frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 25 July 2022
DOI 10.3389/fnins.2022.895637

OPEN ACCESS

EDITED BY
Sung-Ho Lee,
University of North Carolina at Chapel Hill, United States

REVIEWED BY
Rana Muhammad Adnan Ikram,
Hohai University, China
Shijie Zhao,
Northwestern Polytechnical University, China

*CORRESPONDENCE
Mohammad S. E. Sendi
eslampanahsendi@gmail.com
Vince D. Calhoun
vcalhoun@gsu.edu

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 14 March 2022
ACCEPTED 29 June 2022
PUBLISHED 25 July 2022

CITATION
Sendi MSE, Salat DH, Miller RL and Calhoun VD (2022) Two-step clustering-based pipeline for big dynamic functional network connectivity data.
Front. Neurosci. 16:895637.
doi: 10.3389/fnins.2022.895637

Two-step clustering-based pipeline for big dynamic functional network connectivity data

Mohammad S. E. Sendi[1,2,3]*, David H. Salat[4,5], Robyn L. Miller[3,6] and Vince D. Calhoun[1,2,3,6]*

[1]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, United States, [2]Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, United States, [3]Tri-Institutional Center for Translational Research in Neuroimaging and Data Science, Georgia Institute of Technology, Georgia State University, Emory University, Atlanta, GA, United States, [4]Harvard Medical School, Boston, MA, United States, [5]Massachusetts General Hospital, Boston, MA, United States, [6]Department of Computer Science, Georgia State University, Atlanta, GA, United States

Background: Dynamic functional network connectivity (dFNC) estimated from resting-state functional magnetic imaging (rs-fMRI) studies the temporally varying functional integration between brain networks. In a conventional dFNC pipeline, a clustering stage to summarize the connectivity patterns that are transiently but reliably realized over the course of a scanning session. However, identifying the right number of clusters (or states) through a conventional clustering criterion computed by running the algorithm repeatedly over a large range of cluster numbers is time-consuming and requires substantial computational power even for typical dFNC datasets, and the computational demands become prohibitive as datasets become larger and scans longer. Here we developed a new dFNC pipeline based on a two-step clustering approach to analyze large dFNC data without having access to huge computational power.

Methods: In the proposed dFNC pipeline, we implement two-step clustering. In the first step, we randomly use a sub-sample dFNC data and identify several sets of states at different model orders. In the second step, we aggregate all dFNC states estimated from all iterations in the first step and use this to identify the optimum number of clusters using the elbow criteria. Additionally, we use this new reduced dataset and estimate a final set of states by performing a second kmeans clustering on the aggregated dFNC states from the first k-means clustering. To validate the reproducibility of results in the new pipeline, we analyzed four dFNC datasets from the human connectome project (HCP).

Results: We found that both conventional and proposed dFNC pipelines generate similar brain dFNC states across all four sessions with more than 99% similarity. We found that the conventional dFNC pipeline evaluates the clustering order and finds the final dFNC state in 275 min, while this process takes only 11 min for the proposed dFNC pipeline. In other words, the new

pipeline is 25 times faster than the traditional method in finding the optimum number of clusters and finding the final dFNC states. We also found that the new method results in better clustering quality than the conventional approach ($p < 0.001$). We show that the results are replicated across four different datasets from HCP.

Conclusion: We developed a new analytic pipeline that facilitates the analysis of large dFNC datasets without having access to a huge computational power source. We validated the reproducibility of the result across multiple datasets.

KEYWORDS

dynamic functional network connectivity, kmeans clustering, human connectome project, big data, reproducibility

Introduction

In recent decades, blood-oxygenation-level-dependent (BOLD) functional magnetic resonance imaging (fMRI) has provided unique information about brain changes associated with various brain disorders (Heeger and Ress, 2002; Poldrack, 2008; Carbó-Carreté et al., 2020). Functional MRI is a non-invasive imaging technique that identifies localized, time-varying alterations in brain metabolism, such as blood flow and deoxygenated hemoglobin levels (Herberholz et al., 2011). These metabolic changes can be induced by a cognitive task (i.e., task-based fMRI) (Cook et al., 2020) or via unregulated brain fluctuations during rest (i.e., resting-state fMRI). Functional connectivity (FC) or its network analog functional network connectivity (FNC) studies the temporal dependence (typically assessed with correlation) between the BOLD fMRI signal from different brain regions (van den Heuvel and Hulshoff Pol, 2010). The FNC approach uses temporal dependence to infer how various brain networks communicate and may play a significant role in understanding how large-scale neuronal communication in the human brain relates to human behavior (Kalinosky et al., 2019; Cook et al., 2020) and how neurodegenerative diseases alter this relationship (Wang et al., 2019; Yan et al., 2019; Hummer et al., 2020; Quevenco et al., 2020; Vega et al., 2020).

Most previous studies assume FNC is static over time and ignore (average out) brain dynamics (Ioannides, 2007). Indeed, FC is highly dynamic, even during the resting state (Sendi et al., 2021c). In recent years, a new line of research called dynamic functional network connectivity (dFNC) has moved beyond studying the strength of connectivity among brain regions and studied the temporal properties of the FNC (Allen et al., 2014). Dynamic FNC has shown promise as a biomarker for schizophrenia (Sendi et al., 2021a,b), Alzheimer's disease (Sendi et al., 2021c), major depressive disorder (Sendi et al., 2021d), and autism spectrum disorder (Harlalka et al., 2019). It has been shown that dFNC improves the classification of disordered and healthy conditions (Rashid et al., 2016; Saha et al., 2021) and provides more information about neurological and neuropsychiatric disorders pathology than its static counterpart (Menon and Krishnamurthy, 2019).

Figure 1 shows the conventional analytic pipeline that is used for analyzing dFNC information (Rashid et al., 2016; Sendi et al., 2021a,b,c,d). This pipeline contains four main steps. In the first step, we estimate the intrinsic components for the desired brain regions. Second, we calculate the dFNC using a sliding window. In the third step, we concatenate all dFNCs of all subjects and go through an optimization process to find the clustering order based on the elbow criterion. In the fourth step, we estimate the final dFNC for the whole group and state vector for each individual and calculate the dFNC (or temporal) features for statistical analysis.

In the conventional dFNC pipeline, we mainly use a kmeans clustering approach, even though any clustering approach can be used for clustering dFNC information. This is due to kmeans clustering simplicity in implementation and the ability to scale to a large dataset (Fränti and Sieranoja, 2019). Additionally, it has been shown that kmeans clustering is faster than the other methods such as spectral clustering, density-based spatial clustering of applications with noise or DBSCAN, and mean-shift clustering (McInnes and Healy, 2017). But it is still slow and needs substantial computational power when we work on a sizeable dFNC dataset. On the other hand, recently, the availability of extremely large neuroimaging datasets has made the computational burden of clustering dFNC measurements a significant practical challenge. For example, the UK Biobank dataset released neuroimaging data from more than 40,000 participants (Alfaro-Almagro et al., 2021) and has targeted acquiring data from 100,000 individuals (Alfaro-Almagro et al., 2018). Also, it has been discussed that many neuroimaging analytic pipelines are not scalable for massive data sets, including possibly tens, if not hundreds of thousands of participants (van Horn and Toga, 2014).

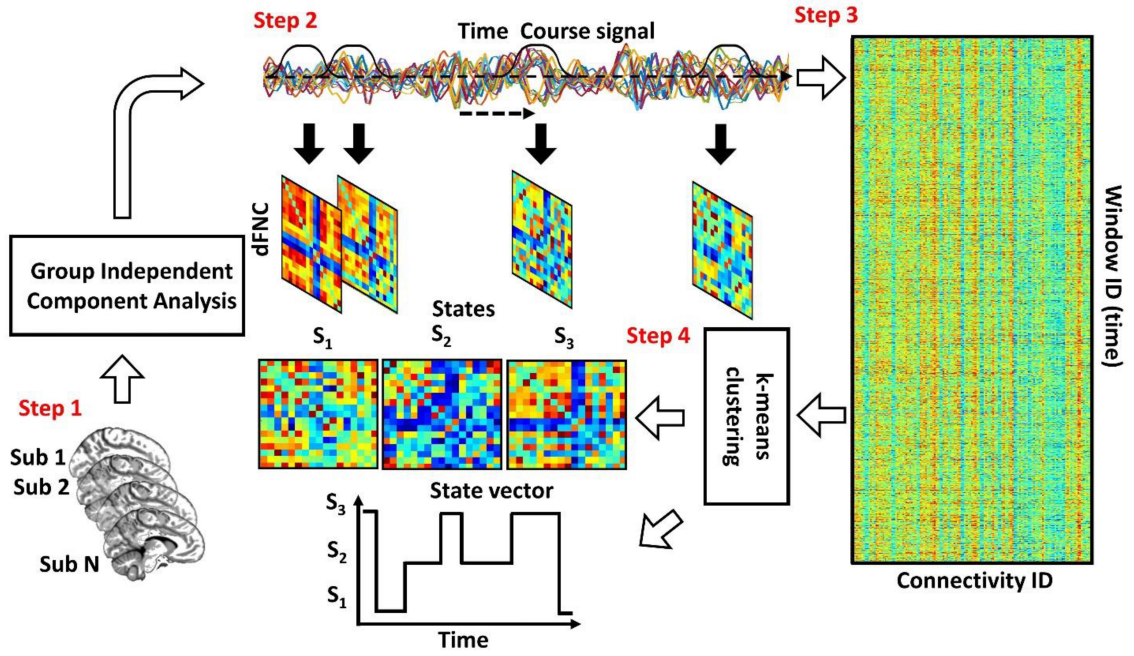

FIGURE 1
The conventional dFNC pipeline. In Step 1, we estimate the independent components using group independent component analysis. In Step 2, we estimate the dFNC using sliding window. In Step 3, we concatenate all dFNCs across all participants. Then, based on elbow criteria, we estimate the cluster order. In Step 4, we use a standard kmeans clustering approach and calculate the dFNC state for group and state vector for everyone.

There are a few disadvantages of using kmeans clustering in the conventional dFNC pipeline. First, we need to load and feed the entire dFNC information to the kmeans clustering to find the final state. Therefore, we need substantial computational power to analyze the sizeable dFNC information. Second, finding the clustering order or the optimum number of dFNC states takes much time in the conventional dFNC pipeline for the large dFNC information. Therefore, developing a framework that can analyze a large dFNC dataset within a reasonable timeframe in a typical cluster computing environment is needed.

This study introduces a new dFNC pipeline that will mitigate the aforementioned issue of the conventional dFNC pipeline in analyzing large dFNC information. The main principle behind the method is to minimize the need to access a huge computational power while we work with large dFNC information. Therefore instead of loading the entire dFNC data to find the optimum state numbers and final dFNC state, we partially load the data through multiple iterations. In more detail, we locally find the states in each iteration and later aggregate all estimated local dFNC states and estimate the final dFNC state for the entire dataset. Therefore, this new approach does not need a large memory to analyze large dFNC data. We evaluated the reproducibility of the results with both standard and proposed dFNC pipelines across four rs-fMRI sessions of human connectome project (HCP) young adults. Additionally, we compared the time needed to find the optimal cluster number with the proposed pipeline vs. the standard one and showed that our approach is faster than the standard method in finding the cluster order. At the same time, both pipelines generate similar dFNC features after finding the final dFNC states.

Material and methods

Our analytic pipeline includes rs-fMRI preprocessing, extracting independent components, calculating dFNC, and estimating the cluster order and dFNC states using the proposed dFNC pipeline. The following subsection describes each step in more detail.

Preprocessing and independent components extraction

We used the statistical parametric mapping (SPM12[1]) running in MATLAB2019 to preprocess the fMRI data. The first five dummy scans were removed before preprocessing. Rigid body motion correction was used to account for the participant's

1 https://www.fil.ion.ucl.ac.uk/spm/

head movement. Then, we used spatial normalization by echo-planar imaging (EPI) template in the standard Montreal Neurological Institute (MNI) space. Finally, a Gaussian kernel was used to smooth the fMRI images using a full width at half maximum (FWHM) of 6 mm. Next, we adapted the Neuromark pipeline to extract intrinsic connectivity networks (ICNs) for each subject (Du et al., 2020). Using this pipeline, we estimated 53 ICNs for each subject and categorized them into seven network domains, including subcortical network (SCN), auditory network (ADN), sensorimotor network (SMN), visual network (VSN), cognitive control network (CCN), the default-mode network (DMN), and cerebellar network (CBN) as shown in **Figure 2**. The details of the extracted ICNs are provided in (Sendi et al., 2021c).

Dynamic functional network connectivity estimation

We used a tapered sliding window and estimated the FC within each window using the Pearson correlation, as shown in Eq. 1.

$$R = \frac{\sum_{n=1}^{N}(x_1 - \overline{x_1})(x_2 - \overline{x_2})}{\sqrt{\sum_{n=1}^{N}(x_1 - \overline{x_1})^2}\sqrt{\sum_{n=1}^{N}(x_2 - \overline{x_2})^2}} \quad (1)$$

where x_1 and x_2 are time-course signals and $\overline{x_1}$ and $\overline{x_2}$ are the mean of x_1 and x_2, respectively. It takes values in the interval $[-1, 1]$ and measures the strength of the linear relationship between x_1 and x_2.

With 53 ICN, the size of each dFNC is 53×53, which equals 1,378 distinct connectivity features. Next, we concatenated dFNC estimates of each window for each subject to form a matrix, called dFNC tensor hereafter, with the size of $T \times F$, where T denotes the number of windows and F donates the number of connectivity features (**Figure 3**).

Proposed dynamic functional network connectivity pipeline

Figure 3 shows the proposed dFNC pipeline for estimating dFNC states of the large dataset. This method includes a few steps. **Step1:** We sub-sample subjects dFNC tensors (m subjects from n subjects per iteration). Then, we run a standard kmeans clustering on the subsampled data with different values of $k = 2, 3, .., L$. The k-means algorithm divides $m \times T$ samples X of each iteration into k disjoint clusters $C_1, C_2, \ldots, C_k$. The cluster centroids μ_i of C_i minimize the within-cluster sum-of-squares criterion as shown in Eq. 2.

$$\min_{\mu_1,..,\mu_k} (\sum_{j=1}^{k}\sum_{i=1}^{mT}(||x_i - \mu_j||^2) \quad (2)$$

We exhaust all subjects by repeating this process r times over disjoint sets of m subjects, where r is equal to $\frac{n}{m}$. In each iteration, we save all cluster centroids for all values of $k \in [2, L]$.

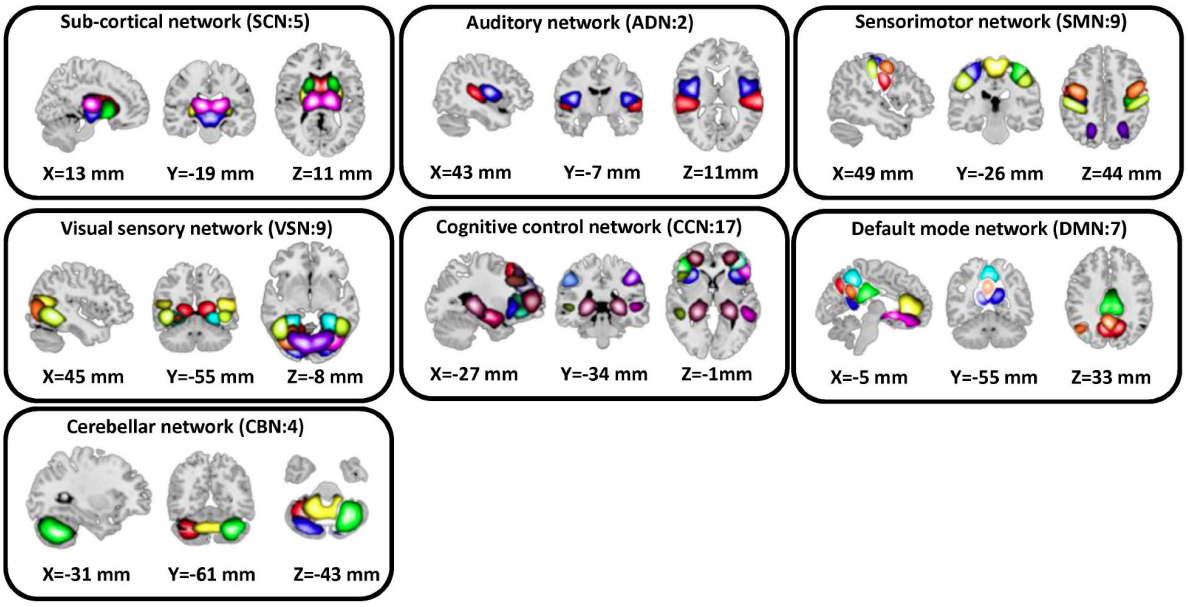

FIGURE 2
Extracted independent components. Fifty three independent components estimated by NeuroMark pipeline. We put them in seven domains including subcortical network (SCN), auditory network (ADN), sensorimotor network (SMN), visual sensory network (VSN), cognitive control network (CCN), default mode network (DMN), and cerebellar network (CBN).

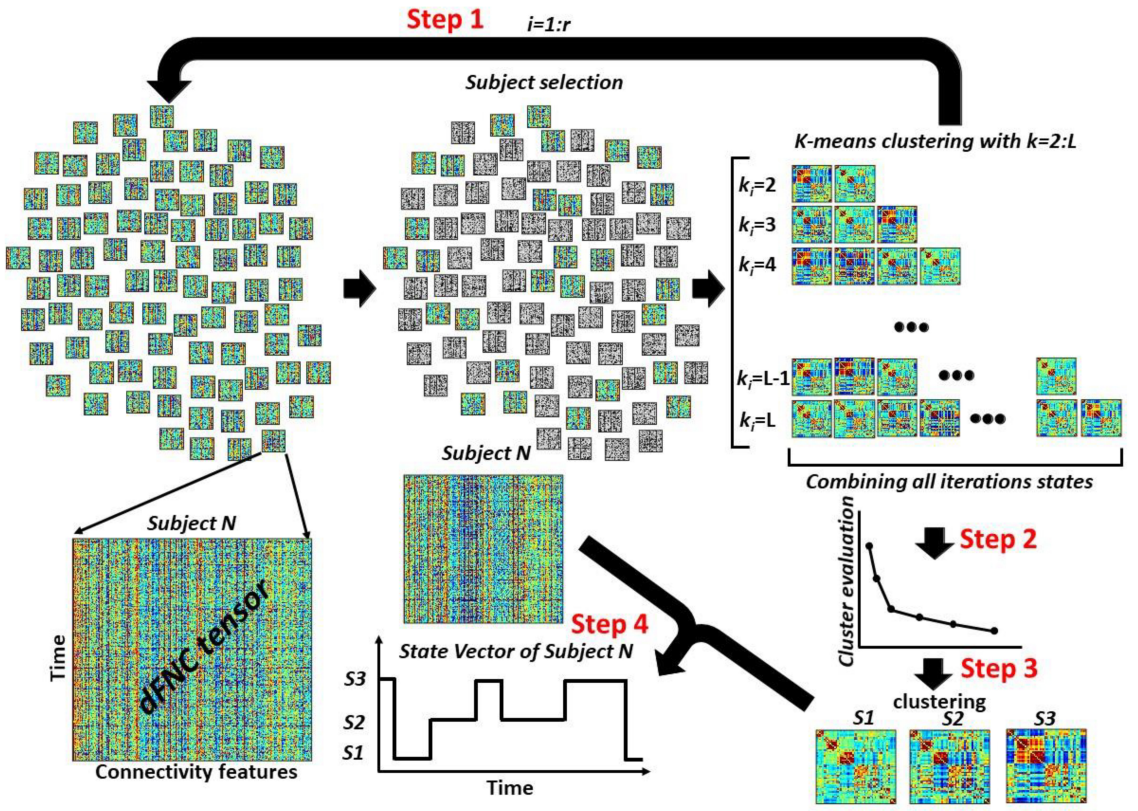

FIGURE 3
The overview of the proposed dFNC pipeline for dFNC state estimation. In Step 1, we select a subsample of dFNC tensor and then used kmeans clustering with k-values from 2 to L and put them into $(\frac{L(L+1)}{2}-1)$. With r iteration, we would have $r(\frac{L(L+1)}{2}-1)$ clusters centroids in total. In Step 2, concatenated all cluster centroids and we use elbow criteria to find the best k-values, called K_{opt} hereafter. In Step 3, using another kmeans clustering approach, we estimated the final dFNC states. In Step 4, we used this final states and found the state vector for each subject.

Therefore, we would have $\frac{L(L+1)}{2}-1$ representative cluster centroids in each iteration. By repeating this process r times, we would have $r(\frac{L(L+1)}{2}-1)$ cluster centroids, reducing the data from the whole dFNC. **Step 2:** We concatenate all centroids estimated from all r iterations. Next, we use the elbow criteria to find the optimum number of clusters using all $r(\frac{L(L+1)}{2}-1)$ observations. **Step 3:** After finding the optimum number of clusters, called K_{opt} hereafter, we use another standard k-means clustering to put all $r(\frac{L(L+1)}{2}-1)$ states into K_{opt} cluster, called final states. **Step 4:** Using the final K_{opt} states, we assign the dFNC of each subject to one of the estimated states and extract the state vector of each participant.

Dynamic functional network connectivity temporal features estimation

We estimated the occupancy rate (OCR) and the number of transitions between states as the representative dFNC temporal features from the state vector. The OCR represents the proportional amount of time each individual spends in a given state for all HCP datasets through both standard and proposed dFNC pipeline.

Clustering quality assessment

To assess the clustering quality for each dFNC data, we calculated the distance between the dFNC data and its associated cluster centroid. Then we calculated the distance between each dFNC sample with the other cluster centroids and then summed them up. Then, we calculated the ratio of the latter to the former one for each dFNC instance, called the distance ratio here. Finally, we averaged all distance ratios out for each participant.

$$R_p = \frac{1}{T}\sum_{i=1}^{T}\frac{d_{i\_sc}}{d_{i\_c}} \quad (3)$$

d_{ic} is the distance between each sample to the cluster centroid of the state the sample belongs. Also, d_{i_sc} is the distance between each sample to other cluster centroids, R_p is the averaged

distance ratio for each participant. It is worth mentioning that a higher ratio means better quality in clustering.

Dataset

To test the proposed method, we used the rs-fMRI and demographic information collected from the 833 young healthy adults (average age: 28.65; range: 22–37 years; female/male: 443/390) from the HCP (Glasser et al., 2016). This dataset is available on the HCP website.[2] The institutional review board from both Washington University and the University of Minnesota approved the study. The rs-fMRI data were collected on a Siemens Skyra 3T with a 32-channel RF receiver head coil. High resolution T2*-weighted functional images were acquired using a gradient-echo EPI sequence with TE = 33.1 ms, TR = 0.72 s, flip angle = 52°, slice thickness = 2 mm, 72 s slices, and 2 mm isotropic voxel, the field of view: 208 × 180 mm (RO × PE), and duration: 14:33 (min: s). For each participant, four separated rs-fMRI sessions (two sessions per day) were acquired that are called HCP1 (session1, day1), HCP2 (session2, day1), HCP3 (session 1, day2), and HCP4 (session2, day12), hereafter. We used all four sessions to evaluate the reproducibility of the result using the proposed dFNC states estimation method. The dFNC size of HCP1, HCP2, HCP3, and HCP4 is 848,827 × 1,378 (8,542 MB), 732,207 × 1,378 (7403 MB), 747,201 × 1,378 (7,555 MB), and 769,692 × 1,378 (7,742 MB), respectively.

Results

Standard and proposed dynamic functional network connectivity pipelines produce similar brain states

The first question we were interested in answering is whether both standard and proposed dFNC pipelines would generate similar dFNC states or not. To test this, we clustered the dFNC data with different L-values in the proposed pipeline (as shown in Figure 3). In the new pipeline, we used 3% of the entire dataset in each iteration. Using elbow criteria, we found that the optimal number of clusters is 2 through both conventional and proposed dFNC pipelines. Then, to evaluate the similarity of dFNC states estimated by the proposed pipeline (with different L) with the states estimated by conventional kmeans, we used the correlation across the matched states as a similarity metric. The similarity between matched states with varying values of L is shown in Figure 4A for all four HCP datasets. We found that the similarity between the matched states generated by both approaches is more than 99%, with any value L of more than five, and the results were reproduced across four HCP datasets. The estimated states with conventional and proposed dFNC pipelines ($L = 6$) are shown in Figure 4B for all HCP datasets.

The proposed dynamic functional network connectivity pipeline finds the optimum cluster number faster than the conventional one

After finding the minimum reliable value of L, we assessed the speed of our method in finding the optimum number of clusters and compared it with the conventional method when it uses the whole dataset. We evaluated the speed of our process with different percentages of data. The result is shown in Figure 5 for HCP1. We found that the new dFNC pipeline is faster when we use a lower percentage of data in each iteration, while the similarity between the matched states estimated with both standard and the proposed pipeline is still more than 98%. Additionally, our proposed method is 25 times faster in funding the cluster order than the traditional method when we use only 0.12% of data (one subject) in each iteration.

Proposed and conventional dynamic functional network connectivity pipelines generate similar dynamic functional network connectivity features

The next question is whether both approaches generate similar dFNC features or not. To assess this, we estimated occupancy rate (OCR), the proportional amount of time each participant spends in a specific state, and the number of between-state transition numbers for each participant in both standard and proposed dFNC pipelines. Both features are estimated from the state vector, which shows the state of the brain at a given time (Figure 3, Step 4). Then, to assess the similarity between the two methods in estimated dFNC features, we calculated the correlation between the result of the two methods. The results are shown in Figures 6A,B for OCR and the number of transitions, respectively, for all four HCP datasets. As Figure 6A shows, the correlation between the estimated OCR by conventional and new dFNC pipelines is more than 0.98 ($p < e^{-10}$). The result was replicated for all four HCP datasets. Additionally, the number of between-state transitions is significantly similar for both methods, and the result was repeated in all HCP datasets. This piece of evidence shows that our new dFNC

2 https://www.humanconnectome.org

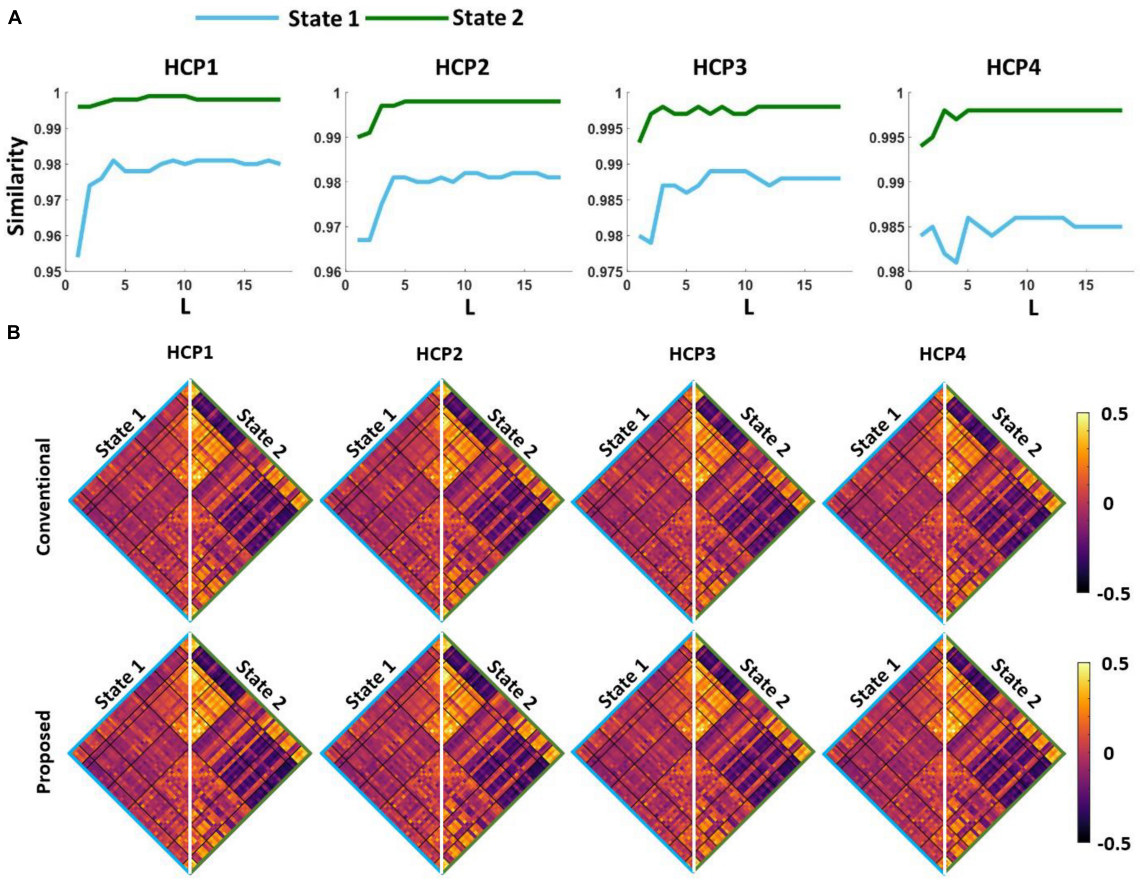

FIGURE 4
The estimated dFNC states with the proposed and conventional pipeline for all HCP datasets. **(A)** We swept the *L*-value in the first kmeans clustering and calculated the similarity between the estimated states with new and conventional method. For any L > 5, we did not find a significant improvement in the similarity between two clustering methods. **(B)** Both new and conventional pipeline generated similar dFNC states in all four HCP datasets.

pipeline produced similar dFNC features as well as the standard kmeans while our method is faster in finding the clustering order and does not require prohibitive levels of computational power.

The proposed dynamic functional network connectivity pipeline has better cluster quality than the standard one

Figure 7 shows the distance ratio of both standard and new dFNC pipelines for the optimum $k = 2$ values in all four HCP sessions. We used a two-sample t-test to compare the distance ratio of the standard vs. the proposed one. We found proposed dFNC pipeline would have better cluster quality than the standard one by having a higher distance ratio ($p < 0.001$, $N = 833$).

Discussion

In this study, we developed an analytic pipeline to analyze large data dFNC information even without having a sophisticated computational resource. There are a few benefits of using this novel framework. (1) In the conventional dFNC pipeline, we need to load the entire dataset regardless of the clustering approach. Loading the entire dFNC data is computationally demanding and slow when using a large dFNC dataset. Our proposed dFNC pipeline does not require loading the entire dataset. This dramatically reduces the required computational resources and the proposed method can be implemented in a computer with small memory size, (2) we showed our method is 25 times faster than the standard method in finding the cluster order and final dFNC states, (3) we validated the reproducibility of the result across four sessions of rs-fMRI data within a population group; and (4) we demonstrated that our approach generates improved clustering quality compared to the standard approach.

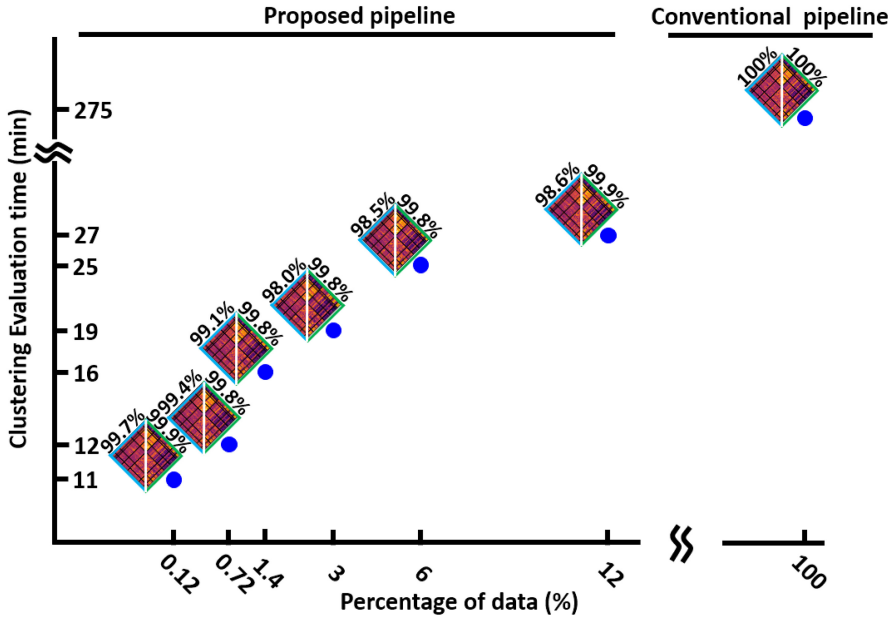

FIGURE 5
The clustering evaluation time with conventional and proposed method. Reducing the percentage of the data used in each iteration of the first step, reduces the evaluation time. The proposed method is 25 times faster the conventional method. The estimated states and their similarity with states estimated from whole data are shown for each percentage of data.

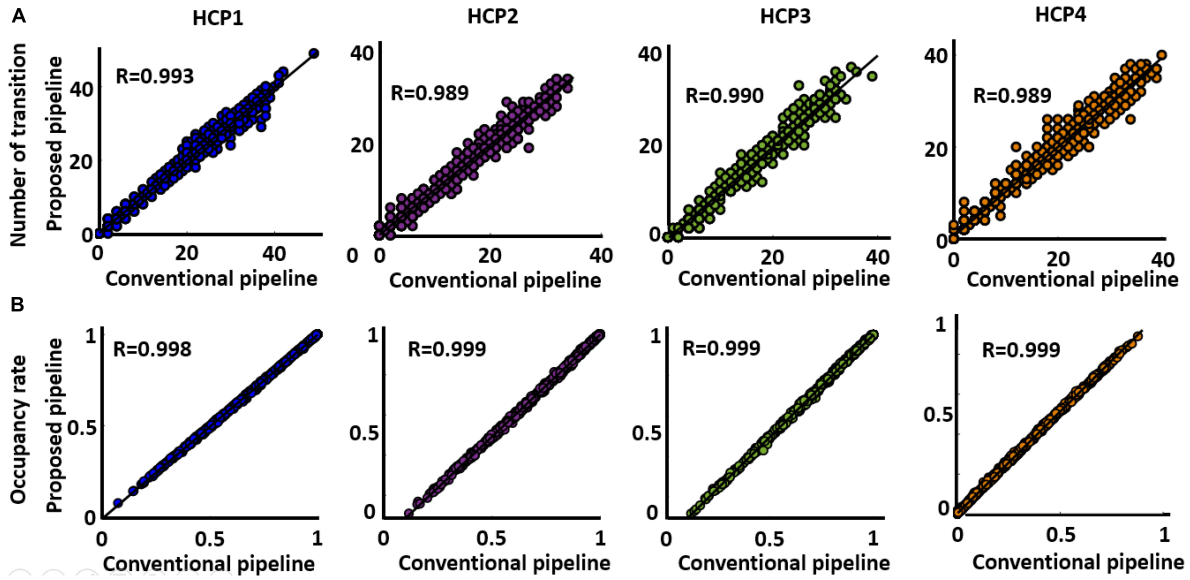

FIGURE 6
Both standard and the proposed dFNC pipeline generated similar dFNC features replicated across four datasets. **(A)** Estimated number of transitions from both standard and proposed pipeline for all HCP datasets. The similarity between the estimated number of transitions from both methods is more than 0.989. **(B)** Estimated occupancy rate (OCR) from both standard and proposed pipeline for all HCP datasets. The similarity between the OCR from both method is more than 0.989 ($p < 0.0001$, $N = 833$).

Unlike the conventional dFNC pipeline in which we need to load the entire dataset, our approach loads a portion of the data in each iteration. Therefore, we reduce both the required memory as well as the computational time. Previous studies proposed the mini-batch kmeans that partially loads the data and does not need expensive computational

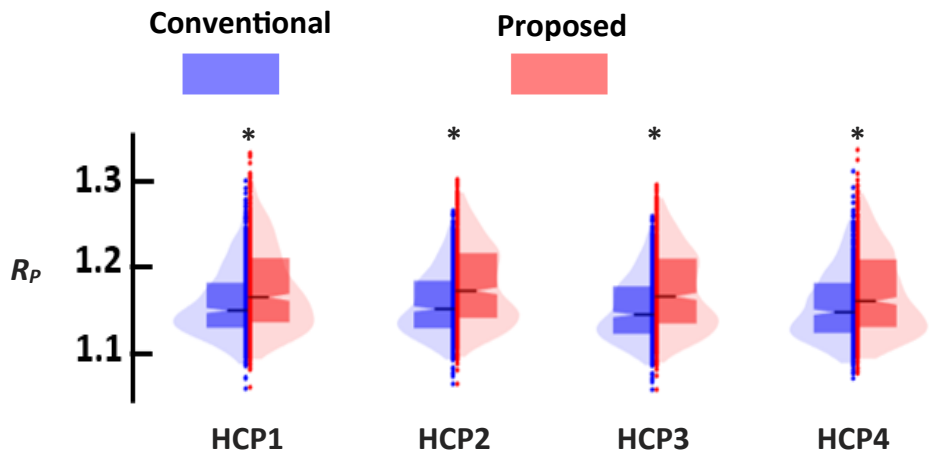

FIGURE 7
The comparison of the cluster quality between standard (blue) and proposed (red) approach. Each column represents that result of each session. In all comparisons, proposed dFNC pipeline had higher cluster quality ($p < 0.001$, $N = 833$). Asterisk (*) represents a significant different between the clustering quality based of old and proposed.

resources (Hicks et al., 2021). But as (Béjar Alonso, 2013) shows, the cluster quality for mini-batch kmeans is reduced compared to standard kmeans clustering, especially when the number of clusters increases. Unlike the mini-batch kmeans approach, our approach reduces the entire clustering process time (**Figure 5**) and increases the clustering quality compared with standard kmeans (**Figure 7**). Additionally, we can adapt mini-batch kmeans or other fast clustering approaches to our proposed dFNC pipeline (Viswanath and Suresh Babu, 2009; Choromanska et al., 2013; Pourkamali-Anaraki and Becker, 2017; Chen et al., 2021). In other words, our new approach is a clustering algorithm agnostic pipeline.

Recent approaches for kmeans clustering of big data have focused on identifying the most informative features for the dataset and then running a kmeans on the reduced set. For example, a recent study reduced the dimension of the data set from p to m ($p > m$) by applying a principal component analysis on the entire dataset followed by a kmeans clustering on the projected dataset (Feldman et al., 2013). This method still needs the whole dataset to be loaded, which requires massive computational power. Additionally, since the kmeans is applied to the project space, we do not have an estimation of the cluster centroid in the original space. However, we can transfer the cluster centroid to the original space, but this estimate is inaccurate and yield lower cluster quality than the standard kmeans approach. But, our approach increases the analysis speed without applying any dimensionality reduction approach. Therefore, our method does not lose any information and yields a lower clustering quality.

Our dFNC pipeline is based on the Neuromark pipeline, a fully automated independent component analysis (ICA) framework that uses spatially constrained ICA to estimate components that are flexible to each subject's data and comparable across individuals (Du et al., 2020). Using the Neuromark pipeline, we calculated the replicated independent components for four HCP sessions. Additionally, we showed that (1) both standard and proposed pipelines generated similar dFNC states in each session of HCP data, and (2) the brain states were replicated across all four sessions using both standard and the proposed dFNC pipeline. The reproducibility of the result across four sessions assessed the robustness of the proposed dFNC pipeline.

There are a few limitations to this study. First, the clustering method in the proposed dFNC pipeline is not limited to kmeans clustering. We can adapt other fast clustering approaches to this pipeline and further improve the computational speed. Second, we did not compare our method's computational speed and clustering quality with different fast clustering approaches (Viswanath and Suresh Babu, 2009; Choromanska et al., 2013; Pourkamali-Anaraki and Becker, 2017; Chen et al., 2021). However, unlike these fast methods, our approach generated a better quality cluster than the standard kmeans clustering method. A future study is needed to compare the results across multiple clustering approaches. Third, we did not propose an algorithmic approach to set the maximum L-value (**Figure 3**). Finding the optimum L-values is done empirically by running the method multiple times to evaluate replicability at different values of L. Future study is needed to develop a mathematical approach to finding the optimum L-values for each dataset. Additionally, we assumed that the preprocessing, group ICA, and estimating dFNC of the large dataset are already done, requiring considerable computation power for a large dataset. A future study is required to develop a methodology to estimate

dFNC information for a large dataset without needing a huge computational power.

Conclusion

Previous dFNC analytics pipelines use standard kmeans clustering, which is ill-suited for big dFNC data. Here, we developed a new dFNC pipeline that reduced the evaluation time for finding the cluster order while we only loaded a portion of the dataset through several iterations. We validated that our method produces similar brain states and dFNC features as the standard method. Additionally, we evaluated the reproducibility of results across four HCP young adult datasets, which showed the high robustness of the proposed method. There are a few advantages of using the proposed approach over the existing method. (1) In the existing pipeline for analyzing dFNC information, we need to load the entire dataset, which requires a huge computational power. But in the proposed dFNC pipeline, we only need to load a small portion of data in each iteration, and it does not need to have access to a big RAM size to analyze the data. (2) The new existing method can find the optimum number of clusters faster than the existing dFNC pipeline, and (3) we showed that the clustering quality is significantly better than what we can get from the conventional dFNC pipeline.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.humanconnectome.org/study/hcp-young-adult.

Ethics statement

This study involving human participants was reviewed and approved by the Washington University and the University of Minnesota.

Author contributions

MS developed the method, analyzed the data, and wrote the manuscript. RM developed the method and provided feedback on the manuscript. DS provided the feedback on the manuscript. VC supervised the study and provided feedback on the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was in part funded by the NSF #2112455 and NIH R01MH123610.

Acknowledgments

We thank those who participated in the HCP study and collected the data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., et al. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* 166, 400–424. doi: 10.1016/j.neuroimage.2017.10.034

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L. R., Bastiani, M., Miller, K. L., et al. (2021). Confound modelling in UK Biobank brain imaging. *Neuroimage* 224:117002. doi: 10.1016/j.neuroimage.2020.117002

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., and Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex* 24, 663–676. doi: 10.1093/cercor/bhs352

Béjar Alonso, J. (2013). *K-means vs Mini Batch K-Means: A Comparison.* Available online at: http://hdl.handle.net/2117/23414 (accessed May, 2013).

Carbó-Carreté, M., Cañete-Massé, C., Peró-Cebollero, M., and Guàrdia-Olmos, J. (2020). Using fMRI to assess brain activity in people with down syndrome:

a systematic review. *Front. Hum. Neurosci.* 14:147. doi: 10.3389/fnhum.2020.00147

Chen, R., Zhao, S., and Liang, M. (2021). A fast multiscale clustering approach based on DBSCAN. *Wirel. Commun. Mobile Comput.* 2021:4071177. doi: 10.1155/2021/4071177

Choromanska, A., Jebara, T., Kim, H., Mohan, M., and Monteleoni, C. (2013). "Fast spectral clustering via the Nyström method," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8139 LNAI, eds S. Jain, R. Munos, F. Stephan, and T. Zeugmann (Berlin: Springer), 367–381. doi: 10.1007/978-3-642-40935-6_26

Cook, M. J., Gardner, A. J., Wojtowicz, M., Williams, W. H., Iverson, G. L., and Stanwell, P. (2020). Task-related functional magnetic resonance imaging activations in patients with acute and subacute mild traumatic brain injury: a coordinate-based meta-analysis. *Neuroimage Clin.* 25:102129. doi: 10.1016/j.nicl.2019.102129

Du, Y., Fu, Z., Sui, J., Gao, S., Xing, Y., Lin, D., et al. (2020). NeuroMark: an automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders. *Neuroimage Clin.* 28:102375. doi: 10.1016/j.nicl.2020.102375

Feldman, D., Schmidt, M., and Sohler, C. (2013). "Turning big data into tiny data: constant-size coresets for k-means, PCA, and projective clustering," in *Proceedings of the 2013 24th Annual ACM-SIAM Symposium Discrete Algorithms*, New Orleans, LA, 1434–1453. doi: 10.1137/18M1209854

Fränti, P., and Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognit.* 93, 95–112. doi: 10.1016/j.patcog.2019.04.014

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L. R., Auerbach, E. J., Behrens, T. E. J., et al. (2016). The Human Connectome Project's neuroimaging approach. *Nat. Neurosci.* 19, 1175–1187. doi: 10.1038/nn.4361

Harlalka, V., Bapi, R. S., Vinod, P. K., and Roy, D. (2019). Atypical flexibility in dynamic functional connectivity quantifies the severity in autism spectrum disorder. *Front. Hum. Neurosci.* 13:6. doi: 10.3389/fnhum.2019.00006

Heeger, D. J., and Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nat. Rev. Neurosci.* 3, 142–151. doi: 10.1038/nrn730

Herberholz, J., Mishra, S. H., Uma, D., Germann, M. W., Edwards, D. H., and Potter, K. (2011). Non-invasive imaging of neuroanatomical structures and neural activation with high-resolution MRI. *Front. Behav. Neurosci.* 5:16. doi: 10.3389/fnbeh.2011.00016

Hicks, S. C., Liu, R., Ni, Y., Purdom, E., and Risso, D. (2021). Mbkmeans: fast clustering for single cell data using mini-batch k-means. *PLoS Comput. Biol.* 17:e1008625. doi: 10.1371/JOURNAL.PCBI.1008625

Hummer, T. A., Yung, M. G., Goñi, J., Conroy, S. K., Francis, M. M., Mehdiyoun, N. F., et al. (2020). Functional network connectivity in early-stage schizophrenia. *Schizophr. Res.* 218, 107–115. doi: 10.1016/j.schres.2020.01.023

Ioannides, A. A. (2007). Dynamic functional connectivity. *Curr. Opin. Neurobiol.* 17, 161–170. doi: 10.1016/j.conb.2007.03.008

Kalinosky, B. T., Vinehout, K., Sotelo, M. R., Hyngstrom, A. S., and Schmit, B. D. (2019). Tasked-based functional brain connectivity in multisensory control of wrist movement after stroke. *Front. Neurol.* 10:609. doi: 10.3389/fneur.2019.00609

McInnes, L., and Healy, J. (2017). "Accelerated hierarchical density based clustering," in *Proceedings of the IEEE International Conference on Data Mining Workshops, ICDMW*, (New Orleans, LA: IEEE), 33–42. doi: 10.1109/ICDMW.2017.12

Menon, S. S., and Krishnamurthy, K. (2019). A comparison of static and dynamic functional connectivities for identifying subjects and biological sex using intrinsic individual brain connectivity. *Sci. Rep.* 9:5729. doi: 10.1038/s41598-019-42090-4

Poldrack, R. A. (2008). The role of fMRI in cognitive neuroscience: Where do we stand? *Curr. Opin. Neurobiol.* 18, 223–227. doi: 10.1016/j.conb.2008.07.006

Pourkamali-Anaraki, F., and Becker, S. (2017). Preconditioned data sparsification for big data with applications to PCA and K-Means. *IEEE Trans. Information Theory* 63, 2954–2974. doi: 10.1109/TIT.2017.2672725

Quevenco, F. C., van Bergen, J. M., Treyer, V., Studer, S. T., Kagerer, S. M., Meyer, R., et al. (2020). Functional brain network connectivity patterns associated with normal cognition at old-age, local β-amyloid, Tau, and APOE4. *Front. Aging Neurosci.* 12:46. doi: 10.3389/fnagi.2020.00046

Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., et al. (2016). Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. *Neuroimage* 134, 645–657. doi: 10.1016/j.neuroimage.2016.04.051

Saha, D. K., Damaraju, E., Rashid, B., Abrol, A., Plis, S. M., and Calhoun, V. D. (2021). A classification-based approach to estimate the number of resting functional magnetic resonance imaging dynamic functional connectivity states. *Brain Connect.* 11, 132–145. doi: 10.1089/brain.2020.0794

Sendi, M. S. E., Zendehrouh, E., Miller, R. L., Fu, Z., Du, Y., Liu, J., et al. (2021c). Alzheimer's disease projection from normal to mild dementia reflected in functional network connectivity: a longitudinal study. *Front. Neural Circuits* 14:593263. doi: 10.3389/fncir.2020.593263

Sendi, M. S. E., Pearlson, G. D., Mathalon, D. H., Ford, J. M., Preda, A., van Erp, T. G. M., et al. (2021a). Multiple overlapping dynamic patterns of the visual sensory network in schizophrenia. *Schizophr. Res.* 228, 103–111. doi: 10.1016/j.schres.2020.11.055

Sendi, M. S. E., Zendehrouh, E., Sui, J., Fu, Z., Zhi, D., Lv, L., et al. (2021d). Abnormal dynamic functional network connectivity estimated from default mode network predicts symptom severity in major depressive disorder. *Brain Connect.* 11, 838–849. doi: 10.1089/brain.2020.0748

Sendi, M. S. E., Zendehrouh, E., Ellis, C. A., Liang, Z., Fu, Z., Mathalon, D. H., et al. (2021b). Aberrant dynamic functional connectivity of default mode network in schizophrenia and links to symptom severity. *Front. Neural Circuits* 15:649417. doi: 10.3389/fncir.2021.649417

van den Heuvel, M. P., and Hulshoff Pol, H. E. (2010). Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* 20, 519–534. doi: 10.1016/j.euroneuro.2010.03.008

van Horn, J. D., and Toga, A. W. (2014). Human neuroimaging as a "Big Data" science. *Brain Imaging Behav.* 8, 323–331. doi: 10.1007/s11682-013-9255-y

Vega, J. N., Taylor, W. D., Gandelman, J. A., Boyd, B. D., Newhouse, P. A., Shokouhi, S., et al. (2020). Persistent intrinsic functional network connectivity alterations in middle-aged and older women with remitted depression. *Front. Psychiatry* 11:62. doi: 10.3389/fpsyt.2020.00062

Viswanath, P., and Suresh Babu, V. (2009). Rough-DBSCAN: a fast hybrid density based clustering method for large data sets. *Pattern Recognit. Lett.* 30, 1477–1488. doi: 10.1016/j.patrec.2009.08.008

Wang, Z., Qiao, K., Chen, G., Sui, D., Dong, H. M., Wang, Y. S., et al. (2019). Functional connectivity changes across the spectrum of subjective cognitive decline, amnestic mild cognitive impairment and Alzheimer's disease. *Front. Neuroinformatics* 13:26. doi: 10.3389/fninf.2019.00026

Yan, C. G., Chen, X., Li, L., Castellanos, F. X., Bai, T. J., Bo, Q. J., et al. (2019). Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proc. Natl. Acad. Sci. U.S.A.* 116, 9078–9083. doi: 10.1073/pnas.1900390116