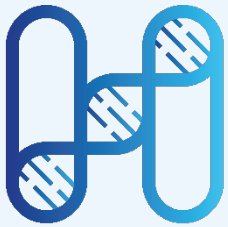

HEREDITARY

HetERogeneous sEmantic Data integratIon for the guT-bRain interplaY

Deliverable D2.3

Linkage and feature extraction from gut-brain, initial evaluation

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by
the European Union

EXECUTIVE SUMMARY

The use case that this deliverable is associated to in the HEREDITARY project aims to replicate and expand earlier findings on gut-brain axis (GBA) interactions by analysing the relationships between brain structure and function, gut microbiota composition, and metabolic parameters. This work leverages multimodal data from two datasets: the Healthy Brain Study (HBS), a cohort of 900 deeply phenotyped individuals aged 30-39 assessed longitudinally across one year, and the Mindset dataset, which includes over 200 psychiatric in-patients with behavioural, clinical, imaging, and microbiome data. The analyses integrate microbiome measures (relative abundance and diversity) with brain connectivity and structural data, using advanced Linked Independent Component Analysis (LICA) and other machine learning methods to investigate multimodal associations. These exploratory methods are meant to set the stage for the discovery of connections between the gut and the brain within HEREDITARY using data analysis techniques on multimodal data.

For this deliverable, preliminary analyses focused on a subsample of 234 healthy participants from the HBS cohort, with exclusions based on incomplete or poor-quality data. After processing, 173 participants were included in the analysis. The LICA model identified 11 independent components (ICs) that demonstrated significant contributions from both resting-state brain connectivity and microbiome relative abundance. These ICs revealed promising GBA patterns, with several components showing multivariate interactions rather than dominance by either modality. For instance, IC 26 highlighted covariance between the relative abundances of specific bacterial genera (e.g., Prevotella, Dialister) and brain networks, including the default mode network (DMN), salience network (SN), and executive control network (ECN). The findings support earlier studies and validate the applicability of LICA for identifying complex GBA interactions.

Future analyses within HEREDITARY will focus on exploring these GBA linkages in the full HBS cohort and conducting hypothesis-driven studies to investigate relationships between stress, diet, microbiome composition, brain connectivity, and eating behaviour. Additionally, the project will expand the integration of GBA data using emerging AI models such as foundation models in fMRI and digital pathology, with a particular focus on extending this type of technology to address gaps in microbiome data analysis. The results of these analyses are expected to advance understanding of gut-brain interactions, particularly in the context of stress, behaviour, and health. Initial findings are encouraging, as they replicate prior results from smaller samples and highlight the potential for multimodal machine learning to uncover novel insights into the GBA. Further work will aim to validate these findings, evaluate longitudinal changes, and explore their implications for understanding stress-related eating and other behaviours. A perspective article on the role of AI in multimodal GBA research is also planned for 2025 to address gaps identified in the literature.

DOCUMENT INFORMATION

Deliverable ID	D2.3
Deliverable Title	Linkage and feature extraction from gut-brain, initial evaluation
Work Package	WP2
Lead Partner	Radboudumc (RUMC)
Due date	31.12.2024
Date of submission	16.12.2024
Type of deliverable	R
Dissemination level	PU

AUTHORS

Name	Organisation
Nils Kohn	Radboudumc
Annemarie Boleij	Radboudumc
Francesco Ciompi	Radboudumc
Giacomo Nebbia (Reviewer)	UCD
Anna Romanovych (Contributor)	UNIPD

REVISION HISTORY

Version	Date	Author	Document history/approvals
0.1	21-11-2024	Nils Kohn, Annemarie Boleij, Francesco Ciompi	Preparation of the first version of the deliverable
0.2	25-11-2024	Giacomo Nebbia	Provision of the feedback on the first version of the document
0.3	08-12-2024	Nils Kohn, Annemarie Boleij, Francesco Ciompi	Changes after review by Giacomo Nebbia From UCD
0.4	10-12-2024	Giacomo Nebbia	Provision of the second round of comments and minor remarks
0.5	12-12-2024	Nils Kohn, Annemarie Boleij, Francesco Ciompi	Making minor changes to the deliverable contents
1.0	16-12-2024	Anna Romanovych	Final edition of the formatting of the document.

Contents

List of Figures

List of Abbreviations

Abbreviation	Original version
HBS	Healthy Brain Study
GBA	Gut-Brain Axis
LICA	Linked Independent Component Analysis
IC	Independent Component (from LICA models)
DMN	Default Mode Network
SN	Salience Network
ECN	Executive Control Network
AI	Artificial Intelligence
fMRI	Functional Magnetic Resonance Imaging
V3-V4	Hypervariable regions of 16S rRNA used for microbiome sequencing
16S rRNA	16S Ribosomal Ribonucleic Acid, used in bacterial taxonomy and microbiome sequencing

1 Project outline

Within the HEREDITARY project, this deliverable is embedded in WP2 and is the initial step in Use Case 4 and Task 2.3. At M12 and M24 of task 2.3, we aimed to provide initial and intermediate results to identify and provide integrated gut-brain linkage and behavioural phenotyping for use in feature extraction for federated learning. For the initial step at M12, we aimed to replicate earlier work on gut-brain linkage in a larger sample and broaden it to the analysis of brain structure and function in relation to gut microbiota relative abundance and diversity measures together with metabolic parameters. For Use Case 4 we will analyse Healthy Brain Study (HBS; Healthy Brain Study Consortium et al., 2022) and for Use Case 5, we will eventually analyse Mindset data (van Eijndhoven et al., 2021). HBS is a cohort of 900 deeply phenotyped, healthy individuals in their 30s measured 3 times throughout one year. Mindset is a dataset of >200 psychiatric in-patients with behavioural, clinical, imaging and faeces sampling. Thus, for this use case and deliverable, we have focused on the HBS dataset.

In future developments, we will integrate the gut microbiota parameters with resting state networks, T1-based grey matter probabilities, Jacobians and diffusion maps using Linked Independent Component Analyses (LICA). In the first step, once the microbiome data is analysed and validated, we aim to characterize gut-brain linkage via LICA with integration of the rich behaviour, clinical and physiological data of the HBS cohort. In a further step, we aim to apply and test other multimodal machine learning algorithms aimed at establishing robustness of modality integration across methods.

1.1 Initial analyses

In an earlier study, we applied a linked independent component analysis (LICA) to microbiota composition and brain connectivity data (Kohn et al., 2021). The LICA has been developed to enable the integration of brain imaging data with other modalities. It is an approach that uses independent component analyses across input modalities to simultaneously explain shared co-variance patterns across brain and non-brain modalities. We previously analysed data from 58 healthy females (mean age = 21.5 years). Magnetic Resonance Imaging data were acquired using resting state functional imaging data. The assessment of gut microbial composition from faeces was based on sequencing of the V4 16S rRNA gene region. We used the LICA model to simultaneously factorize the subjects' large-scale brain networks and microbiome relative abundance data. We found that the abundance of the genus *Prevotella* was associated with the strength of expression of all targeted resting state networks, whereas the genus *Bifidobacterium* was associated with the default mode and frontoparietal-attention networks.

We will, within HEREDITARY, in a first step replicate the LICA association in the full HBS cohort. In the initial analysis reported here, we used a sub-sample of the HBS cohort, which focused on healthy individuals aged 30-39 years from the Nijmegen area in the Netherlands (Healthy Brain Study Consortium et al., 2022). The HBS aims to understand the relationship between the human brain and behaviour, considering various environmental factors such as biological, social, and other contextual influences. To achieve this, the HBS employs an interdisciplinary, longitudinal, and cohort-based sampling approach, collecting multidimensional data in laboratory and ecological settings. The cohort includes cognitive, affective, behavioural, and physiological

assessments, neuroimaging, bio-sampling, questionnaires, ecological momentary assessments, and real-world evaluation measures using wearable devices (Healthy Brain Study Consortium et al., 2022).

The relevant sample in this study consisted of Dutch-speaking healthy individuals aged 30-39 from the Nijmegen region, assessed during 1 year at three sampling time points between 2020 and 2021. An earlier funding had enabled analyses of the gut microbiome of the HBS study in a subsample comprising 234 healthy individuals. The lower age limit of 30 was chosen to exclude neurodevelopmental disorders that might not have fully developed before this age (Sowell et al., 2003). Only individuals with no history of psychiatric illness, no current brain-affecting diseases, no use of brain-targeting medications, and no history of brain surgery were included. Detailed exclusion criteria and ethical considerations for the HBS cohort are available in the Methods/Design section of Healthy Brain Study Consortium et al. (2022).

The multi-modal data relevant to this study included three data sources: 1) neuroimaging data (rs-fMRI), 2) stool samples for gut microbiome analysis, and 3) questionnaires compiling demographic, behavioural, cognitive, and affective information from the HBS sample. The LICA in the current study included data from 189 individuals after the exclusion of forty-five individuals (from the N = 234) due to incomplete MRI data (n = 8), poor-quality MRI data (n = 27), and incomplete faecal data (n = 10). Additionally, in the subsequent analysis of correlations between GBA patterns (independent components (ICs) of the LICA) and behavioural data, the contribution of each subject to each IC had to be aligned with their respective behavioural data through a concatenation process. Consequently, the data from sixteen participants was excluded due to unmatched codes with their behavioural data (n = 9) and incomplete behavioural data that corresponded to their neuroimaging and microbiome data collection session (n = 7). Finally, 173 participants (55 males, 94 females, and 24 of unknown gender) were included in the examination of GBA gender differences and association analyses linking GBA patterns (ICs) to mental health or behaviour and cognition measures.

1.1.1 Initial results

This study selected a LICA model with 47 components, following recommendations by Groves et al. (2011), referred to as “LICA 47”. Of the 47 ICs in LICA 47, 10 showed significant contributions from both resting-state connectivity and microbiome relative abundance (ICs 26, 34, 35, 37, 38, 41, 42, 44, 45, and 47)(Figure 1). These ICs surpassed the minimum contribution threshold for both modalities. Additionally, IC 32, with nearly 20% contribution of each modality, was included due to its balanced feature contributions and due to the exploratory nature of the study, bringing the total to 11 ICs of interest. These ICs suggest potential gut-brain multivariate interactions, unlike unimodal patterns dominated by either brain connectivity or microbiome relative abundance alone, as also shown in Figure 1.

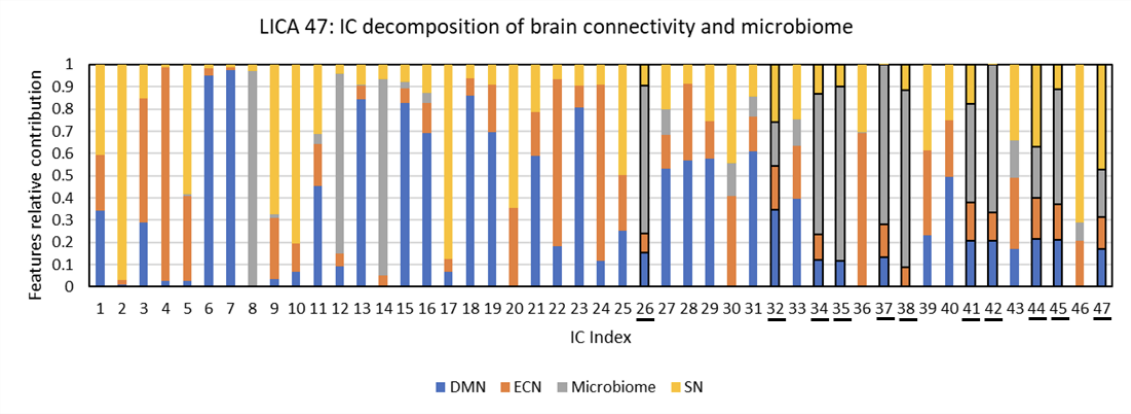

Figure 1:

LICA47 decomposition of the resting state fMRI connectivity and microbiome relative abundance. Displayed are the component indices with the relative contribution of the input modalities. Modalities that mix with at least 10% of brain or microbiome are highlighted. DMN= default mode network, ECN= executive control network, SN= salience network, microbiome = relative abundance at genus level.

Of these eleven components none showed to be explained or highly dominated by a single subject. This means that considering LICA's ICA basis, the method did not force any IC of interest to be solely explained by an outlier type of individual, which would not be representable for the definition of a "healthy GBA pattern" within the general healthy population. Therefore, none of these components was disregarded for further analysis or interpretation. Additionally, sanity checks on brain connectivity renderings for each IC revealed no issues, such as uniform voxel values across the brain data (Kohn et al., 2021). As a result, all eleven ICs were considered valid for further analyses.

The variability in ICs 26, 34, 35, 37, 38, 41, 42, and 45 was primarily influenced by microbiome relative abundance. In contrast, IC 32 was mainly driven by the Default Mode Network (DMN), while ICs 44 and 47 were predominantly influenced by the Salience Network (SN). As this initial analysis does not include all subjects and thereby should be regarded as preliminary, we will limit our report of results to the first numerical component as an example.

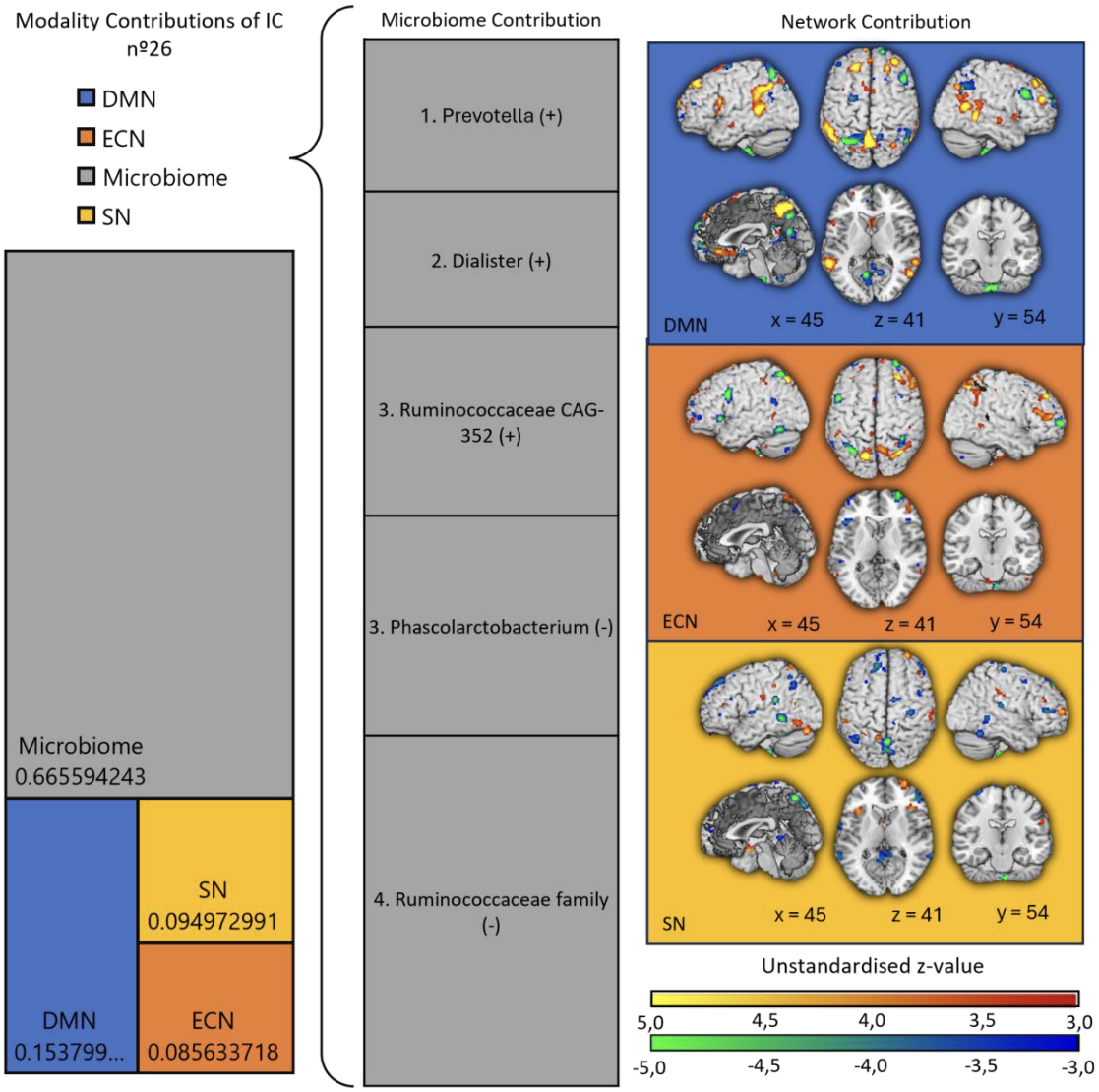

Figure 2:

Report of component number 26. on the left hand the modality contribution is displayed, in the middle panel the bacterial genera that load on this component and on the right hand then brain connectivity of the respective networks is displayed. Microbiome = relative abundance at genus level.

IC 26 demonstrates contributions from the microbiome's relative abundance and three brain networks—DMN, SN, and Executive Control Network (ECN)—in descending order of covariance. The microbiome contributed over 50%, with significant covariance observed in the genera *Prevotella* (+), *Dialister* (+), *Ruminococcaceae* CAG-362 (+), *Phascolarctobacterium* (-), and *Ruminococcaceae* family (-) in order of co-variance. DMN-output voxels showed positive covariance in the medial prefrontal cortex, precuneus angular gyrus and dorsolateral prefrontal cortex which are linked to the ECN, while negative covariance appeared in the bilateral superior parietal lobule (linked to ECN). Negative covariance was sparse, mainly in the left superior parietal lobule and anterior insula (linked to SN). SN hubs showed positive covariance in the bilateral anterior insula, with notable negative voxel intensities in the precuneus and angular gyrus (linked to DMN). The GBA pattern in IC 26 did not show clear network-specific

intra-connectivity but involved overlapping covariance between network hubs and bacterial relative abundances.

In summary, the number of components that show a mixing of gut and brain is promising as well as the pattern of combining microbiome and brain, which are partly replicating earlier results from smaller samples.

1.2 Outlook

The presented analysis and results are meant to set the stage for the larger-scale analysis of multi-modal data for the gut-brain axis use case within HEREDITARY. In March 2024, we have submitted all faecal samples of HBS for DNA extraction and 16s rRNA sequencing of the V3-V4 region with Illumina Novaseq 6000. The raw sequence data have been received and have been pre-processed for further analysis. We are currently finishing the analyses of the complete 1900 microbiome samples of the HBS cohort and expect to have compositional data including diversity matrices and relative abundances ready for integration before February 2025. Based on the microbiome composition results, we will evaluate potential targets for metabolic parameters. This will enable the intermediate analysis as described for M24 of D2.3.

In the upcoming year, the HEREDITARY project will conduct an exploration of the gut brain linkage with LICA presented in this deliverable in the full sample (M24 of D2.3) and conduct two more hypotheses driven studies investigating the relationships between stress, diet, microbiome composition, brain connectivity, and eating behaviour, which were not specified in the deliverables, but aid the aim of integrated gut-brain linkage and behavioural phenotyping in Use Case 4. Study 1 aims to determine whether elevated hedonic food valuation and choices are driven by stress, anxiety, and dietary factors, mediated by multivariate associations between the gut microbiome and reward-related brain connectivity. Study 2 will explore whether these associations explain stress-related eating in daily life. The analysis will employ data from the Healthy Brain Study, focusing on multimodal datasets including resting-state fMRI, gut microbiota profiles, stress responses, and behavioural measures from a food auction task. LICA will integrate these data to uncover joint contributions of brain and gut features in predicting food choices and stress-induced eating patterns. Longitudinal and exploratory approaches will further validate findings and assess robustness.

In parallel with these initial analyses and in order to achieve the final deliverables of Use Case 4, we have been investigating the role of recently developed foundation models in the fMRI and digital pathology space, to be used as data encoders for feature extraction (D2.3 M48). For fMRI, we have identified the recent BrainLM model (Ortega et al. (2023)), whereas for pathology we have explored the UNI (Chen et al. (2024)) and the GigaPath (Xu et al. (2024)) models. We have not found an existing foundation model for gut microbiome data (16s ribosomal RNA), indicating both a gap and an opportunity for this project. In this first year of the project, we have also identified gaps in the literature and perspectives of research on gut-brain interplay, and in particular on the role of AI in this topic. To address this need, we have started working on a literature review and preliminary work on, for example, the role of AI agents on multi-modal data in the gut-brain axis (inspired by the work of Gap et al. (2024) in the biomedical domain), to be used as a base for a perspective manuscript that we are planning to write in 2025.

REFERENCES

The bibliographic entries are arranged in lexicographical order based on the key, following the APA style. This enables us to place the entries and citations in the table and text in any sequence, allowing for later sorting while ensuring consistency.

Key	Reference
Healthy Brain Cohort (2022)	Aarts, E., Akkerman, A., Altgassen, M., Bartels, R., Beckers, D., Bevelander, K., Bijleveld, E., Davidson, E. B., Boleij, A., & Bralten, J. (2022). Protocol of the Healthy Brain Study: An accessible resource for understanding the human brain and how it dynamically and individually operates in its bio-social context (vol 16, e0260952, 2021). *PLOS ONE*, *17*(4). https://doi.org/10.1371/journal.pone.0260952
Van Eijndhoven et al. (2021)	van Eijndhoven, P. F., Collard, R. M., Vrijsen, J. N., Geurts, D. G., Arias-Vasquez, A., Schellekens, A. F., ... & Schene, A. H. (2021). Measuring integrated novel dimensions in neurodevelopmental and stress-related mental disorders (MIND-Set): a cross-sectional comorbidity study from an RDoC perspective. medRxiv, 2021-06. https://doi.org/10.1101/2021.06.05.21256695
Groves et al. (2011)	Groves, A. R., Beckmann, C. F., Smith, S. M., & Woolrich, M. W. (2011). Linked independent component analysis for multimodal data fusion. *NeuroImage*, *54*(3), 2198–2217. https://doi.org/10.1016/j.neuroimage.2010.09.073
Kohn et al. (2021)	Kohn, N., Szopinska-Tokov, J., Llera, A., Beckmann, C., Arias Vasquez, A., & Aarts, E. (2021). Multivariate associative patterns between the gut microbiota and large-scale brain network connectivity. *Gut Microbes*, https://doi.org/10.1101/2020.08.25.266122
Sowell et al. (2003)	Sowell, E. R., Peterson, B. S., Thompson, P. M., Welcome, S. E., Henkenius, A. L., & Toga, A. W. (2003). Mapping cortical change across the human life span. *Nature Neuroscience*, *6*(3), 309–315. https://doi.org/10.1038/nn1008
Ortega et al. (2023)	Ortega Caro, J., Oliveira Fonseca, A.H., Averill, C., Rizvi, S.A., Rosati, M., Cross, J.L., Mittal, P., Zappala, E., Levine, D., Dhodapkar, R.M. and Han, I., 2023. BrainLM: A foundation model for brain activity recordings. bioRxiv, pp.2023-09. https://doi.org/10.1101/2023.09.12.557460
Chen et al. (2024)	Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M. and Williams, M., 2024. Towards a general-purpose foundation model for computational pathology. Nature Medicine, 30(3), pp.850-862. https://doi.org/10.1038/s41591-024-02857-3
Xu et al. (2024)	Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y. and Xu, Y., 2024. A whole-slide foundation model for digital pathology from real-world data. Nature, pp.1-8. https://doi.org/10.1038/s41586-024-07441-w
Gao et al. (2024)	Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J.R., Ektefaie, Y., Kondic, J. and Zitnik, M., 2024. Empowering biomedical

Key	Reference
	discovery with ai agents. *Cell*, *187*(22), pp.6125-6151. https://doi.org/10.1016/j.cell.2024.09.022