Molecular systems biology approaches to investigate mechanisms of gut-brain communication in neurological diseases

Vandemoortele Boris[1,2,3] & Vermeirssen Vanessa[1,2,3,*]

[1] Lab for Computational Biology, Integromics and Gene Regulation (CBIGR), Cancer Research Institute Ghent (CRIG), Ghent, Belgium

[2] Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium

[3] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium

*Corresponding author: vanessa.vermeirssen@ugent.be

Total word count: 6916 (4998 main text)

Short running title: Systems biology for the gut-brain axis

Abstract (250)

While the incidence of neurological disease is increasing worldwide, treatment remains mostly limited to symptom management. The gut-brain axis, which encompasses the communication routes between microbiota, gut, and brain, has emerged as a crucial area of investigation for identifying new preventive and therapeutic targets in neurological disease. Due to the inter-organ, systemic nature of the gut-brain axis, together with the multitude of biomolecules and microbial species involved, molecular systems biology approaches are required to accurately investigate the mechanisms of gut-brain communication. High-throughput omics profiling, together with computational methodologies such as dimensionality reduction or clustering, machine learning, network inference and genome-scale metabolic models, allow to discover novel biomarkers and elucidate mechanistic insights. In this review, we introduce the general concepts of experimental and computational methodologies for gut-brain axis research and discuss their applications, mainly in human cohorts. We further highlight important aspects concerning rational study design, sampling procedures and data modalities relevant for gut-brain communication, strengths and limitations of methodological approaches and some future perspectives. In conclusion, we review how multi-omics analysis, together with advanced data mining, are essential to functionally characterize the gut-brain axis in neurological disease and finally put forward novel preventive or therapeutic strategies.

Introduction

Despite decades of research, treatment of neurological diseases is limited to symptom management, and most drugs achieve only moderate efficacy. Hence, there is an urgent unmet medical need for novel, cost-efficient disease-modifying treatments, of which patients are most likely to benefit if administered early in disease progression. In this respect, the gut-brain axis (GBA) is an exciting research area that opens up possibilities for preventive and therapeutic strategies. The GBA refers to the communication routes and interrelationship between microbiota and inflammation in the gut, and neuroinflammation and neurological disease in the brain[1]. Due to the inter-organ, systemic nature of the GBA, together with the diversity of molecular features and microbial species involved, systems biology approaches are required to capture underlying molecular mechanisms.

Technological advancements in high-throughput molecular profiling enable the cost-efficient, high-throughput analysis of multiple biomolecules and molecular interactions in parallel, such as genome, epigenome, transcriptome, proteome, metabolome and interactome. In addition, latest developments in computational methodologies allow performing multi-omics data integration to capture a multi-modal view and resolve the flow of signaling information across multiple regulatory layers. In this way, information is obtained that exceeds that of the sum of the individual omics, such that GBA molecular systems biology emerges as an exciting new framework to study complex neurological diseases. However, data integration and interpretation have emerged as new bottlenecks. Integrating multi-omics datasets is challenging due to heterogeneity in terms of size, format, dimensionality, noisiness and information content. Generally, we distinguish between multi-stage and multi-dimensional integration, where the former merges different modalities in subsequent steps and the latter combines them at once. The most ideal scenario is to preserve specific properties of each data modality and to integrate different omics data and environmental context simultaneously to identify coordinated behavior between the different levels. Broadly, multi-omics data integration can be based on pairwise statistical association, clustering or dimensionality reduction, network inference, machine learning and composite methodologies. Pairwise statistical association focusses on the interaction between pairs of omics data; clustering or dimensionality reduction transforms the data into a common space of lower dimensions to find patterns in the data; network inference maps the data on to graphs representing interactions between biomolecules; and machine learning predicts or classifies through a model that is optimized iteratively using input data. In addition, prior information in the form of expert biological knowledge can be taken into account, such as in genome-scale metabolic models (GEMs), which are of particular interest in the context of GBA and use gene-protein reaction

rules to link genes encoding enzymes to curated metabolic pathways. These GEMs can be constructed for both the host and the gut microbiome, enabling to model host-microbe and microbe-microbe interactions[2].

Multi-omics integration serves several goals in GBA research: understanding the molecular mechanisms at play, classifying disease versus normal, subtyping within a disease, predicting biomarkers for diagnosis and prognosis, identifying causal effects and detecting molecular drivers. Some methodologies are more suited for a given purpose than others: while clustering or dimensionality reduction and machine learning lean more towards biomarker discovery, patient classification and subtyping, statistical association, network inference and GEMs allow to elucidate molecular mechanisms. Therefore, in the choice of methodology, the biological question has to be kept in mind. In this work we review how omics and multi-omics analysis, together with advanced bioinformatics and machine learning, are essential to functionally characterize the GBA and put forward novel preventive or therapeutic strategies for neurological disorders. First, we address specific experimental design challenges in the context of GBA studies with a focus on human cohorts, and summarize the characteristics of various data modalities. Next, we introduce statistical concepts behind computational methodologies, and show their application in recently published studies focusing on neurological disease. In addition, we discuss the strengths and weaknesses of the investigated approaches, as well as potential challenges and future directions of molecular systems biology in GBA research.

Experimental study design along the gut-brain axis

Different routes have been identified by which crosstalk between gut and brain, and between the immune system and the nervous system, occurs. Gut microbiota and their metabolites can directly activate the vagus nerve, which connects to the central nervous system. They can also influence the generation, maturation and function of immune cells, which can migrate to the brain[1]. In addition, microbial products can induce the release of neurotransmitters and peptide hormones from enteroendocrine cells. Moreover, several microbes and microbial metabolites can pass through the intestinal barrier, enter systemic circulation, cross the brain barriers and act as neuroimmunomodulatory signals in the brain[3,4]. This especially occurs upon dysbiosis and inflammation when permeability of gut and brain barriers is increased. Both cell-autonomous and circulating metabolites serve as signals that transmit information about environmental changes to individual cells to induce appropriate adjustments in gene regulation. This implicates that constructing a complete and accurate GBA model will require multi-omics molecular data sampled from multiple tissues and liquids such as blood, stool and cerebrospinal fluid (CSF)[4,5] (**Figure 1A**). Ideally, one has access to both gut and brain tissue

biopsies, a challenging objective in human cohorts, hence researchers often turn to animal models. In addition, perturbation experiments, which are better suited to demonstrate causality, are more feasible in animal models. The initial design and the types of data generated in a given study thus determine which analyses can be performed, and which biological questions can be answered. Next to the multi-omics data to profile, an additional consideration regarding experimental design is where, in which individuals and when to sample these data.

Extensive and heterogenous patient cohorts better represent the overall population, but will result in a higher within-group variability. Moreover, multi-omics data, including microbial diversity, are influenced by a myriad of confounding factors besides disease status, e.g. age, sex, geographical location, diet, past and current drug treatments, and even physical exercise, often leading to contradictory findings across studies[6,7]. A rigorous analysis of and correction for potential confounding factors is thus essential, and findings should not readily be extrapolated across heterogenous populations. Next, also sampling location and time points must be considered. Longitudinal sampling over a given time period is recommended and allows to monitor the dynamics of disease progression, biomarker presence and the effect of environmental factors. Molecular snapshots taken at a single time point lack these features, but are much easier to obtain, especially if sampling requires invasive procedures or essential tissues. Whole blood and serum, next to CSF, represent an established GBA communication route, making it an interesting resource to identify potential messenger-molecules[4]. Stool samples on the other hand can be considered a functional readout of the gut microbiota[4]. Circulating immune cells can be isolated from blood and CSF, after which they can be phenotypically profiled using multi-omics[8]. Finally, tissue biopsies can be analyzed by high-throughput molecular profiling in different anatomical locations, i.e. gut and brain. Preferably, all omics are measured on the same subjects, resulting in coupled data that is well suited for integration[5]. All samples ideally originate from the same biological material i.e. from the same tissue or liquid biopsy at the same time, in order to avoid batch effects between the different omics data. In a so-called split sample study design, samples taken from the same biological material are divided for different omics analyses. In a replicate-matched study design, samples from different biological replicates within the same experiment are used for different omics analyses e.g. mutually exclusive omics analyses in which sample preparation for one omics impedes profiling of a second omics in the same sample. In the GBA context, often a source-matched study is conducted, where different samples of the same individual are chosen for different analyses e.g. transcriptomics on gut tissue and metabolomics on plasma[5]. Whether these samples are best taken at a single time point depends on the biological question at hand, as different omics modalities are subject to different time scales of change. For example,

changing metabolite levels might result from fast post-translational feedback mechanisms within metabolic pathways, while transcriptomic changes resulting from these altered metabolite levels are likely observable only after a given period of time.

Different data modalities synergistically define the neurological disease phenotype

Uncovering the unknown function of a single gene is a monumental task, but gives only limited insight. Genes do not act in isolation but are embedded in highly complex biological systems. Transcription is facilitated by regulatory factors such as transcription factors, chromatin-modifying enzymes and nucleosome remodeling complexes. Furthermore, metabolites, as nutrients, products of the host proteome or derived from the gut microbiota, have broader roles in cellular signaling than simply being sources of fuel and building blocks. Metabolites are known to influence gene regulation as ligands for signaling or regulatory factors, and as substrates or cofactors of DNA or histone modifying enzymes and chromatin remodelers[9] (**Figure 1B**). Especially short-chain fatty acids (SCFAs) modulate brain homeostasis and neuroinflammation by affecting microglia activation, astrocyte and oligodendrocyte function and Treg expansion through chromatin remodeling, histone deacetylase inhibition and ligand binding to G protein-coupled receptors[10]. As an example, butyrate acts as an endogenous inhibitor of histone deacetylases[11]; and butyrate or butyrate-producing bacteria have also been reported to be differentially abundant in numerous neurological diseases such as multiple sclerosis (MS)[12], major depressive disorder (MDD)[6,7], Alzheimer's disesase (AD)[13] and Parkinson's disease (PD)[14]. Interestingly, these SCFAs are not produced by the host but by specific microbial species, thus linking the gut microbiome to human gene regulation and neurological disease.

Specifically in the GBA context, **microbiomics/metagenomics and metatranscriptomics** data are an essential resource. The analysis of 16S ribosomal RNA , referred to as microbiomics, characterizes the presence of micro-organisms up to genus-level accuracy. However, microbiomics provides insufficient resolution, as distinct species within the *Bacteroides* genus differentially impact depression-like behavior[15]. Whole metagenome shotgun sequencing, although more expensive, provides up to strain-level sensitivity as well as insights into the functional potential of the identified micro-organisms[16]. Metatranscriptomics in its turn reveals which microbial genes are actively being transcribed in a given context or individual. **Metabolomics** data represent the phenotype of an entire biological regulatory cascade due to its implicit integration of genomics, epigenomics, transcriptomics, proteomics and even metagenomics data. Although often profiled in easily accessible bodily fluids such as serum or stool samples, untargeted metabolomics data are extremely complex to interpret and require highly-skilled scientists to annotate mass spectrometry peaks to biological

metabolites[17]. Targeted metabolomics on the other hand are more easily interpretable, but are limited to known metabolites, and thus represent only a fraction of the true metabolic diversity. In addition to metabolomics and microbiome data which can be easily obtained from liquid biopsies, transcriptomes and epigenomes can be profiled from tissue biopsies and circulating immune cells. These data can reveal potential gene regulatory mechanisms through association of gene expression with specific transcription factor binding sites within regions of open chromatin. If proteomics data are added, which are also a complex data type, the effect of post-translational modifications on protein activity and half-life can be incorporated.

Molecular data resulting from transcriptomics, proteomics, metabolomics, epigenomics and microbiomics/metagenomics assays are highly dimensional, implying the amount of molecular features i.e. genes or metabolites measured greatly outnumbers the number of observations. This is referred to as the curse of dimensionality[18]. Additionally, data encounter sparsity at several levels. First, there is sparsity at the sample or patient level, as not all omics are profiled across all samples. Second, there is sparsity within each omics dataset, as not all molecular features are measured in all samples due to technical limitations. Furthermore, different omics suffer from different degrees of sparsity, with mass spectrometry-based techniques often resulting in more sparse data matrices. Data of different modalities result in discrete or continuous, numerical or categorical variables with different ranges that cannot be directly compared such that data first need to be transformed. In addition, more omics data are collected at single cell level instead of at tissue level (i.e. bulk), allowing to dissect complex tissues into specific cell states. However, these data present with high heterogeneity, dimensionality, overdispersion and sparsity[19]. Hence, omics data preprocessing should be performed with great care. This includes quality control by removal of batch effects resulting from technical artefacts and removal of low signal features, normalization across individuals, scaling of the different data modalities, and selection of highly variable features within each dataset.

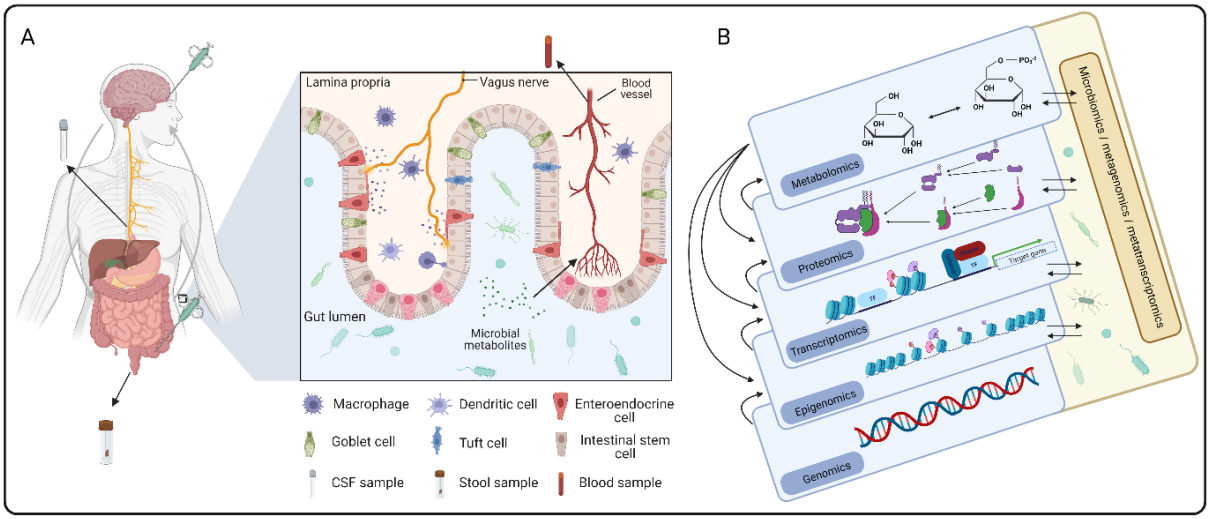

Figure 1: The gut-brain axis encompasses systems-level, inter-organ regulatory processes. A: *The gut-brain axis connects different organs and tissues through three main communication routes. First, the vagus nerve connects the brain to the enteric nervous system, which in its turn is connected to the central nervous system. Second, epithelial and immune cells are in direct contact with microbial cells and molecules, after which immune cells can travel to anatomically distant organs. Third, microbial metabolites can enter the bloodstream and reach the blood brain barrier through systemic circulation. Ideally, omics data are sampled from gut and brain tissue, stool, blood and cerebrospinal fluid.* ***B:*** *Biological systems are regulated through feedback mechanisms between multiple molecular levels, and need to be profiled by different omics techniques. Genetic information is encoded in the DNA, which is tightly packaged into chromatin. Epigenetic modification of histones such as acetylation and methylation can result in open chromatin, which can be actively transcribed by RNA polymerase in combination with transcriptional regulators. mRNA is then translated into protein, which often function in multi-protein complexes. These proteins facilitate cellular metabolism, and metabolites in turn regulate the activity of proteins such as transcription factors and histone modifiers. Created with BioRender.com*

Computational strategies for biomarker discovery and patient stratification

Neuroinflammation is well known to present with gut inflammation and dysbiosis, and increasing evidence indicates gut dysfunction may precede neurological symptoms even by decades[20]. This opens up new possibilities towards early screening . Screening methods find their origin in the differential abundance of specific molecular markers in a patient compared to control group or in healthy versus inflamed tissue. Based on the data modalities, the number of individuals enrolled and the eventual goal of the study, three classes of methods can be used to classify patients and identify descriptive molecular compounds (**Figure 2A, Table 1**).

The first class of methods, already widely applied to GBA, aims to identify **differential abundances** of specific individual molecular features based on a single omics. These statistical tests assess for each individual feature whether it is present in a higher or lower abundancy in one group compared to the other, and whether the observed difference reflects biological signal or random noise. The microbiota can be additionally characterized by two distinct diversity measures: alpha- and beta-diversity. Alpha-diversity is a measure for the richness of a microbial community, i.e. the number of different genera/species/strains, and is represented by indices such as the Shannon or Fisher index[7]. Beta-diversity reflects differences in microbial abundance of specific genera/species/strains between groups. In a mouse model of autism spectrum disorder, differences in stool beta-diversity were reported together with a >40-fold increase of the metabolite 4-ethylphenylsulfate (4EPS) in serum metabolomics compared to a healthy control group, an effect that could be restored by treating mice with *Bacteriodes Fragilis*[21]. A follow-up study reported that gut-derived 4EPS could enter systemic circulation and subsequently the brain, where it altered oligodendrocyte maturation, neuronal myelination and increased anxiety-like behavior[3]. Similar studies using metabolomics, microbiomics/metagenomics and/or transcriptomics have been performed in extensive patient cohorts for other neurological disorders, for example depletion of the *Coprococcus* and *Dialister* taxa has been reported across multiple MDD patient cohorts[6]. However, methods designed to identify differentially abundant molecular features suffer from the dimensionality curse, i.e. the fact that each dataset contains many more features than observations. Also, they fail to identify interactions between features and cannot integrate multiple data modalities.

A second class of methods is **dimensionality reduction and clustering**. During Principle Component Analysis (PCA), the most intuitive approach, each data point in a high-dimensional space is mapped onto a novel set of axes, or principal components (PC), which are chosen through rotation of the original axes to capture maximal variance in the original data. Since the first PC captures the highest proportion of the total variance, a number of PCs can be chosen

to represent the data in lower dimensional space. This implicates that data points in the reduced dimension matrix are actually linear combinations of the original data points, so that they lose their one-on-one relationship with individual molecular features. Other methods such as Principal Coordinate analysis (PCoA), UMAP or t-SNE have been specifically optimized for high-dimensional and sparse biological data and can be used to create intuitive visualizations of underlying data clusters, trajectories and patterns of interest[18]. PCoA is commonly used to visualize microbiome beta-diversity due to its intrinsic visualization of dissimilarity between samples and robust performance on highly sparse data. Using a human discovery cohort, microbiomics and PCoA, Liang and colleagues found significant microbiome differences in individuals characterized by mild or no cognitive impairment, which were mainly caused by species belonging to the *Firmicutes*, *Bacteriodetes* and *Proteobacteria* phyla[22]. Dimensionality reduction techniques can further be used to project distinct data modalities into a common low-dimensional space, facilitating data integration across omics, not only capturing signals shared by all omics data, but also those emerging from the complementarity of the various omics[23]. Multi-omics factor analysis (MOFA) infers a set of (hidden) factors that capture both technical and biological sources of variability across multi-omics data, allowing the identification of sample subgroups through clustering, and identification of highly informative features across multiple omics at once[24]. Clark et al. profiled metabolome, proteome, lipidome, one-carbon metabolism and inflammatory markers in the CSF of a cohort comprising adults with normal cognition, mild cognitive impairment and mild dementia. MOFA identified 5 hidden factors within the multi-omics datasets, to which protein 14-3-3 zeta/delta, clusterin, interleukin-15 and transgelin-2 contributed substantially. Addition of these four MOFA-selected features to a reference classification model for AD pathology resulted in an increased sensitivity and specificity[25].

Machine learning comprises methods such as logistic regression, support vector machines, Bayesian models, random forests (RFs) and boosting, which identify the most informative features in a given dataset by assigning them a higher weight in the final model. Samples in the original dataset are typically divided into a well-balanced training and testing set, after which a model can be iteratively trained until it achieves both good prediction specificity and sensitivity. Logistic regression is designed for binary classification tasks, and model parameters can be interpreted as feature importance. Levi et al. applied logistic regression to serum metabolomics data and achieved near-perfect separation of MS patients and healthy controls[12]. RFs are ensembles of decision trees, which are flowchart-like decision structures that aim to maximize the differences between groups at each additional split. A major advantage of RFs is their high interpretability through feature prioritization methods, which estimate the effect of removing or replacing a given molecular feature during the classification

procedure[26]. Removing or replacing highly-informative features will results in a substantial drop in classification performance, thus identifying the degree to which specific features are characteristic to the groups under study. Finally, gradient boosting decision tree models are a variation to RFs, in which new trees are added to the ensemble of classifiers additively instead of randomly. These have classified dysbiotic and non-dysbiotic microbiomes using predicted metabolite secretion fluxes in a gut inflammation cohort, after which chorismate, D-ribose, L-lactate and phenol were identified as the most informative metabolites[27]. Levi and colleagues further used gradient boosting to predict metabolite levels using only microbiome data, and found 26 metabolites to be associated with the microbiome. One of these, indolepropionate, was present in lower concentrations in the serum of MS patients compared to controls, although there was no significant difference in the abundance of indolepropionate-producing microbiota. However, MS patients' microbiomes had a lower abundance of indolelactate-producing species, an intermediate of tryptophan to indolepropionate catabolism[12].

Computational strategies to elucidate molecular mechanisms of GBA communication in neurological disease

Differential molecular abundances, hidden data patterns or features informative for machine learning in themselves lack biological interpretability and do not result in disease pathology insights. In order to achieve a mechanistic understanding of regulatory processes and to identify causal effects or molecular drivers, the biological relationships between different omics data types needs to be considered (**Figure 2B, Table 1**).

Pairwise integration of biological data can identify interactions between different omics types and thus across regulatory layers in the cell and the organism. Two broad categories can be distinguished, namely genetics of intermediate trait analysis, and correlation analysis between two modalities such as the microbiome and metabolome. The analysis of expression quantitative trait loci (eQTL) and DNA variants links genetic variations to transcriptomic alterations by assessing statistical associations between genomic polymorphisms and the expression levels of, often nearby, genes[28]. However, the analysis of metabolic and microbial trait loci is likely more informative in the GBA context, although this has not yet been reported. Correlation analyses on the other hand can reveal functional interactions between microbiota, metabolites and genes. Within the IRONMET MDD patient cohort, lower dietary and circulating proline levels were associated to lower Patient Health Questionnaire-9 (PHQ9) scores. Circulating proline levels could further be positively associated to species from the *Parabacteroides* and *Prevotella* genera, and negatively with *Actinobacteria* and SCFA-producing species. Remarkably, patients with high dietary proline but low circulating proline

levels had a microbial signature associated with low PHQ9 scores, suggesting systems-level interactions between the microbiome, metabolome and MDD symptoms[7]. Of particular interest in the GBA context is the analysis of correlations between microbial abundances and circulating metabolites. Such analyses revealed three 'metabolite type-bacterial taxa' correlated pairs in a model of MDD, next to associations between SCFAs and differentially abundant microbial genera[29]. Using longitudinal multi-omics data in an MS patient cohort, Cantoni et al. found significant correlations between the gut microbiome and host blood immune profiles in healthy controls but not in patients, suggesting disruptions of immune-microbiome homeostatic interactions in MS[30]. In an AD patient cohort, correlations between differentially abundant microbial genera and CSF biomarkers have been reported, such that easily accessible microbiome data might provide an alternative for the invasive lumbar puncture in the future[31]. Finally, Liang et al. found significant correlations between serum metabolites and metagenomic pathways enriched in individuals with impaired cognition[22], which were identified using a LASSO logistic regression model[22]. However, although statistical associations reveal insightful interactions between the microbiome, metabolome, gene expression and phenotypic traits, they cannot identify causative features. Additional (perturbation) experiments, longitudinal data or more advanced bioinformatics frameworks, often including expert biological knowledge as prior information, are needed to identify causality and true biological mechanisms of action.

Integrated regulatory networks provide an intuitive method to study molecular interactions across omics types. These networks are constructed of nodes, which represent omics features, and edges between the nodes reflecting correlations, regulatory or functional interactions. Uncovering these edges in a given biological context comes down to unraveling statistical dependencies between molecular features, using methods such as Bayesian statistics, regression, mutual information and correlation on transcriptome data and requires a large number of bulk observations. Single cell omics datasets on the other hand inherently contain the required statistical variability between features such that patient-specific regulatory networks can be constructed from a single sample, and regulatory programs can be inferred in a cell-type specific manner. The inclusion of multi-omics in the network inference process, such as regulator binding information or protein-protein interactions, results in more accurate biological networks[32,33]. Although the search for the best method is still subject of research, we and others have shown that different methods add complementary information to the inference of robust regulatory networks, which advocates for the construction of ensemble networks[34]. The popular methodology Weighted Gene Coexpression Network Analysis (WGCNA) constructs coexpression modules from single omics, after which associations between coexpression modules, other omics or phenotypic traits can be assessed[35]. This has allowed

clustering of both serum and CSF metabolites into coexpression modules that could be associated to MS severity[8] or neurotoxicity[36]. Similarly, coexpression modules have been constructed from blood metabolomics in an AD patient cohort, revealing significant correlations between amino acids, short-, medium- and long-chain acylcarnitines on the one hand, and AD severity and cognitive traits on the other hand. Further integration with transcriptome data highlighted *cpt1a*, which encodes a protein involved in the rate-limiting step of acylcarnitine transport into mitochondria, and might thus account for the observed accumulation of medium/long-chain acylcarnitines during AD progression[37]. Badam et al. compared several module detection tools such as WGCNA and clique-based methods to propose a framework for multi-omics and genetic risk factor integration in MS[38]. Zheng and colleagues made use of non-human primates displaying depressive-like behavior and WGCNA to construct separate metagenomics and metabolomics coexpression modules, which were correlated to depressive-like behavior, revealing a functional interaction between the gut microbiome, lipid metabolism and depressive-like behavior[39]. Interestingly, WGCNA can also be applied on different tissues and/or liquids, making it appealing in the GBA context. However, similar to studies that assess correlations between features in omics datasets, WGCNA represents merely associations. Directed regulatory networks, in which upstream regulators and downstream targets are characterized, add a more causal interpretation of the data. The integrative multi-omics module network inference algorithm Lemon-Tree builds coexpression modules across samples using a model-based Gibbs sampler, after which potential regulators are assigned through an ensemble of decision trees[40]. Interestingly, expression data need not come from a single tissue nor need potential regulators come from the same data modality, as they can be any omics profiled on the same samples[40,41]. This is particularly useful in the GBA context, as it allows to model gene expression in both gut and brain, and assign microbial signatures or circulating metabolites as potential regulators. Regulatory network inference on neuroinflammation cohorts can thus theoretically be exploited to predict causal relationships between the gut microbiota, circulating metabolites and gene expression patterns in the brain. Overall, a critical aspect is to go beyond statistical associations and identify direct causal relationships. Indeed, as insightful as associations between microbial abundances, metabolite levels and transcriptional programs are, these fail to provide information on the directionality of the interaction and cannot identify driving mechanisms. Transcription dynamics information which is inherently present in single cell-or bulk time series transcriptome data, enables causal network inference, as does the inclusion of other omics data, such as chromatin accessibility data or regulator binding information.

Finally, expert-curated biological databases allow to exploit biological knowledge to achieve context-specific causality. Several databases provide curated knowledge on transcriptional

regulatory interactions (DoRothEA, OmniPath), interactions between metabolites and transcriptional regulators (STITCHdb, ASD), kinase-substrate interactions and metabolic interactions (Recon3D), or interactions between genes and metabolic pathways (KEGG), and these can be used as prior knowledge in integrated regulatory networks or pathway-based models[32]. **GEMs**, such as the human Recon3D, use gene-protein-reaction associations based on known genomic information and metabolic pathways to construct a stoichiometry matrix of metabolites and reactions, creating a mathematical representation of a given metabolic network. Through the addition of constraints such as nutrient input, upper- and lower-bound reaction fluxes, and additional omics data, mainly transcriptomics, an optimization approach can be applied to infer context-specific metabolic pathway activity[42]. Especially exciting in the GBA context, GEMs can also be constructed for the microbiota, allowing to model microbe-microbe and host-microbe metabolic interactions[43,44]. The inclusion of personalized microbiome, transcriptome and metabolome data allows contextualizing GEMs to patient-specific models, in which the effects of dysbiosis on for example secreted and circulating metabolites can be studied[27]. These models can be further extended towards whole-body metabolic reconstructions including tissue-specific GEMs and transport routes between anatomically distant organ systems[42]. Using microbiome data and resulting personalized microbial community models, Baldini and colleagues identified PD associated changes in the predicted secretion potential of nine microbial metabolites including GABA, an effect mainly explained by the *Akkermansia*, *Acidaminococcus* and *Roseburia* genera[45]. Furthermore, personalized microbial community GEMs have been used to describe increased circulating homoserine levels and altered sulfur metabolism in PD patients. These changes could largely be explained by differences in *Akkermansia muciniphila* abundance, as this species accounted for over 50% of the variance in predicted secretion potential of methionine and hydrogen sulfide[14].

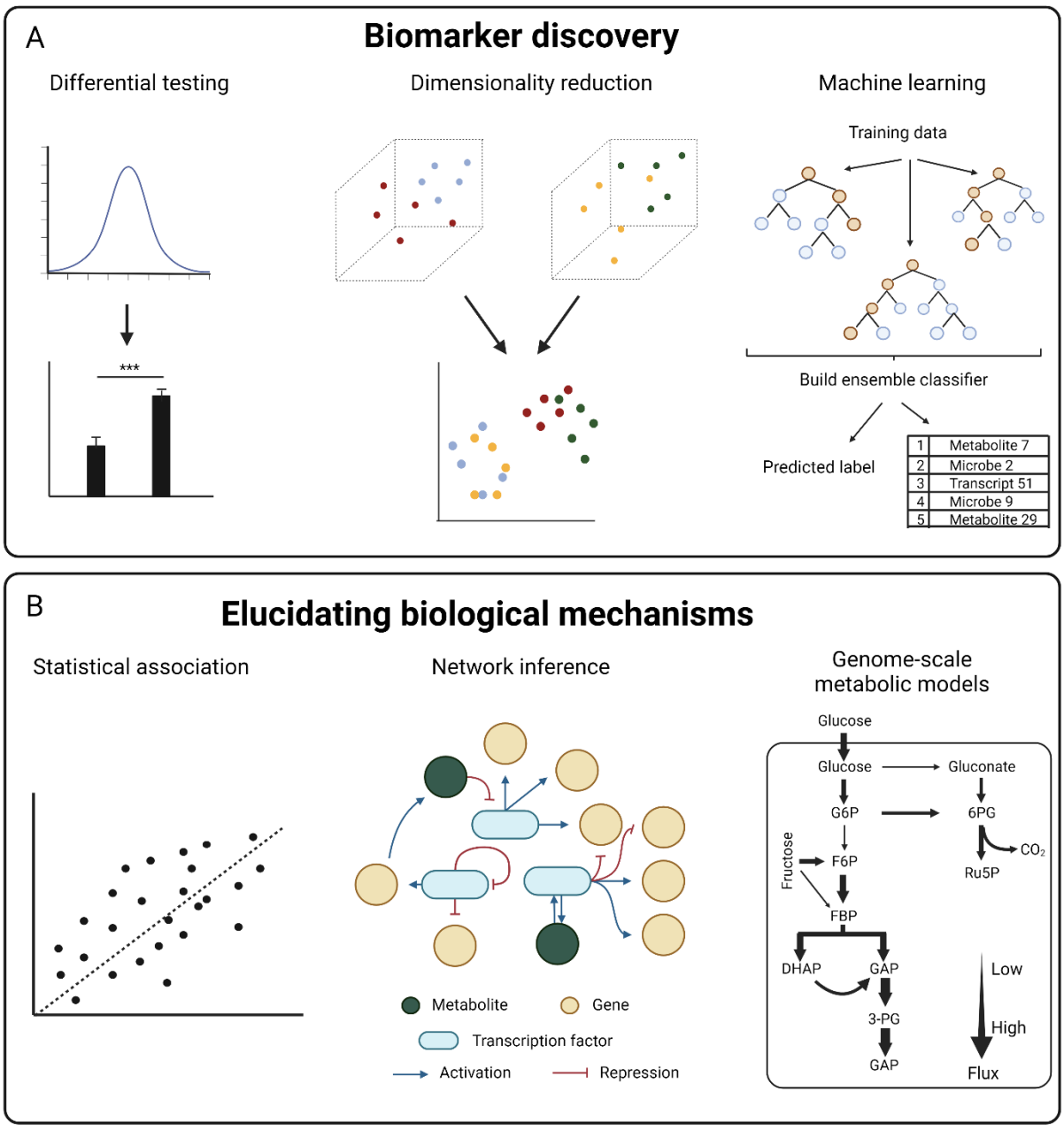

Figure 2: Overview of computational multi-omics analysis methods for biomarker discovery or the elucidation of biological mechanisms. A: *Differential testing identifies features that are differentially abundant in one group compared to another, and assesses whether this is the effect of biological signal or random noise. Dimensionality reduction methods allow to project different data modalities in a common latent space, and thus analyze them together. Machine learning allows to pinpoint biomarkers and classify observations into (sub)groups. A major advantage is the interpretability through feature prioritization methods.* ***B:*** *Pairwise statistical association models interactions between different omics but cannot identify causality. Network inference tools uncover regulatory interactions in a data-driven manner. Genome-scale metabolic models allow to study flux through metabolic networks and model host-microbe interactions. Created with BioRender.com*

Conclusion & future perspectives

Due to technological advancements in high-throughput biology and computer science, GBA is becoming an exciting field of study in neurological disease. Gut-brain communication involves multiple organs, and matched sampling of the appropriate tissues and liquids at suitable time points in extensive patient cohorts is recommended. Experimental design must be carefully considered before and during each study to adequately deal with sample and omics heterogeneity and confounding factors. Ideally, all omics are longitudinally profiled in all individuals, and metadata must be collected and shared conform ethical standards. The study design determines the downstream analysis and the conclusions that can ultimately be drawn. Different data modalities complement one another in the description of the complete neurological disease phenotype and allow for an enhanced biomarker discovery and the elucidation of mechanistic insights. Since the generation of multi-omics data from patient cohorts is a costly process often requiring valuable biological material, data should be shared and made publicly available according to FAIR principles (Findable, Accessible, Interoperable and Reusable)[46]. This should allow for a more robust biomarker identification, as current studies often suffer from limited sample sizes and incomplete control for confounding factors[6,7]. Advancements in state-of-the art omics profiling techniques at single molecule, single cell and spatial level hold great promise, as these enable to study lowly-abundant molecular features, rare cell types and spatial heterogeneity within tissues. Furthermore, improved untargeted metabolomics analysis techniques will results in many more metabolites being profiled, as today only a fraction of the total metabolomic diversity can accurately be identified. In addition, stable isotope tracing in vivo with 13C-labeled nutrients or metabolites[47] is a powerful methodology to demonstrate the causality and route of gut-brain communication in neurological disease.

Novel joint dimensionality reduction and machine learning are increasingly being applied in GBA to discover novel biomarkers. With the increase in multi-omics data for GBA, deep learning approaches bear great potential to detect novel biomarkers. Integrated network inference methods like WGCNA and Lemon-Tree are especially interesting in the GBA context, since they enable multi-omics integration across tissues revealing mechanistic insights. Efforts such as the Virtual Metabolic Human[49] database provide an invaluable resource for data integration, but knowledge-based models by themselves are restricted to curated biological information, and are therefore limited in uncovering novel biology. Ideally data integration is a combination of supervised and unsupervised learning, such as the combination of unsupervised network inference and supervised GEMs, thus contextualizing and extending expert biological knowledge in a data-driven manner[50]. This should result in more accurate

GBA models, allowing wet-lab researchers to prioritize hypotheses and most efficiently make use of available resources.

Overall, integration of different data modalities, especially when profiled in distinct organ systems, is still in its infancy and often lacks robustness. Currently, there is no optimal tool that is broadly applicable to different types of research questions, and general guidance in the field is lacking. Furthermore, there is an urgent need for the development of novel computational approaches tailored towards the study design and data complexity inherent to GBA research. Today, molecular systems biology and omics integration have already revealed significant insights regarding gut-brain communication for numerous neurological diseases, as reviewed here. The GBA can certainly be considered a basis for treating neurological diseases, and likely holds great potential towards the development of disease-modifying therapeutics and personalized medicine.

Acknowledgements

None.

Conflict of interest

Disclosure: None.

Funding

This work was supported by a grant from the Ghent University Special Research Fund BOF/STA/201909/030 'Multi-omics data integration to elucidate the causes of complex diseases'.

Author contributions

Author contributions are reported according to the CRediT taxonomy (BV = Boris Vandemoortele, VV = Vanessa Vermeirssen). Conceptualization: BV, VV; Supervision: VV; Visualization: BV; Writing - original draft: BV, VV; Writing – review and editing: BV, VV.

Data availability statement

Not applicable.

ORCID

Boris Vandemoortele https://orcid.org/0000-0002-4352-0765

Vanessa Vermeirssen https://orcid.org/0000-0002-1975-0712

References

1. Agirman, G., Yu, K. B. & Hsiao, E. Y. Signaling inflammation across the gut-brain axis. *Science* **374**, 1087–1092 (2021).
2. Richelle, A. *et al.* Model-based assessment of mammalian cell metabolic functionalities using omics data. *Cell Rep. Methods* **1**, 100040 (2021).
3. Needham, B. D. *et al.* A gut-derived metabolite alters brain activity and anxiety behaviour in mice. *Nature* 1–7 (2022) doi:10.1038/s41586-022-04396-8.
4. Lai, Y. *et al.* High-coverage metabolomics uncovers microbiota-driven biochemical landscape of interorgan transport and gut-brain communication in mice. *Nat. Commun.* **12**, 6000 (2021).
5. Cavill, R., Jennen, D., Kleinjans, J. & Briedé, J. J. Transcriptomic and metabolomic data integration. *Brief. Bioinform.* **17**, 891–901 (2016).
6. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4**, 623–632 (2019).
7. Mayneris-Perxachs, J. *et al.* Microbiota alterations in proline metabolism impact depression. *Cell Metab.* **34**, 681-701.e10 (2022).
8. Fitzgerald, K. C. *et al.* Multi-omic evaluation of metabolic alterations in multiple sclerosis identifies shifts in aromatic amino acid metabolism. *Cell Rep. Med.* **2**, 100424 (2021).
9. Schvartzman, J. M., Thompson, C. B. & Finley, L. W. S. Metabolic regulation of chromatin modifications and gene expression. *J. Cell Biol.* **217**, 2247–2259 (2018).
10. Dalile, B., Van Oudenhove, L., Vervliet, B. & Verbeke, K. The role of short-chain fatty acids in microbiota–gut–brain communication. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 461–478 (2019).
11. Davie, J. R. Inhibition of histone deacetylase activity by butyrate. *J. Nutr.* **133**, 2485S-2493S (2003).
12. Levi, I. *et al.* Potential role of indolelactate and butyrate in multiple sclerosis revealed by integrated microbiome-metabolome analysis. *Cell Rep. Med.* **2**, 100246 (2021).

13. Liu, J. *et al.* Neuroprotective Effects of Clostridium butyricum against Vascular Dementia in Mice via Metabolic Butyrate. *BioMed Res. Int.* **2015**, 412946 (2015).
14. Hertel, J. *et al.* Integrated Analyses of Microbiome and Longitudinal Metabolome Data Reveal Microbial-Host Interactions on Sulfur Metabolism in Parkinson's Disease. *Cell Rep.* **29**, 1767-1777.e8 (2019).
15. Zhang, Y. *et al.* Bacteroides species differentially modulate depression-like behavior via gut-brain metabolic signaling. *Brain. Behav. Immun.* **102**, 11–22 (2022).
16. Pérez-Cobas, A. E., Gomez-Valero, L. & Buchrieser, C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microb. Genomics* **6**, mgen000409 (2020).
17. De Paepe, E. *et al.* A validated multi-matrix platform for metabolomic fingerprinting of human urine, feces and plasma using ultra-high performance liquid-chromatography coupled to hybrid orbitrap high-resolution mass spectrometry. *Anal. Chim. Acta* **1033**, 108–118 (2018).
18. Malepathirana, T., Senanayake, D., Vidanaarachchi, R., Gautam, V. & Halgamuge, S. Dimensionality reduction for visualizing high-dimensional biological data. *Biosystems* **220**, 104749 (2022).
19. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
20. Pellegrini, C. *et al.* Prodromal Intestinal Events in Alzheimer's Disease (AD): Colonic Dysmotility and Inflammation Are Associated with Enteric AD-Related Protein Deposition. *Int. J. Mol. Sci.* **21**, 3523 (2020).
21. Hsiao, E. Y. *et al.* The microbiota modulates gut physiology and behavioral abnormalities associated with autism. *Cell* **155**, 1451–1463 (2013).
22. Liang, X. *et al.* Gut microbiome, cognitive function and brain structure: a multi-omics integration analysis. *Transl. Neurodegener.* **11**, 49 (2022).
23. Cantini, L. *et al.* Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).

24. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
25. Clark, C., Dayon, L., Masoodi, M., Bowman, G. L. & Popp, J. An integrative multi-omics approach reveals new central nervous system pathway alterations in Alzheimer's disease. *Alzheimers Res. Ther.* **13**, 71 (2021).
26. Fabris, F., Doherty, A., Palmer, D., de Magalhães, J. P. & Freitas, A. A. A new approach for interpreting Random Forest models and its application to the biology of ageing. *Bioinformatics* **34**, 2449–2456 (2018).
27. Heinken, A., Hertel, J. & Thiele, I. Metabolic modelling reveals broad changes in gut microbial metabolism in inflammatory bowel disease patients with dysbiosis. *Npj Syst. Biol. Appl.* **7**, 1–11 (2021).
28. Patel, D. *et al.* Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl. Psychiatry* **11**, 1–17 (2021).
29. Tian, T. *et al.* Multi-omics data reveals the disturbance of glycerophospholipid metabolism caused by disordered gut microbiota in depressed mice. *J. Adv. Res.* **39**, 135–145 (2021).
30. Cantoni, C. *et al.* Alterations of host-gut microbiome interactions in multiple sclerosis. *eBioMedicine* **76**, (2022).
31. Vogt, N. M. *et al.* Gut microbiome alterations in Alzheimer's disease. *Sci. Rep.* **7**, 13537 (2017).
32. Dugourd. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol. Syst. Biol.* **17**, e9730 (2021).
33. Loers, J. U. & Vermeirssen, V. SUBATOMIC: a SUbgraph BAsed mulTi-OMIcs clustering framework to analyze integrated multi-edge networks. *BMC Bioinformatics* **23**, 363 (2022).
34. Vermeirssen, V., De Clercq, I., Van Parys, T., Van Breusegem, F. & Van de Peer, Y. Arabidopsis Ensemble Reverse-Engineered Gene Regulatory Network Discloses Interconnected Transcription Factors in Oxidative Stress[W]. *Plant Cell* **26**, 4656–4679 (2014).

35. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **1**, 54 (2007).
36. Ntranos, A. *et al.* Bacterial neurotoxic metabolites in multiple sclerosis cerebrospinal fluid and plasma. *Brain* awab320 (2021) doi:10.1093/brain/awab320.
37. Horgusluoglu, E. *et al.* Integrative metabolomics-genomics approach reveals key metabolic pathways and regulators of Alzheimer's disease. *Alzheimers Dement.* **n/a**, (2021).
38. Badam, T. V. S. *et al.* A validated generally applicable approach using the systematic assessment of disease modules by GWAS reveals a multi-omic module strongly associated with risk factors in multiple sclerosis. *BMC Genomics* **22**, 631 (2021).
39. Zheng, P. *et al.* The gut microbiome modulates gut–brain axis glycerophospholipid metabolism in a region-specific manner in a nonhuman primate model of depression. *Mol. Psychiatry* **26**, 2380–2392 (2021).
40. Bonnet, E., Calzone, L. & Michoel, T. Integrative Multi-omics Module Network Inference with Lemon-Tree. *PLOS Comput. Biol.* **11**, e1003983 (2015).
41. Erola, P., Björkegren, J. L. M. & Michoel, T. Model-based clustering of multi-tissue gene expression data. *Bioinformatics* **36**, 1807–1813 (2020).
42. Thiele, I. Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Mol. Syst. Biol.* **16**, e8982 (2020).
43. Magnúsdóttir, S. & Thiele, I. Modeling metabolism of the human gut microbiome. *Curr. Opin. Biotechnol.* **51**, 90–96 (2018).
44. Baldini, F. *et al.* The Microbiome Modeling Toolbox: from microbial interactions to personalized microbial communities. *Bioinformatics* **35**, 2332–2334 (2019).
45. Baldini, F. *et al.* Parkinson's disease-associated alterations of the gut microbiome predict disease-relevant changes in metabolic functions. *BMC Biol.* **18**, 62 (2020).
46. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

47. Lund, P. J. *et al.* Stable isotope tracing in vivo reveals a metabolic bridge linking the microbiota to host histone acetylation. *Cell Rep.* **41**, (2022).
48. Li, Y., Ge, X., Peng, F., Li, W. & Li, J. J. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biol.* **23**, 79 (2022).
49. Noronha, A. *et al.* The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Res.* **47**, D614–D624 (2019).
50. Chung, C. H., Lin, D.-W., Eames, A. & Chandrasekaran, S. Next-Generation Genome-Scale Metabolic Modeling through Integration of Regulatory Mechanisms. *Metabolites* **11**, 606 (2021).