models.

(2) Additional analysis shows that "BCG is not significantly associated with cognition itself". This is the most perplexing result. This is like saying Brain Age Gap is not related to chronological Age. It is counterintuitive since the Brain Age Gap is calculated by chronological age minus actual age, and most research has shown a strong relationship between the Brain Age Gap and age.

If the brain cognition gap is not related to cognition, is it possible that the results found are mainly due to the predictive model not fitting well with another dataset? Regardless, the lack of association between BCG and cognition deserves a discussion.

(3) I still do not fully understand the rationale of the mediation analysis. The analysis and findings are still not related to aims 1 and 2, since DA and entropy are not part of the prediction models. But I appreciate the explanation that this part is related to the authors' previous work, and that the authors attempted to link to them somehow.

https://doi.org/10.7554/eLife.104053.2.sa1

Author response:

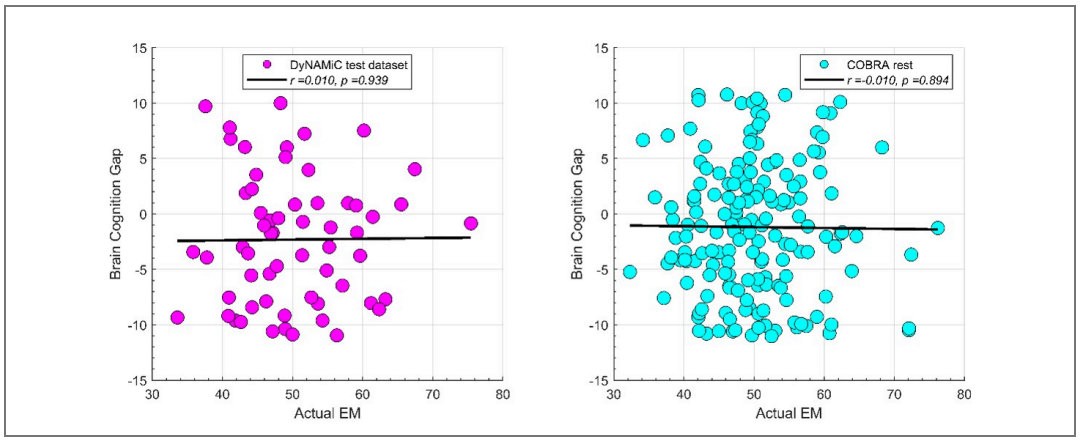
The following is the authors' response to the original reviews.

In the revised version, our primary focus has been to more clearly demonstrate the unique contribution of the brain-cognitive gap (BCG) beyond what is captured by cognitive performance alone, and to show that the BCG is not trivially driven by the observed cognitive scores. Additional analyses now demonstrate that the BCG provides complementary and nuanced information regarding factors associated with cognitive resilience, above and beyond the cognitive measures themselves.

In response to the comment regarding the inclusion of a baseline predictive model, we would like to clarify that the central aim of our study is to compare predictive utility across different cognitive states (resting state, movie watching, and n-back), rather than to establish a single universally optimal prediction model. Several previous studies have already systematically compared deep learning approaches with more traditional machine learning methods for functional connectome-based prediction. In contrast, the goal of the present study is to examine how brain state modulates the ability of AI-based functional connectome models to capture individual differences in working memory and episodic memory.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *The authors attempted to identify whether a new deep-learning model could be applied to both resting and task state fMRI data to predict cognition and dopaminergic signaling. They found that resting state and moving watching conditions best predict episodic memory, but only movie watching predicts both episodic and working memory. A negative 'brain gap' (where the model trained on brain connectivity predicts worse performance than what is actually observed) was associated with less physical activity, poorer cardiovascular function, and lower D1R availability.*
>
> *Strengths:*
>
> *The paper should be of broad interest to the journal's readership, with implications for cognitive neuroscience, psychiatry, and psychology fields. The paper is very well-written and clear. The authors use two independent datasets to validate their findings, including two of the largest databases of dopamine receptor availability to link brain functional connectivity/activity with neurochemical signaling.*
>
> *Weaknesses:*
>
> *The deep learning findings represent a relatively small extension/enhancement of knowledge in a very crowded field.*
>
> *It's unclear from these results how much utility the brain gaps provide above and beyond observed performance. It would be helpful to take a median split of the dataset on observed performance and plot aside the current Figure 3 results to see how the cardiovascular and physical activity measures differ based on actual performance. Could the authors perform additional analyses describing how much additional variance is explained in these measures by including brain gaps?*

We thank the reviewer for raising this important point. In response to their request, we first examined the relationship between the BCG and the cognitive measure itself. We did not find any significant relationship in either the DyNAMiC sample (r =0.01, p =0.939) or the COBRA sample ((r =0.01, p=0.894) (see Author response image 1).



Author response image 1.

We then conducted additional analyses, splitting the sample into high and low EM performers, and compared their levels of physical activity and Framingham cardiovascular disease (CVD) risk scores. We found no significant difference in physical activity (DyNAMiC: p =0.56, 95% CI: –14.99 - 8.13; COBRA: p =0.29, 95% CI: –3.54 - 1.05) or Framingham CVD risk score (DyNAMiC: p =0.11, 95% CI: –1.08 - 10.72; COBRA: p =0.41, 95% CI: –1.86 - 4.58) between high and low EM perfprmers. Given the significant difference in physical activity and Framingham CVD risk score between positive and negative BCG groups, our results support that BCG provides unique information, beyond the observed cognitive measure (episodic memory score), regarding factors that contribute to cognitive resilience. These results have been added to Section 2.4, and Figure 3 has been updated.

> *Some of the imaging findings require deeper analysis. For Figure 1f - Which default mode regions have high salience? DMN is a huge network with subregions having differing functions.*

Grad-CAM provides a coarse, gradient-based attribution that reflects how the learned feature maps contribute to the model output. It is not designed to produce specific input-level interpretations, such as symmetric edge-wise importance values. Therefore, the primary interpretation remains at the network level rather than at the level of individual FC edges.

> *Along the same lines, were the striatal D1R findings regionally specific at all? It would be informative to test whether the three nuclei (Accumbens, Caudate, Putamen) and/or voxelwise models would show something above and beyond what is achieved from averaging D1R across the striatum. What about cortical D1R, which is highly abundant, strongly associated with cognitive (especially WM) performance, and has much unique variance beyond striatal D1R? https://www.science.org/doi/full/10.1126/sciadv.1501672. The PET findings are one of the unique strengths of this paper and are underexplored. It's also unclear if the measure of brain entropy should simply be averaged across all regions.*

In this study, we focused on D1DR/ D2DR averaged across the caudate and putamen, which has been reported in our previous work to be more strongly associated with cognitive functions (Johansson et al., 2023, Nyberg et al., 2016), compared to the nucleus Accumbens,

which tends to show lower D1DR/D2DR levels and limited association with these cognitive domains. Following the Reviewer's suggestion, we examined regional variations and found that while both caudate and putamen D1DR showed significant associations with BCG, there were no significant associations for D1DR in the nucleus accumbens or DLPFC with BCG. For D2DR, we observed a significant association between caudate/putamen D2DR and BCG.

D1DR:

Partial correlation between:

Caudate_Bilateral vs. NegGap, (r =0.37, p =0.02

Putamen_Bilateral vs. NegGap, r =0.34, p =0.03

Accumbens_Bilateral vs. NegGap, r =0.07, p =0.69

Mean (LRCaud, LRput, LRacc) vs NegGap, r =0.35, p =0.03

DLPFC_Bilateral vs NegGap, r =0.21, p =0.21

Striatum_Bilateral (Mean (LRCaud, LRput)) vs. NegGap, r =0.40, p =0.01

Caudate_Bilateral vs. PosGap, r=–0.37, p=0.02

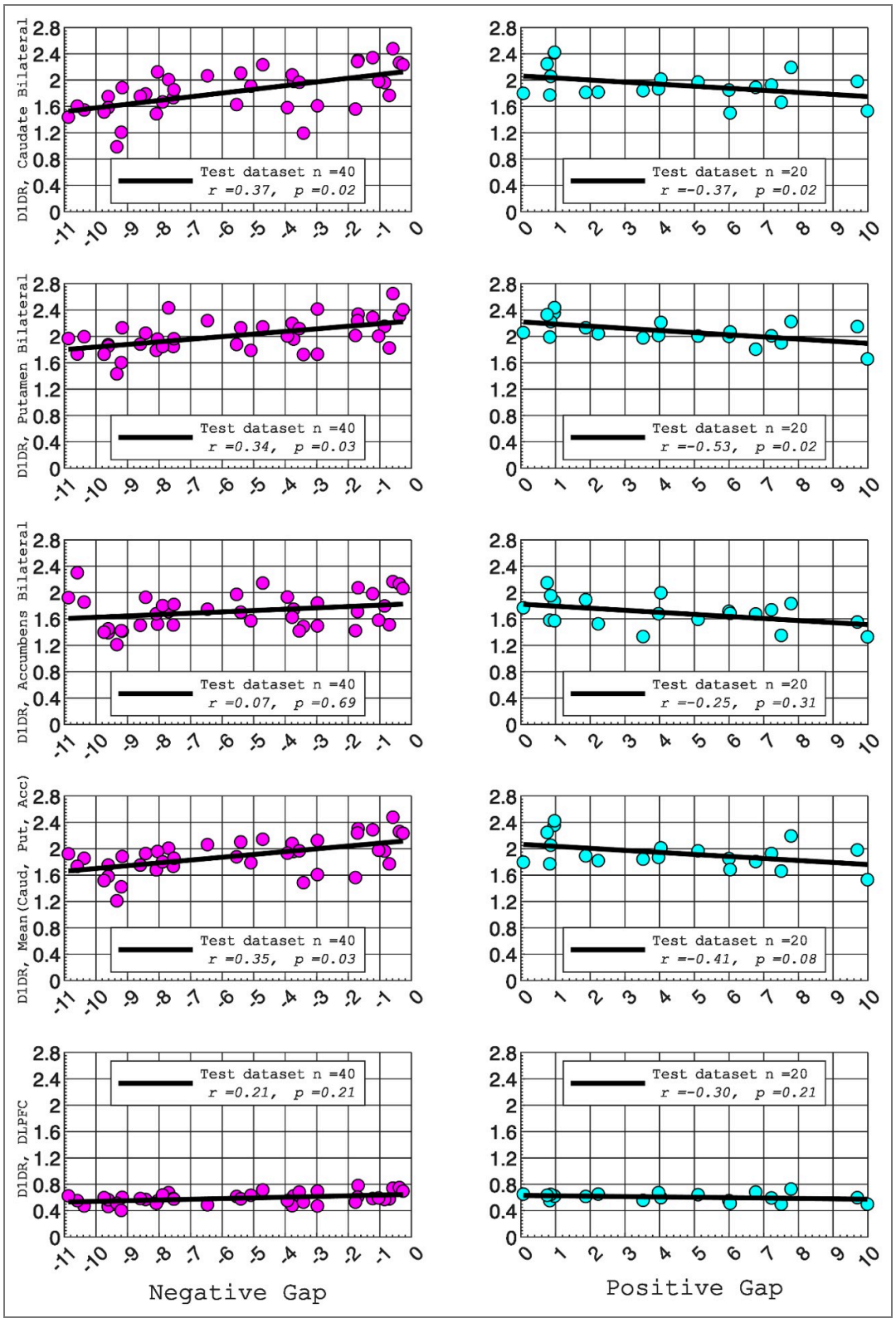
Putamen_Bilateral vs. PosGap, r=–0.53, p=0.02

Accumbens_Bilateral vs. PosGap, r=–0.25, p=0.31

Mean (LRCaud, LRput, LRacc) vs PosGap, r=–0.41, p=0.08

DLPFC_Bilateral vs. PosGap, r=–0.30, p=0.21

Striatum_Bilateral (Mean (LRCaud, LRput)) vs. PosGap, r=–0.49, p=0.03



Author response image 2.

D2DR:

Correlation between:

Caudate_Bilateral vs. NegGap, r=0.36, p=0.0003

Putamen_Bilateral vs. NegGap, r=0.22, p=0.03

Accumbens_Bilateral vs. NegGap, r= –0.01, p=0.91

Mean (LRCaud, LRput, LRacc) vs PosGap, r= –0.24, p=0.01

Striatum_Bilateral vs. NegGap, r=0.39, p=0.0001

Caudate_Bilateral vs. PosGap, r= –0.34, p=0.004

Putamen_Bilateral vs. PosGap, r= –0.37, p=0.002

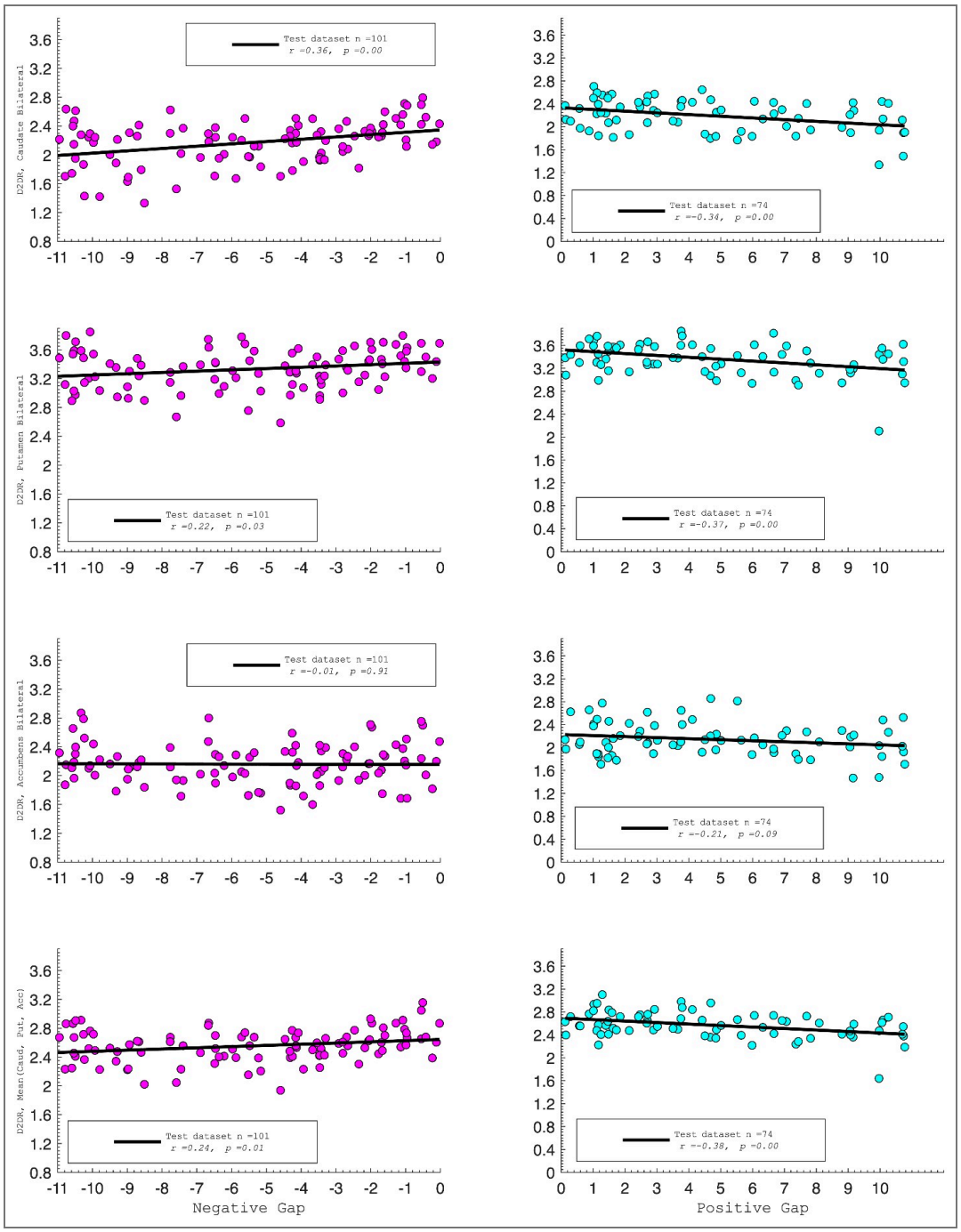
Accumbens_Bilateral vs. PosGap, r= –0.21, p=0.09

Mean (LRCaud, LRput, LRacc) vs PosGap, r= –0.38, p=0.001

Striatum_Bilateral vs. PosGap, r= –0.49, p=0.0001

We have added the following sentence to the Results section to highlight these regional differences in D1DR/D2DR in relation to BCG.

“Both D1DR and D2DR availability in the striatum were associated with BCG, such that lower dopamine receptor availability was linked to a greater behavioral-cognitive gap. However, these associations varied by region. For D1DR, significant correlations with BCG were observed in the caudate (positive gap: r = –0.37, p =0.02; negative gap: r= 0.37, p =0.02) and putamen (positive gap: r = –0.53, p=0.02; negative gap:r=0.34, p=0.03), but not in the nucleus accumbens (positive gap: r= –0.25, p= 0.31; negative gap: r =0.07, p=0.69) or the DLPFC (positive gap: r = –0.30, p=0.21; negative gap: r =0.21, p=0.21). For D2DR, both caudate (positive gap: r = –0.34, p=0.004; negative gap: r =0.36, p=0.0003) and putamen (positive gap: r = –0.37, p=0.002; negative gap: r =0.22, p=0.03) showed significant associations with BCG.”



Author response image 3.

> *It is not clear from the text that the authors met the preconditions for mediation analysis (that is, demonstrating significant correlations between D1R and entropy, in addition to the correlation with brain gap. The authors should report this as well.*

This is a fair question. We recalculated entropy in the striatum, given that D1DR is more strongly expressed in this region and, therefore, reduced striatal D1DR may have a more pronounced impact on local entropy (as the reviewer suggested, it may not be appropriate to compute entropy across all brain regions). Our analyses showed that lower D1DR/D2DR levels were associated with higher entropy, which in turn was related to higher BCG.

DyNAMiC; negative gap:

Partial correlation between:

Entropy and D1DR, r = –0.33, p=0.04.

Entropy and NegGap, r = –0.36, p=0.03.

DyNAMiC; positive gap:

Partial correlation between:

Entropy and D1DR, r = –0.56, p=0.01.

Entropy and PosGap, r r =0.47, p=0.04.

COBRA; negative gap:

Correlation between:

Entropy and D2DR, r = –0.22, p=0.03.

Entropy and NegGap, r = –0.27, p=0.007.

COBRA; positive gap:

Correlation between:

Entropy and D2DR, r = –0.26, p=0.03.

Entropy and PosGap, r = 0.25, p=0.03.

We have added these results under the result section 2.6. We have further updated Figure 4 in the revised manuscript, reporting these correlation results.

> *Was age controlled for in the mediation analysis? I would not consider this result valid unless that is the case.*

We utilized the mediation package in *R*, and to control for a covariate age in the mediation analysis, we added age as a covariate in both the mediator model and the outcome model. The following information has been added in the method section in the revised version of the manuscript.

"To assess the statistical significance of this mediation effect, we employed the bootstrapping method as outlined by Preacher and Hayes (145) and age has been controlled for in all statistical analysis."

> *The discussion section is long, but the authors would do better to replace some less helpful sections (e.g., the paragraph on methodological tweaks to parcellations and model alignment) with a couple of other important points, including:*
>
> *(1) Discuss the 'sweet-spot' of movie watching for behavior prediction in the context of studies showing that task states 'quench' neural variability: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007983. This may not be mutually exclusive of the discussion on dopamine and signal-to-noise ratio, but it would be helpful for the authors to discuss their potential overlap vs. unique contributions to the observed findings.*

Thank you for the comment. We have now eliminated the section about methodological tweaks and extended the discussion on the sweet-spot of the task for behavioral prediction by referencing the paper that the reviewer suggested. Here comes the paragraph discussing this topic:

“Additionally, previous research showed that movie-watching alters the propagation of activity across cortical pathways (105), particularly within and between regions involved in audiovisual processing and attention. These alterations lead to a less segregated and more integrated network organization (106). Similarly, the n-back task has been associated with increased integration of task-positive cortico-cortical connectivity (104, 107) and striato-cortical connectivity (102). Our findings also suggest that certain task contexts strike an optimal balance between reducing neural variability and maintaining sufficient richness to capture individual differences. Prior work shows that task states quench neural variability, leading to a more reliable and predictable neural signal (108). In this context, movie watching may represent such a sweet spot constraining neural dynamics through shared audiovisual stimulation, while simultaneously engaging a broad range of cognitive processes that preserve individual differences.”

> *(2) The argument that dopamine signaling increases signal-to-noise ratio is based on some preclinical data as well as correlational data using fMRI with pharmacological challenges. It is less clear how PET-derived estimates of D1R and D2R availability equate to 'dopamine signaling' as it is thought of in this context. Presumably, based on these data, higher D1R or D2R availability would be related to greater levels of tonic dopaminergic signaling. However, in the case of the COBRA dataset with D2R estimates, those are based on raclopride -- which competes with endogenous dopamine for the D2 receptor. Therefore, someone with higher levels of endogenous dopamine signaling should theoretically have lower raclopride binding and lower D2R estimates. I'm not arguing that the authors' logic is flawed or that D1R and D2R are not good measures of dopamine signaling, but I'd ask the authors to dig into the literature and describe more direct potential links for how greater receptor availability might be associated with greater dopamine signaling (and hence lower entropy). Adding this to the discussion would be very valuable for PET research.*

Thank you for raising this important point. We agree that D1R and D2R availability should not be taken as direct proxies of dopamine signaling. However, prior work has suggested meaningful associations between pre- and post-synaptic markers. For instance, a well-powered study demonstrated a significant correlation between D2R availability and dopamine synthesis capacity measured by FMT (Berry et al., 2018). This finding supports the idea that postsynaptic receptor markers may, under certain conditions, serve as an indirect proxy for dopaminergic signaling. Moreover, the number of dopamine-producing neurons innervating the striatum during development has been proposed to shape the structural maturation and arborization of dendrites (McAllister, 2000; Whitford et al., 2002), potentially providing a structural and functional basis for observed associations between pre- and post-synaptic measures.

At the same time, smaller-scale studies have yielded mixed findings, reporting either non-significant associations (Heinz et al., 2005; Kienast et al., 2008) or negative correlations (Ito et al., 2011). Importantly, the latter studies employed [18F]FDOPA to index dopamine synthesis, which has been argued to provide a less reliable estimate of synthesis capacity compared to FMT, as used in Berry et al. (2018). These inconsistencies underscore that the relationship between pre- and post-synaptic markers is not straightforward and requires further examination in larger, well-powered samples. The following paragraph has been added to the discussion.

"An important caveat is that D1DR and D2DR availability do not provide a direct measure of dopamine signaling. Instead, they reflect receptor availability, which interacts with endogenous dopamine in a complex manner. PET measures of D1R and D2R availability reflect the density of unoccupied dopamine receptors and the degree to which endogenous dopamine competes with radioligand binding. D2R binding potential is sensitive to competition from synaptic dopamine, such that higher ambient dopamine generally reduces tracer binding; D1R binding, however, is less affected by endogenous dopamine under physiological conditions, reflecting more directly receptor expression levels. Previous studies demonstrated a significant association between D2R availability and dopamine synthesis capacity measured by FMT (117, 118), suggesting that postsynaptic receptor markers may, under certain conditions, serve as a proxy for dopaminergic signaling. Developmental factors, such as the number of dopamine-producing neurons innervating the striatum, may further influence the structural and functional relationship between pre- and post-synaptic markers. By contrast, smaller studies have reported non-significant (119, 120) or negative (121) associations, although these studies relied on [18F]FDOPA, which is considered a less precise index of dopamine synthesis than FMT. Taken together, these reports indicate that the relationship between pre- and post-synaptic markers is complex and not necessarily linear. Accordingly, our observation that lower receptor availability is associated with greater neural variability should not be interpreted as direct evidence of weaker dopaminergic signaling, but rather as reflecting the interplay between receptor density and endogenous dopamine occupancy, particularly in the case of D2DR."

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *The authors developed a deep learning model based on a DenseNet CNN architecture to predict two cognitive functions: working memory and episodic memory, from functional connectivity matrices. These matrices were recorded under three conditions: during rest, a working memory task, and a movie, and were treated as images for the CNN algorithm. They tested their model's performance across different conditions and a separate dataset with a different age distribution (using the same MRI scanner, scanning configurations, and cognitive tests). They also calculated the "brain cognition gap" based on the model trained on resting functional connectivity to predict working memory. Extending from the commonly used index "brain age," the brain cognition gap was defined as the difference between the working memory score predicted by their model (predicted working memory) and the working memory score based on the working memory test itself (observed working memory). This brain cognition gap was found to be associated with physical activity, education, and cardiovascular risk. The authors also conducted additional mediation tests to examine whether regional functional variability mediated the relationship between PET-derived measures of dopamine and the brain cognition gap.*
>
> *Strengths:*
>
> *The major strength of this manuscript is the extensive effort the authors have put into creating a new 'biomarker' that links deep learning with fMRI, PET, physical activity, education, and cardiovascular risk across two studies. This effort is impressive.*
>
> *Weaknesses:*
>
> *There are several weaknesses in the current methods and results, making many of the claims unconvincing. These weaknesses include:*
>
> *(1) The lack of baseline models to benchmark the predictive performance of their DenseNet models.*

> *(2) The inappropriate calculation of the brain cognition gap due to the lack of control for regression-toward-the-mean and the influence of the working memory itself (a common practice in brain age studies).*
>
> *(3) The lack of benchmarking of the brain cognition gap against the 'corrected' brain age gap and the direct prediction of physical activity, education, and cardiovascular risk.*
>
> *(4) Minimal justification for their PET mediation analysis.*

We appreciate the reviewer's constructive comments on the strengths and weaknesses of our study. In this revised version, we've addressed the concerns regarding the calculation of the brain-cognitive gap, clarified the unique variance that the brain-cognitive gap contributes beyond cognition itself, and provided additional justification for the PET mediation analysis. For the lack of a baseline model, it is important to highlight that our aim has never been to compare the predictive power of different deep learning or machine learning approaches. Therefore, the text in the introduction and discussion has been amended to avoid miscommunication on this topic.

> *Regarding the impact of the work on the field and the utility of the methods and data to the community, I see its potential. However, addressing all the weaknesses listed above is crucial and likely to change the conclusions of the results.*
>
> *It is important to note that many statements in the manuscript are overstated, making the contribution of the manuscript seem exaggerated.*

We have run additional analysis based on the reviewer's suggestions. The effect sizes and statistical values were adjusted due to the corrections; the overall conclusions remain largely consistent. The relationships between the brain-cognition gap and key factors such as physical activity, and cardiovascular risk persisted. We have updated the manuscript accordingly and revised the relevant sections to reflect these refinements and the resulting interpretations.

> *For instance, the abstract claims "there is a lack of objective biomarkers to accurately predict cognitive function," and the discussion states, "across various studies, the correlation between predicted and actual fluid intelligence typically hovers around 0.25 (98-100)." However, a meta-analysis by Vieira and colleagues (2022 https://doi.org/10.1016/j.intell.2022.101654) found over 37 studies up to 2020 predicting cognitive abilities from fMRI with machine learning, with 24 studies published in 2019-20 alone. Since 2020, with the rise of machine learning and AI, even more studies have likely been published on this topic, all claiming to show objective biomarkers to accurately predict cognitive function. Vieira and colleagues also found an average performance of these objective biomarkers in predicting general cognition at r = .42, similar to what was found in this manuscript. Based on this alone, it is unclear how novel or superior their method is without a proper systematic benchmark.*

We appreciate the opportunity to clarify our study's contribution relative to prior work. We have revised the introduction and discussion to highlight the contribution of other methods when it comes to biomarkers. As for the comment related to the work by Vieira and colleagues, Vieira et al. (2022) indeed present a comprehensive meta-analysis of studies predicting general and fluid intelligence using neuroimaging and machine learning. However, there are two critical differences between ours verus previous work:

Target Cognitive Domains:

Our study does not focus on general or fluid intelligence, but rather on comprehensive EM (3 tests) and WM (3 tests), two distinct cognitive domains that are critically important for aging

research. These distinct abilities, in this context (measured by three independent tests to boost the reliability) are less frequently studied as predictive targets in the existing fMRI-ML literature, particularly using deep learning methods.

Critically, our study explicitly compares predictive power across different cognitive states (rest, movie watching, n-back), with the aim of identifying the states that best capture individual differences across domains. Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach.

Our primary objective is to test how brain state influences the ability of functional connectivity to predict domain-specific cognitive performance, using a deep learning framework. As now stated explicitly in the revised manuscript, this objective is operationalized through three clearly defined aims:

(1) To compare the predictive utility of functional connectomes derived from different brain states (resting state, movie watching, and n-back task) for EM and WM;

(2) To introduce and evaluate a brain-cognition gap as a marker of individual differences beyond chronological age; and

(3) To examine the contribution of dopaminergic integrity to variability in connectome uniqueness and brain-cognition gaps.

We have revised the manuscript text to make this focus clearer and to avoid any misinterpretation of our aims. Specifically, we removed statements in the Discussion that could be read as suggesting that our deep learning approach outperforms prior machine learning methods. While we compared our model with the connectome predictive modeling (CPM) approach and observed better performance with our deep learning framework for some of the prediction models, we did not conduct a comprehensive benchmark across all available machine learning methods nor was this the aim of the present study. Accordingly, we have adjusted the text to avoid implying methodological/biomarker superiority beyond the scope of our analyses.

Modeling Approach:

While Vieira et al. show that the majority (76%) of prior studies used linear modeling approaches, including CPM and penalized regressions, these models are often vulnerable to overfitting, especially when applied to high-dimensional fMRI data. Our use of a DenseNet-based CNN architecture is motivated by the need to leverage inductive biases suited to functional connectivity data, and we evaluate this approach across multiple cognitive tasks and independent datasets.

Vieira and colleagues report that studies predicting general intelligence from fMRI (particularly from the HCP dataset) average around r =0.42, while those predicting fluid intelligence average around r =0.15. Our original claim about the correlation hovering around 0.25 is therefore not incorrect – and aligns with the Vieira meta-analysis. We have, however, nuanced this statement in the manuscript, now stating that correlations are higher for general intelligence than fluid intelligence.

Altogether, we considered the reviewer's comments and therefore conducted a careful revision of the manuscript text to moderate and clarify statements that may have come across as overstated. We have refined the language throughout the Introduction and Discussion sections to better align with the strength of the evidence and the scope of our contributions. A few examples are:

“Our study explicitly compares predictive power across different cognitive states (rest, movie watching, n-back), with the aim of identifying the states that best capture individual differences across domains. The relative performance of deep learning and other non-linear approaches depends on multiple factors, including sample size, model architecture, feature representation, and domain-specific characteristics of the prediction target. In this context, deep learning was employed as a flexible framework capable of modeling high-dimensional functional connectivity patterns across cognitive states, rather than as a claim of inherent methodological superiority. Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach.”

Also in page 14.

“Our study introduces a deep neural network architecture that features dense connections and incorporates an attentional mechanism. While our findings demonstrate that a deep learning framework can provide reasonable predictive accuracy, it is important to note that other machine learning approaches (e.g., tree-based models) may offer comparable predictive power, as suggested by prior benchmarking work (29, 30).”

> *Similarly, the authors claim superior performance of deep learning and mischaracterize machine learning algorithms: "In particular, deep neural networks (DNN) methods have been successfully applied to behavioral and disease prediction (24-26), and have been found to outperform other machine learning approaches (27-29)," and "Deep learning approaches overcome the limitation of predictive techniques that solely rely on linear associations between connectivity and behavioral phenotypes (17)." However, the superiority of deep learning is debatable. Studies show comparable performance between machine learning (such as kernel regression) and deep learning (such as fully-connected neural networks, BrainNetCNN, Graph CNN (GCNN), and temporal CNN), e.g., He and colleagues (2019) and Vieira and colleagues (2024) https://doi.org/10.1016/j.neuroimage.2019.116276 and Vieira and colleagues' https://doi.org/10.1101/2024.03.07.583858.*

We agree that the performance gap between traditional machine learning models and deep learning (which is a subcategory of machine learning) in neuroimaging is debatable and task-dependent. Indeed, both He et al. (2019) and Vieira et al. (2024) offer evidence that kernel regression can achieve performance on par with deep learning models, applied to appropriate datasets.

We have therefore nuanced the statements in the revised version of the manuscript as follows:

Introduction:

“In particular, deep neural networks (DNN) methods have been successfully applied to behavioral and disease prediction (24-26), and were initially expected to outperform other machine learning approaches (27-29). However, this superiority remains debatable, as recent studies have reported comparable performance between DNNs and traditional methods (He et al.,2019; Vieira et al.,2024). Accordingly, the present study does not aim to benchmark deep learning against traditional machine learning approaches, but instead uses a consistent predictive framework to examine how brain state influences the utility of FC for cognitive prediction.”

“Deep learning approaches offer a flexible modeling framework capable of capturing complex non-linear associations in high-dimensional data with potentially less sensitivity to training on a smaller subsample (Vieira et al., 2024)”.

Discussion:

We agree that traditional methods, such as kernel-based models, tree ensembles, and non-linear SVRs, can also effectively capture such relationships. The relative performance of our model and other non-linear approaches depends on several factors, including data size, model architecture, and domain-specific considerations. We have included additional explanations in the discussion to address this.

> *Moreover, many non-deep learning predictive techniques are non-linear, e.g., XGBoost, CatBoost, random forest, kernel ridge, and support vector regression with non-linear kernels (such as RBF and polynomial). Thus, stating that machine learning can only model linear relationships is incorrect. Moreover, for the small amount of data the authors had, some might argue that a linear algorithm might be more appropriate to balance the bias-variance trade-off in prediction. Again, without a proper systematic benchmark, it is unclear how well their DenseNet algorithm performs compared to other algorithms.*

Thank you for bring this up. We have now removed statements implying that machine learning can only model linear relationship.

> *Regarding the Brain Age literature, the authors also misinterpreted recent findings: "However, a recent study suggests that brain age predictions contribute minimally compared to chronological age for explaining cognitive decline (65), implying that cognitive predictions are more reliable." In this study, Tetereva and colleagues (2024) (https://doi.org/10.7554/eLife.87297.4) showed that non-deep-learning machine learning can make good predictions from MRI on both chronological age (with r up to .88) and fluid cognition (with r up to .627). Using the combination of functional connectivity matrices across rest and tasks to predict fluid cognition, they found performance at r = .565, comparable to what was found in the current manuscript with deep learning. Nonetheless, while brain age predicted chronological age well (and brain cognition predicted fluid cognition well), it was problematic to predict fluid cognition from brain age. They showed that, because brain age, by design, shared so much common variance with chronological age, brain age and chronological age captured the same variance of fluid cognition. When chronological age was controlled for in the prediction of fluid cognition, brain age no longer had high predictive ability. In the case of the current manuscript, the brain cognition gap is not appropriately controlled for cognition (to be more precise, a working memory score). I expect the performance in predicting physical activity, education, and cardiovascular risk will drop dramatically once cognition is controlled for. There are at least two ways to control cognition according to Tetereva and colleagues' study (see more in the recommendations).*

We thank the reviewer for breaking down the findings in the study by Tetereva and colleagues (2024). It was not our intention to suggest that Tetereva et al. showed brain age has little predictive value in general. Our understanding of the findings reported in that study is on par with the reviewers' clarifications. We have now revised the introductions to avoid any misunderstanding:

"A recent study demonstrated that while brain age can predict chronological age with high accuracy from MRI, its utility for predicting cognition is limited. Specifically, Tetereva and colleagues (2024) showed that brain age strongly tracks chronological age and that brain cognition (using functional connectivity) can predict fluid cognition. Yet, when used to predict cognition, brain age largely overlapped with chronological age, such that controlling for chronological age eliminated the predictive contribution of brain age. This finding suggests that brain-age models may provide little unique explanatory power for cognitive decline beyond what is already captured by chronological age. Building on this observation and

extending the concept of a brain-age gap to a brain-cognition gap (BCG, defined as the discrepancy between predicted and observed cognitive performance), we propose that a BCG may serve as an informative marker of individual differences."

In addition, in response to the first comment from Reviewer 1, we have extended our results in the manuscript. We first showed that BCG is not significantly associated with cognition itself (see Author response image 1). Moreover, we conducted additional analyses, splitting the sample into high and low EM performers, and compared their levels of physical activity and Framingham cardiovascular risk scores. We found that no significant difference in physical activity (DyNAMiC: p =0.56, 95% CI: -14.99 – 8.13; COBRA: p =0.29, 95% CI: -3.54 – 1.05) or Framingham CVD risk score (DyNAMiC: p =0.11, 95% CI: -1.08 – 10.72; COBRA: p =0.41, 95% CI: -1.86 – 4.58) between high and low EM performers. Given the significant difference in physical activity and Framingham CVD risk score between positive and negative BCG groups, our results support that BCP provides unique information, beyond cognitive measures, regarding factors that contribute to cognitive resilience. This text has been added into the result section, and Figure 3 has been updated in the manuscript.

> *The authors mentioned, "The third aim of the current study is to uncover the contribution of dopamine (DA) integrity to brain-cognition gaps." However, I fail to see how mediation analysis would test this. The authors also mentioned, "Insufficient DA modulation can affect neurocognitive functions detrimentally (69, 74, 76-78)." They should test if DA levels are related to working memory scores in their study, and if so, whether the relationship is mediated by the "corrected" brain-cognition gaps. Note see more on the recommendation for the calculation of the "corrected" brain-cognition gaps.*

Our mediation was not designed to test whether DA predicts episodic memory performance directly, nor whether BCG mediates such a relationship. Instead, we specifically investigated whether the effect of DA on BCG operates through functional variability, the theoretical framework emphasizing the role of DA on neuronal grain and signal-to-noise ratio (see our recent work in Korkki et al., 2025). We agree that future work could extend our approach by directly examining whether BCG mediates the link between DA and cognitive outcomes. However, in the present study, our primary focus was on testing the mechanistic pathway of DA → entropy → BCG.

In line with this aim, we found that lower DA receptor availability was associated with larger BCGs (Figure 4). We then asked whether this relationship is mediated by functional signal variability, such that lower DA is linked to reduced signal-to-noise ratio (i.e., greater entropy), which in turn contributes to less reliable prediction of cognition and, consequently, larger BCGs. Our mediation analysis supports this pathway (please see also our reply to Reviewer 1, Comment 6).

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *This paper by Esmaeili and co-authors presents a connectome prediction study to predict episodic memory and relate prediction errors to other phonotypic variables.*
>
> *Strengths:*
>
> *(1) A primary and external validation dataset.*
>
> *(2) Novel use of prediction errors (i.e., brain-cognitive gap).*
>
> *(3) A wide range of data was investigated.*
>
> *Weaknesses:*

> *(1) Lack of comparisons to other methods for prediction.*
>
> *(2) Several different points are being investigated that don't allow any particular one to shine through.*
>
> *(3) Some choices of analysis are not well-motivated.*
>
> *(4) How do the n-back connectomes perform for prediction if the authors do not regress task activations from the n-back task?*

We thank the reviewer for raising these important points. For the lack of comparisons to other methods, it is important to highlight that our aim has never been to compare the predictive power of different deep learning or machine learning approaches. Rather, our primary objective was to test how brain state influences the ability of functional connectivity to predict domain-specific cognitive performance, using a deep learning framework.Therefore, the text in the introduction and discussion has been amended to avoid miscommunication on this topic.

We chose to regress out task-evoked activations based on prior work demonstrating that failing to do so can produce spurious but systematic inflation of task functional connectivity estimates (Cole et al., 2019). In that study, as well as subsequent reports (e.g., Gao et al., 2020; Gonzalez-Castillo & Bandettini, 2018), connectomes derived without activation regression tended to capture task-evoked coactivations rather than background task functional interactions, which can artificially boost predictive performance but limit interpretability (whether it is co-activation or intrinsic connectivity during an entire goal-oriented task) and generalizability. For this reason, our analyses focused on the more conservative approach of regressing out task activations. Accordingly, we compared predictive performance only under this preprocessing strategy.

We have added the following sentence to clarify this in the method: “To avoid spurious inflation of task functional connectivity by task-evoked activations, we regressed out task activation patterns from the n-back data prior to estimating functional connectivity, following recommendations by Cole et al. (2019) and related work.”

> *(5) I am a little concerned about overfitting with the convolutional neural net. For example, the drop-off in prediction performance in the external sample is stark. How does the deep learning approach used here compare to something simpler, like a connectome-based predictive model or ridge regression?*
>
> *(6) It may be nice to try the other models in the validation dataset. This would also provide a sense of the overfitting that may be going on with overfitting.*

We thank the reviewer for raising this point. The prediction performance indeed dropped for episodic memory when models trained on the DyNAMiC sample were applied to the COBRA sample, whereas performance for working memory remained nearly identical across datasets. Moreover, our prediction power is on par with previous studies reporting reliable prediction of intelligence using deep learning approach (Vieira et al., 2021; Fan et al.,2020). While we compared our model with the connectome predictive modeling (CPM) approach and observed better performance with our deep learning framework, we did not conduct a comprehensive benchmark across all available machine learning methods nor was this the aim of the present study.

We have revised the manuscript text to make this focus clearer and to avoid any misinterpretation of our aims. Specifically, we removed statements in the Discussion that could be read as suggesting that our deep learning approach outperforms prior machine learning methods. Finally, We have added the following paragraph to the discussion:

"Our study used a deep neural network architecture that features dense connections and incorporates an attentional mechanism. While our findings demonstrate that a deep learning framework can provide reasonable predictive accuracy, it is important to note that other machine learning approaches (e.g., tree-based models) may offer comparable predictive power, as suggested by prior benchmarking work (29, 30). Our study explicitly compares predictive power across different cognitive states (rest, movie watching, n-back) to identify the states that best capture individual differences across domains. The relative performance of deep learning and other non-linear approaches depends on multiple factors, including sample size, model architecture, feature representation, and domain-specific characteristics of the prediction target. In this context, deep learning was employed as a flexible framework capable of modeling high-dimensional functional connectivity patterns across cognitive states, rather than as a claim of inherent methodological superiority. Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach."

> *(7) While predictive models increase the power over association studies, they still require large samples to prevent overfitting. Do the authors have a sense of the power their main and external validation sample sizes provide?*

We thank the reviewer for this important point. Our main sample size, together with the external validation in COBRA, is moderate for deep learning applications. To reduce the risk of overfitting, we employed several strategies, including external validation, early stopping, dropout, and regularization. As noted, performance for episodic memory decreased in the external sample, which we acknowledge, but key associations such as the link between BCG and resilient factors remained significant. Importantly, prediction of working memory was maintained across datasets, reducing the likelihood that the observed findings are driven by overfitting. We have added a statement in the Discussion to reflect on the limitations of sample size and the implications for generalizability.

We added the following sentence to the discussion:

"We acknowledge that our main and validation samples are moderate in size for deep learning, which constrains statistical power and generalizability. Although external validation, early stopping, dropout, and regularization help mitigate overfitting, larger samples will be needed in future work to fully establish the robustness of these predictive models."

> *(8) I am not sure that the Mann-Whitney is the correct test for comparing the distributions of prediction performances. The distributions are dependent on each other as they are each predicting the same outcomes. Using the typical degrees of freedom formula would overestimate the degrees of freedom.*

We appreciate the reviewer's comment and agree that applying statistical tests directly to bootstrapped samples can lead to inflated or misleading *p*-values, as the degrees of freedom are determined by the number of bootstrap iterations rather than the actual number of independent observations.

In our analysis, the Mann-Whitney *U* test was applied to 1000 bootstrapped correlation coefficients (r) for each model. While this number is relatively low and was chosen to limit overestimation of significance, we recognize that these bootstrapped samples are not independent, and thus the use of a Mann-Whitney *U* test can still be problematic. To address this concern, we have revised our statistical analysis. Rather than applying the Mann-Whitney *U* test to the bootstrapped r distributions, we now compute the difference in correlation coefficients ($\Delta r = r_{actual} - r_{rest}$) for each bootstrap iteration. We then calculate a 95% confidence interval for Δr. If this interval does not include zero, we consider the

difference statistically significant. This approach avoids artificially inflating the sample size and adheres more closely to proper statistical inference.

We have updated the Methods (the following text) and Results sections accordingly and clearly stated the limitations regarding the degrees of freedom for all tests.

"For the bootstrap-based comparison of model performance (bootstrap resampling with 1000 iterations), no test statistic with an associated degree of freedom is reported. Instead, statistical inference is based on the bootstrap distribution of the difference in correlation coefficients (Δr) and its 95% confidence interval. As bootstrap confidence-interval–based inference does not rely on an analytic sampling distribution, degrees of freedom are not defined for this procedure." This has now been explicitly stated in the Methods section to avoid ambiguity.

In the result section, we have reported with corresponding CI.

> *(9) The brain cognition gap is interesting. It is very similar conceptually to the brain age gap. When associating the brain age gap with other phenotypes, typically age is regressed from the brain age gap and the other phenotype. In other words, age is typically associated with a brain age gap as individuals at the tail ages often show the largest gaps. Is the brain cognition gap correlated with episodic memory and do the group differences hold if episodic memory is controlled for?*

We thank the reviewer's comment regarding the relationship between the brain cognition gap and episodic memory.

Since this question was raised by all reviewers, we have conducted additional analyses. We did find that BCG is independent from the cognitive measure and provided additional information, beyond cognition alone, about factors contributing to resilience. Please visit our response to the first comment of Reviewer 1.

> *(10) I have the same question for the dopamine results. Particularly, in the correlations that are divided by brain cognition gap sign. I could see these types of patterns arise due to a correlation with a third variable.*

For dopamine results, we explored whether age or cognition alone might confound the dopamine–brain cognition gap relationships. However, neither was significantly correlated with the brain cognition gap groups. The associations remained significant after controlling for age, suggesting that the observed patterns are not likely due to these potential third-variable confounder. This is also inline with our observation of significant associations between DA and GAP in an age-homogeneous COBRA sample. That said, we found that entropy, indeed, mediates the direct link between DA and BAG, suggesting that individuals with lower DA exhibit greater regional variability, and in turn larger BCG.

These results have now been embedded into the manuscript. We also highlighted that age has been controlled for in reported correlation and mediation analyses.

> ***Recommendations for the authors:***
>
> ***Reviewing Editor Comment:***
>
> *We particularly recommend that the authors: (a) compare the performance of their deep learning model with other baseline models, and (b) adjust for cognitive performance within the brain-cognition gap. These steps would strengthen the evidence base.*

We thank the editor for their comments. As for the first comments, our study explicitly compares predictive power across different cognitive states (rest, movie watching, n-back), with the aim of identifying the states that best capture individual differences across domains.

Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach. We have revised the manuscript text to make this focus clearer and to avoid any misinterpretation of our aims. Specifically, we removed statements in the Discussion that could be read as suggesting that our deep learning approach outperforms prior machine learning methods. While we compared our model with the connectome predictive modeling (CPM) approach and observed better performance with our deep learning framework, we did not conduct a comprehensive benchmark across all available machine learning methods, nor was this the aim of the present study. Accordingly, we have adjusted the text to avoid implying methodological superiority beyond the scope of our analyses. Finally, we have added the following paragraph to the discussion:

"Our study used a deep neural network architecture that features dense connections and incorporates an attentional mechanism. While our findings demonstrate that a deep learning framework can provide reasonable predictive accuracy, it is important to note that other machine learning approaches (e.g., tree-based models) may offer comparable predictive power, as suggested by prior benchmarking work (29, 30).

Our study explicitly compares predictive power across different cognitive states (rest, movie watching, n-back) to identify the states that best capture individual differences across domains. The relative performance of deep learning and other non-linear approaches depends on multiple factors, including sample size, model architecture, feature representation, and domain-specific characteristics of the prediction target. In this context, deep learning was employed as a flexible framework capable of modeling high-dimensional functional connectivity patterns across cognitive states, rather than as a claim of inherent methodological superiority. Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach."

As for the second comment, we followed the instructions by Reviewer 1. In response to their request, we first examined the relationship between the Brain-Cognitive Gap (BCG) and the cognitive measure itself. Surprisingly, we did not find any significant relationship in either the DyNAMiC sample (r =0.01, p =0.939) or the COBRA sample (r =0.01, p =0.89) (see Author response image 1).

We then conducted additional analyses, splitting the sample into high and low EM performers, and compared their levels of physical activity and Framingham cardiovascular disease (CVD) risk scores. We found no significant difference in physical activity (DyNAMiC: p =0.56, 95% CI: –14.99 - 8.13; COBRA: p =0.29, 95% CI: –3.54 - 1.05) or Framingham CVD risk score (DyNAMiC: p =0.11, 95% CI: –1.08 - 10.72; COBRA: p =0.41, 95% CI: –1.86 - 4.58) between high and low EM perfprmers. Given the significant difference in physical activity and Framingham CVD risk score between positive and negative BCG groups, our results support that BCG provides unique information, beyond the observed cognitive measure (episodic memory score), regarding factors that contribute to cognitive resilience. These results have been added to Section 2.4, and Figure 3 has been updated.

> ***Reviewer #1 (Recommendations for the authors):***
>
> *(1) The top and bottom triangles of the saliency maps, particularly in Figure 2, do not look symmetrical (this is most notable in the hotspot representing the between-network correlation of DMN and FPN). What is going on here? Was the image compressed or altered in some way, or is this a visual artifact of the interpolation method?*

We appreciate the reviewer's insightful comment. Minor differences in the saliency maps between the upper and lower triangles of the FC matrix can arise due to several factors. For instance, Grad-CAM generates saliency maps at the resolution of the convolutional feature

maps, which are then upsampled to match the input matrix dimensions. We initially used the default bilinear interpolation, which may have introduced slight asymmetries or blurring, resulting in interpolation artifacts. In response, we have reprocessed the saliency maps using spline interpolation in MATLAB. The updated saliency figures have been included in the revised version of the manuscript.

> *(2) Pages 11-12. Please make it explicit in the text that the brain gap-education association was not significant in the COBRA dataset.*

Thanks for pointing this out. We added the following sentence to the discussion.

"Note that the association with education was significant only in the DyNAMiC sample and did not reach significance in the COBRA dataset."

> *(3) Please overlay individual data points onto the boxplots in Figure 3 so that we can appropriately evaluate the data distributions.*

Figure 3 has now been updated.

> *(4) Section 2.6: Was entropy calculated on movie-watching data, resting data, or all fMRI data? Please specify.*

We thank the reviewer for pointing this out. We have updated the text (Section 2.6) to clarify that entropy was calculated from the resting-state data. We intended to examine the mediating role of regional variability in the relationship between dopamine and the BCG of the winning model for episodic memory. Because resting state and movie-watching were the winning conditions for EM prediction, but movie-watching was not available in COBRA, we focused on entropy during rest, which exists in both datasets.

> *(5) Was entropy during the resting state correlated with entropy during the task state, across individuals?*

We agree this is an interesting question. However, investigating the correlation of entropy between rest and task states goes beyond the scope of the present study. Our aim here was to test whether regional variability mediates the effect of dopamine on the BCG. Specifically, we examined whether individuals with lower striatal D1DR show higher local variability, which in turn relates to less accurate prediction and a larger gap. We assessed both the relationship between D1DR and entropy and the association between entropy and the gap, and these results have now been added to the manuscript (see also our response to Reviewer 1's public comment).

> ***Reviewer #2 (Recommendation for authors):***
>
> *(1) The lack of baseline models to benchmark the predictive performance of their DenseNet models makes their results hard to interpret. This problem is quite common across ML literature. For instance, many DL-based algorithms were developed for tabular data without proper benchmarking against other ML algorithms. When they were properly tested, most weren't better than many tree-based ML algorithms (e.g., https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf). I can see that a similar problem might happen here.*
>
> *For this particular manuscript, the authors made strong statements without doing a proper benchmark, e.g., from the discussion, "Indeed, the predictive power in the current study is stronger than for CPM-based predictions reported before." And "Unlike the BrainNet convolutional neural network, which focuses on staged transformations, our densely connected model promotes extensive feature reuse, possibly leading to more robust feature extraction." I hope to see the performance of the proposed algorithm*

> *against 1) other DL algorithms (e.g., fully-connected neural networks, BrainNetCNN, Graph CNN (GCNN), temporal CNN, GRU, and LSTM, see https://doi.org/10.1016/j.neuroimage.2019.116276 and https://doi.org/10.1002/hbm.26415), 2) ML algorithms (e.g., SVR with linear, RBF and polynomial kernels, Elastic Net, XGBoost, random forest, CPM), 3) data reduction algorithms (e.g., PCA regression, Partial Least Square). The results of this benchmark will substantiate the claims made by the authors.*

Our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach. We have revised the manuscript text to make this focus clearer and to avoid any misinterpretation of our aims. Specifically, we removed statements in the Discussion that could be read as suggesting that our deep learning approach outperforms prior machine learning methods. While we compared our model with the connectome predictive modeling (CPM) approach and observed better performance with our deep learning framework, we did not conduct a comprehensive benchmark across all available machine learning methods, nor was this the aim of the present study. Accordingly, we have adjusted the text to avoid implying methodological superiority beyond the scope of our analyses. Finally, we have added the following paragraph to the discussion:

"Our study used a deep neural network architecture that features dense connections and incorporates an attentional mechanism. While our findings demonstrate that a deep learning framework can provide reasonable predictive accuracy, it is important to note that other machine learning approaches (e.g., tree-based models) may offer comparable predictive power, as suggested by prior benchmarking work (29, 30). Our study explicitly compares predictive power across different cognitive states (rest, movie watching, n-back) to identify the states that best capture individual differences across domains. The relative performance of deep learning and other non-linear approaches depends on multiple factors, including sample size, model architecture, feature representation, and domain-specific characteristics of the prediction target. In this context, deep learning was employed as a flexible framework capable of modeling high-dimensional functional connectivity patterns across cognitive states, rather than as a claim of inherent methodological superiority. Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach."

> *(2) From Figure 6b, it looks like the functional connectivity matrices were converted to different images, and each of the four images (in grey, blue, yellow, and red) was treated as a separate channel. What are these grey, blue, yellow, and red images?*

In our study, the inputs to the deep learning models were subject-specific FC matrices of size 273×273. To augment the data, we created different versions of each FC matrix by reordering specific brain networks within the matrix. To visualize that the inputs were augmented, we used different color codings (grey, blue, yellow, and red) in Figure 6b. These colors were intended solely to represent different augmented versions of the same subject's FC matrix. They were not treated as separate channels in the model. To avoid any confusion or misinterpretation, we have revised this part of the figure and now use only grey coloring to represent the augmented FC matrices.

> *(3) The differences in performance between within vs. outside studies might simply be due to the fact that the models trained from DyNAMiC captured the brain variation due to age, which is also related to cognitive abilities. I was wondering if age is controlled for, would performance be more similar across the studies? The authors should provide the performance of models that are controlled for age.*

We initially conducted partial correlation between FC features and cognitive measures while controlling for age. This is further supported by the fact that the model trained on the age-

heterogeneous DyNAMiC sample provided a fairly reasonable prediction in the age-homogeneous COBRA dataset, particularly for working memory (see figure 2d). Moreover, in our post hoc analyses, we additionally controlled for age when examining associations, for example, between GAP and dopamine measures.

> *(4) Related to point (3), from the discussion, "Validation outcomes thus affirm that the models, particularly those constructed from rest data, are robust to the particulars of the dataset." The performance dropped around half, so I am not sure if this conclusion is warranted.*

We thank the reviewer for raising this point. The prediction performance indeed dropped for episodic memory when models trained on the DyNAMiC sample were applied to the COBRA sample, whereas performance for working memory remained nearly identical across datasets. Although both EM and WM are sensitive to age, the divergence in cross-dataset performance suggests that factors beyond age alone may contribute to these differences. To address this, we have revised the discussion as follows:

“Differences between the DyNAMiC and COBRA datasets make cross-dataset prediction a harder problem, as the age ranges of samples significantly vary, and prior studies highlight the importance of individual characteristics like age in predicting behavior from FC (33). In line with this, model performance decreased when predicting EM in the COBRA sample whereas prediction of WM remained largely unchanged. Thus, validation outcomes suggest that the models, particularly those predicting WM, show robustness across datasets, whereas the reduced EM performance highlights potential data-specific influences that limit generalizability.”

> *(5) Please report the degree of freedom in all of the statistical analyses. Was the Mann-Whitney U test done on the bootstrapped r? If so, the degree of freedom was arbitrarily set by the number of bootstrapping, and hence the p-value can be higher or lower depending on the number of bootstrapping. This could lead to misleading conclusions.*

We appreciate the reviewer’s comment and agree that applying statistical tests directly to bootstrapped samples can lead to inflated or misleading p-values, as the degrees of freedom are determined by the number of bootstrap iterations rather than the actual number of independent observations.

In our analysis, the Mann-Whitney *U* test was applied to 1000 bootstrapped correlation coefficients (r) for each model. While this number is relatively low and was chosen to limit overestimation of significance, we recognize that these bootstrapped samples are not independent, and thus the use of a Mann-Whitney *U* test can still be problematic. To address this concern, we have revised our statistical analysis. Rather than applying the Mann-Whitney *U* test to the bootstrapped r distributions, we now compute the difference in correlation coefficients ($\Delta r = r_{actual} - r_{rest}$) for each bootstrap iteration. We then calculate a 95% confidence interval for Δr. If this interval does not include zero, we consider the difference statistically significant. This approach avoids artificially inflating the sample size and adheres more closely to proper statistical inference.

We have updated the Methods (the following text) and Results sections accordingly and clearly stated the limitations regarding the degrees of freedom for all tests.

"For the bootstrap-based comparison of model performance (bootstrap resampling with 1000 iterations), no test statistic with an associated degree of freedom is reported. Instead, statistical inference is based on the bootstrap distribution of the difference in correlation coefficients (Δr) and its 95% confidence interval. As bootstrap confidence-interval–based inference does not rely on an analytic sampling distribution, degrees of freedom are not defined for this procedure." This has now been explicitly stated in the Methods section to avoid ambiguity.

In the result section, we have reported with corresponding CI.

> *(6) For predictive performance, the correlation was reported in the table, while R^2 is reported in the text. This is confusing. Also, could you clarify if the R^2 is calculated using the sum square definition, not Pearson r squared? If Pearson r squared was used, then R^2 of a negative Pearson r would be positive, which is misleading (see 10.1001/jamapsychiatry.2019.3671). Also, other performance indices apart from Pearson r and R^2 should be reported (e.g., MSE and MAE, again see 10.1001/jamapsychiatry.2019.3671). This will allow a better understanding of the models' performance.*

We thank the reviewer for this helpful comment. We acknowledge the inconsistency in reporting predictive performance metrics and have revised the manuscript for clarity. In the text, we have reported the *r* value, whereas in the table, we have reported r^2 using the sum-of-squared definition. Specifically, we now consistently report Pearson correlation (*r*), mean squared error (MSE), and mean absolute error (MAE) across both the text and Tables 1 and 2.

Regarding r^2, we confirm that it was calculated using the sum-of-squares definition (i.e.,

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

rather than as the square of the Pearson correlation coefficient. This ensures that negative correlations do not result in misleading positive R^2 values, as pointed out by the reviewer and discussed in Poldrack et al. (2020). All performance metrics (*r*, r^2, MSE, and MAE) are now reported in Tables 1 and 2 to allow a more comprehensive and interpretable comparison of model performance.

We have included a description of the method under section 4.9. Statistical significance analysis.

> *(7) Could you clarify how data are standardized across training, validation, and tests (including Z-standardization for the cognitive tests)? This is to prevent data leakage.*

Thanks for the comments. We did standardization the cognitive test from both training and test, separately.

We have added the following paragraph to the method section:

"A composite score of performances across the three tests was calculated and used as the measure of the cognitive domain in question (i.e., episodic memory, working memory). For each of the three tests, scores were summarized across the total number of trials. The three resulting sum scores were z-standardized and averaged to form one composite score for each domain. The standardization has been carried out independently for the training (DyNAMiC) and test (COBRA) samples."

> *(8) There is really no ground truth to confirm that Grad-CAM provides actual feature importance used by the models. Perhaps the authors should compare that with Haufe transformation, which is commonly used in the predictive model for cognition*

> *(e.g., https://doi.org/10.1016/j.neuroimage.2021.118648 and https://doi.org/10.1016/j.neuroimage.2023.120115).*

We appreciate the reviewer's comment and the suggested references. The Haufe transformation is primarily applied in traditional machine learning models, particularly in cognitive neuroscience, to interpret linear predictive models by mapping classifier weights back to the input space. However, its direct applicability to deep learning models, especially convolutional neural networks, remains an open research area with no widely established methodologies. Furthermore, the Haufe transformation does not provide feature importance in the same manner as Grad-CAM. Grad-CAM highlights spatial regions within an image that contribute to a model's decision, making it particularly useful for interpreting convolutional networks in vision tasks. In contrast, the Haufe method offers a weight transformation that is more suited for understanding linear models and may not be as intuitive for feature attribution in complex hierarchical representations such as those learned by deep neural networks.

While we acknowledge that Grad-CAM, like other interpretability methods, does not provide absolute ground truth validation for feature importance, it remains one of the most widely used and validated techniques for deep learning interpretability, particularly in medical imaging applications. Given its integration with frameworks such as Keras and TensorFlow and its ability to provide spatial attributions aligned with domain knowledge, we believe it is a suitable choice for our study. Future work may explore additional interpretability techniques, including adaptations of the Haufe transformation if applicable to deep learning architectures.

We have added more details on Grad-CAM implementations in the Method.

> *(9) Related to Grad-CAM, "These edges, indicated by a salience intensity of {greater than or equal to}.5, exert a significant influence on the model (Figure 1f)." What does 'significant' in this context mean? And how did the authors come up with the .5 threshold? Is it based on permutation or bootstrapping tests?*

We appreciate the reviewer's comment and the opportunity to clarify our approach. In this context, the term "significant" refers to the regions' relative contribution to the model's decision, as shown by the Grad-CAM saliency map. However, to avoid implying statistical testing, we will revise the term to "highly contributing."

Regarding the 0.5 threshold, this value was selected empirically based on the normalized Grad-CAM activation values, where saliency scores range between 0 and 1. A threshold of 0.5 was used as a heuristic to highlight regions with relatively strong activation. However, this was not determined through statistical methods such as permutation or bootstrapping tests. We recognize the importance of rigorous threshold selection and will clarify this in the text. Future work could incorporate statistical methods to define thresholds more objectively.

We have included the following text in the Method section:

"Grad-CAM saliency maps were interpreted qualitatively, with a heuristic threshold (≥ 0.5) applied to highlight regions with relatively higher contribution to the model's predictions. These values do not reflect statistical significance and should therefore be interpreted descriptively."

> *(10) Still related to the saliency map, I believe the upper and lower triangles of the functional connectivity matrix are the same. If so, why are there some differences in saliency? While the difference is not prominent, this might affect the accuracy of Grad-CAM.*

Minor differences in the saliency maps between the upper and lower triangles of the FC matrix can arise due to several factors. For instance, Grad-CAM generates saliency maps at the resolution of the convolutional feature maps, which are then upsampled to match the input matrix dimensions. We initially used the default bilinear interpolation, which may have introduced slight asymmetries or blurring, resulting in interpolation artifacts. In response, we have reprocessed the saliency maps using spline interpolation in MATLAB. The updated saliency figures have been included in the revised version of the manuscript.

> *(11) Why did the authors only report the cross-study for EM on rest, and for WM on n-back? This is a bit unexpected since COBRA has both rest and n-back. If there is no good justification, please report both.*

We focused on reporting cross-study results for EM using rest because rest was the winning condition for predicting EM in the DyNAMiC sample. Importantly, n-back did not significantly predict EM in DyNAMiC, and rest did not significantly predict WM. For this reason, we highlighted only the conditions that showed meaningful predictive power in the original analyses.

> *(12) Are codes, trained models, and data available? To ensure transparency and reproducibility, I hope to see the code from preprocessing to modeling and statistical analyses.*

The analysis code is openly available on our GitHub page *https://github.com/MorEsm/AI-based-Prediction-of-Cognitive-Function*. Due to ethical considerations and GDPR restrictions in the European Union, we are not permitted to publicly share the raw data. However, we can provide detailed information about preprocessing steps and analysis pipelines to facilitate reproducibility.

> *(13 &14) The authors did not appropriately control for regression-toward-the-mean and the influence of the working memory itself when calculating the brain cognition gap. This is commonly done to brain age (see https://doi.org/10.7554/eLife.87297.4, https://doi.org/10.1002/hbm.25533, https://doi.org/10.1016/j.nicl.2020.102229, https://doi.org/10.3389/fnagi.2018.00317). Otherwise, the brain cognition gap still depends on the cognition/working memory score itself. Based on Tetereva et al., "If, for instance, Brain Age was based on prediction models with poor performance and made a prediction that everyone was 50 years old, individual differences in Brain Age Gap would then depend solely on chronological age (i.e., 50 minus chronological age)." Because of this, Tetereva and colleagues found that the 'uncorrected' brain age gap that predicted chronological age the worst became the best index to predict fluid cognitive abilities. This shows the pitfall of the 'uncorrected' brain age gap. You can apply the same logic to the brain cognition gap.*
>
> *(14) Additionally, another way to show the unique contribution of brain cognition, over and above cognition per se, is to add both brain cognition and cognition together to predict physical activity, education, and cardiovascular risk.*

We thank the Reviewer for raising this important point. In response to their request and also the request from Rev. 1, we first examined the relationship between the Brain-Cognitive Gap (BCG) and the cognitive measure itself. Surprisingly, we did not find any significant relationship in either the DyNAMiC sample (r =0.01, p =0.939) or the COBRA sample (r =0.01, p =0.894) (see Author response image 1).

We then conducted additional analyses, splitting the sample into high and low EM performers, and compared their levels of physical activity and Framingham cardiovascular risk scores. We found that no significant difference in physical activity (DyNAMiC: p =0.56, CI: -14.99 – 8.13; COBRA: p =0.29, CI: -3.54 – 1.05) or Framingham CVD risk score (DyNAMiC: p

=0.11, CI: -1.08 – 10.72; COBRA: p =0.41, CI: -1.86 – 4.58) between high and low EM perfprmers. Given the significant difference in physical activity and Framingham CVD risk score between positive and negative BCG groups, our results support that BCP provides unique information, beyond cognitive measure, regarding factors that contribute to cognitive resilience. These results have been added to Section 2.4, and Figure 3 has been updated.

> *(15) Related to the brain age gap, the brain cognition gap is actually just another way to quantify how generalizable models are to another sample, similar to MAE or MSE. If the models built from DyNAMiC don't fit well with samples from COBRA, you will get a higher (i.e., wider) brain cognition gap, which means a poor fit. The authors should discuss this interpretation - should your biomarker's performance be due to a fit of the model?*

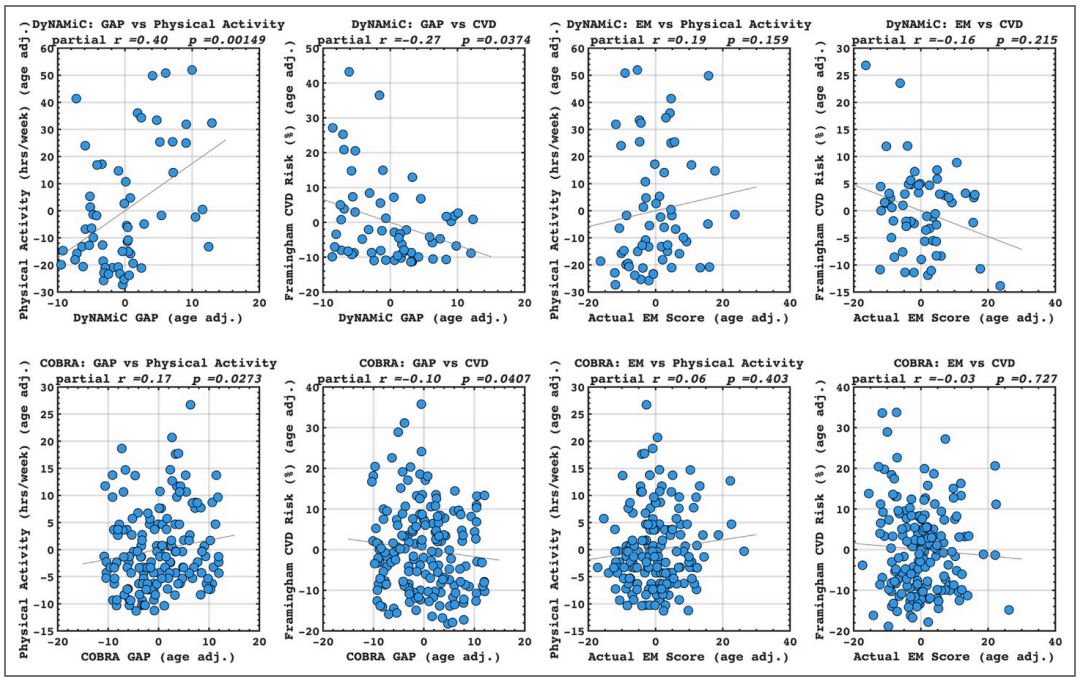
We appreciate this insightful comment. We agree that BCG can be interpreted not only as a marker of individual differences and resilience factors but also as a measure of model fit, analogous to error metrics, such as MAE or MSE. A higher gap may, in part, reflect poorer generalizability of models across samples. We have now revised the Discussion to explicitly acknowledge this alternative interpretation and to emphasize that BCG should be viewed both as a candidate biomarker and as a reflection of model performance.

We added the following paragraph in the discussion:

“An important caveat is that BCG can also be conceptualized as an error metric, similar to mean absolute error or mean square error, reflecting the extent to which models trained in one sample generalize to another. From this perspective, a larger gap may not only indicate individual differences related to resilience factors and dopaminergic function, but also reduced model fit or generalizability across datasets. Thus, BCG likely reflects a combination of meaningful biological variability and methodological variance.”

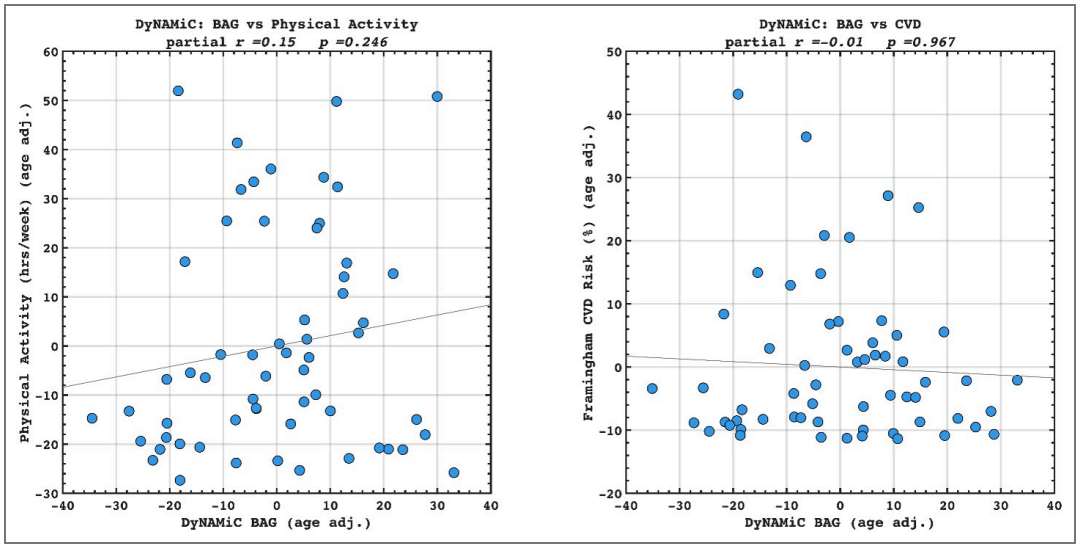
> *(16) It is unclear why the authors binarized the brain cognition gap when predicting physical activity, education, and cardiovascular risk, and not doing so with the striatal D1DR. It is rarely a good idea to binarize a continuous variable (see 10.1136/bmj.332.7549.1080). In this case, people who had a bigger negative brain cognition gap were treated equally to people who had a smaller negative brain cognition gap. I also do not think it is necessary to separately analyze positive and negative gaps. Perhaps the authors should correlate the corrected brain cognition gap with physical activity, education, and cardiovascular risk and provide scatter plots and effect sizes.*

Following the reveiwer suggestion, we directly correlated BCG with physical activity and cardiovascular risk. Our results confirmed our initial analysis that individuals with a negative gap exhibited lower physical activity and higher Framingham CVD risk across both COBRA and DyNAMiC datasets. We have reported these results on page 10.



Author response image 5.

> *(17) Given that the motivation is to move away from brain age, the authors should benchmark the corrected brain cognition gap against the corrected brain age gap, as well as against the performance when directly predicting physical activity, education, and cardiovascular risk from the functional connectivity metrics.*



Author response image 6.

We agree that benchmarking BCG against BAG in predicting lifestyle and vascular risk factors would be valuable. We have calculated adjusted BAG and related it to lifestyle and vascular risk factors. Interestingly, we did not find any significant association, suggesting that BCG might be more sensitive to cognitive resilience. However, this investigation was beyond the scope of the present study. Our aim was not to compare BCG with BAG, but rather to examine whether BCG provides information beyond cognition itself. We also note that introducing BAG would open a separate line of investigation, namely, which cognitive state (rest, movie-

watching, n-back) best estimates biological age. While this is an interesting question in its own right, addressing it here would considerably broaden the scope and complexity of an already dense manuscript. To prevent misunderstanding, we have clarified this point in the Discussion and added a caveat noting that future work should explicitly benchmark these approaches. That said, if the Reviewer and/or the Editor incline to add these additional findings into the manuscript, we are open to doing so in a revision.

We have added the following sentence to the Discussion.

"While our focus was to investigate whether the brain–cognition gap provides information about factors contributing to cognitive resilience, we acknowledge that benchmarking BCG against the brain-age gap in predicting lifestyle and vascular risk factors would be valuable. However, addressing this question lies beyond the scope of the present study, and future work should systematically compare these approaches."

> *(18) Why was only the working memory score used to create brain cognition, and not episodic memory as well? Including both could provide a more comprehensive measure.*

We initially attempted to predict both episodic memory (EM) and working memory (WM). However, EM prediction was only reliable within and across samples for the resting state, whereas WM prediction generalized most strongly from the movie-watching condition. Because COBRA does not include a movie-watching paradigm, we could not evaluate WM prediction across datasets. For this reason, we focused on EM when examining the brain–cognition gap.

> *(19) The PET mediation analysis seemed to come out of the blue. Is there existing literature showing the relationship between striatal D1DR and cognition? If so, did the authors find a similar relationship in the current data? I also suggest rewriting this section to strengthen the justification for the PET mediation analysis.*

We have previously conducted studies in which DA found to be associated with memory (Johansson et al., 2023, Nyberg et al., 2016).

The third aim of our study was to examine whether DA integrity is implicated in brain–cognition gaps (BCG), which we propose as a marker of cognitive resilience. In line with this aim, we found that lower DA receptor availability was associated with larger BCGs (Figure 4). We then asked whether this relationship is mediated by functional signal variability, such that lower DA is linked to reduced signal-to-noise ratio (i.e., greater entropy in functional connectivity), which in turn contributes to less reliable prediction of cognition and, consequently, larger BCGs. Our mediation analysis supports this pathway (see also our reply to Reviewer 1, Comment 6).

Thus, our mediation was not designed to test whether DA predicts episodic memory performance directly, nor whether BCG mediates such a relationship. Instead, we specifically investigated whether the effect of DA on BCG operates through functional variability. We agree that future work could extend our approach by directly examining whether BCG mediates the link between DA and cognitive outcomes. However, in the present study, our primary focus was on testing the mechanistic pathway of DA → entropy → BCG.

> *Minor recommendations:*
>
> *(1) Task-based connections are not truly task-based, as they are around 70-80% related to the resting state, capturing non-task-specific functional connectivity. Task-based connections should refer to techniques that derive task-related connectivity, such as psychophysiological interaction and beta-series correlation. Perhaps use terms like "functional connectivity during tasks."*

Thank you. This has been corrected throughout the manuscript.

> *(2) Are there really two studies? The same MRI was used with the same configurations, and participants were from the same city. The only difference is the age range. It may be more appropriate to refer to this as "across age groups" rather than "cross-datasets."*

Thank you for this comment. While the two samples share some similarities, there are also several marked differences beyond age range. For example, Movie-watching was administered in DyNAMiC but not collected in COBRA. The resting-state fMRI sequence was 12 minutes in DyNAMiC but only 6 minutes in COBRA. Moreover, DyNAMiC included dopamine D1-receptor PET, whereas COBRA assessed dopamine D2-receptor availability. Even the questionnaires used to measure physical activity differed between the two studies. Given these methodological and measurement differences, we believe that referring to them as “cross-datasets” rather than “across age groups” more accurately captures the distinction.

> *(3) What kind of movie is "Cockpit"? Can you explain? Different movies may elicit different patterns of connectivity.*

We apologize for not providing information about the movie, which has been presented in our recent work (Johansson et al., 2023).

The participants’ reactions to the content of the movie were not monitored, but the clips were selected to be as neutral in their content as possible. The content of the movie: Following his termination as a pilot and the end of his marriage, Valle embarks on a quest to secure new employment. Faced with desperation in the job market, he resorts to disguising himself as a woman with the intention of obtaining a position at a company specially seeking a female pilot.

This information is added to the method section.

“During the fMRI session, participants viewed a 12-minute segment from the Swedish comedy film Cockpit (2012). We did not monitor participants’ responses to the movie, and the chosen clips were selected to be relatively neutral in emotional content. The storyline follows Valle, a recently fired pilot whose marriage has ended, as he struggles to find new employment. In a desperate attempt to secure a job at an airline specifically recruiting a female pilot, he presents himself as a woman.”

> *(4) There is a typo in the equation numbering (i.e., two equations are designated as #1).*

We have now corrected the typo.

> *(5) From the discussion: "Importantly, this prediction generalizes across conditions." This is not surprising given the similarity between conditions, with around 70-80% variance.*

We agree with the reviewer that the high similarity of FC across states likely increases the chance of cross-condition generalizability. However, this generalization is not guaranteed for all models. For example, the model trained on FC during movie-watching successfully predicted episodic memory during rest, but it did not generalize to episodic memory during the n-back condition, although movie-watching and n-back FC patterns are themselves highly correlated. Thus, the observed generalization is meaningful in demonstrating that not all models transfer equally well across states.

That said, we have added the following sentence to the Discussion:

“Importantly, this prediction generalizes across conditions and datasets, suggesting that features derived from resting state FC serve as a relatively stable marker of individual differences in EM, though with reduced strength in COBRA. While such generalization is

partly facilitated by the similarity of functional connectivity across states, it is not a trivial outcome. For instance, the model trained on movie-watching data generalized to EM prediction during rest but failed to do so for the n-back condition, even though movie-watching and n-back connectivity patterns are themselves highly correlated. This indicates that successful generalization depends not only on shared variance across states but also on the cognitive processes most relevant to the target behavior."

> *(6) It might be helpful to include some figures for the cognitive tasks used. The description is a bit hard to follow without visual aids.*

Thanks for the comment. We have had a figure describing this in the initial paper about DyNAMiC (Nordin et al., 2022). We have added the Supplementary Figure (Fig S3) in the manuscript.

Fig S3. Overview of the cognitive tests included in the DyNAMiC study. Adopted from Nordin et al. with permission.

> *(7) It may not be appropriate to use the term "cross-validation" here, as one dataset was used for testing and the other for training, but not vice versa (so no "cross" per se).*

We thank the reviewer for pointing this out. We agree that the term "cross-validation" is not precise in this context, since we trained the model in one dataset and tested it in another without performing the reverse. We have revised the manuscript to use the term "external validation" instead of "cross-validation" to more accurately describe our cross-dataset approach.

> *(8) I don't have access to the supplementary materials or code/data, so all of the comments here are based on the main text.*

We have added the supplementary materials and inserted the GitHub link to the code.

> ***Reviewer #3 (Recommendations for the authors):***
>
> *I suggest benchmarking against other simpler algorithms and controlling for memory in the brain cognition gap analyses.*
>
> *The authors might also want to simplify some aspects of the paper. There is a lot going on, which leaves less space to go into enough details for some analyses to warrant claims in the discussion. For example, the authors only compare the deep net to CPM and kernel ridge based on the literature. Direct comparisons would be needed.*

Thanks for the comment. We have made an attempt to address the concerns outlined in the public recommendation. Our study explicitly compares predictive power across different cognitive states (rest, movie watching, n-back), with the aim of identifying the states that best capture individual differences across domains. Thus, our goal was not to propose a universally superior prediction model, but rather to test how brain state influences predictive utility for WM and EM using a deep learning approach. We have revised the manuscript text to make this focus clearer and to avoid any misinterpretation of our aims. Specifically, we removed statements in the Discussion that could be read as suggesting that our deep learning approach outperforms prior machine learning methods. While we compared our model with the connectome predictive modeling (CPM) approach and observed better performance with our deep learning framework, we did not conduct a comprehensive benchmark across all available machine learning methods, nor was this the aim of the present study. Accordingly, we have adjusted the text to avoid implying methodological superiority beyond the scope of our analyses. Furthermore, we have controlled for memory as suggested by the reviewer and outlined in response to reviewer 1.

https://doi.org/10.7554/eLife.104053.2.sa0