



Sep 02, 2025

Investigating HARs in Neurodevelopment via Host–Microbe–Metabolite Interactions: A Computational Protocol

DOI

<https://dx.doi.org/10.17504/protocols.io.ewov11dmkvr2/v1>

Siddharth Singh¹

¹Indian Institute of Technology Indore



Siddharth Singh

Indian Institute of Technology Indore

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.ewov11dmkvr2/v1>

Protocol Citation: Siddharth Singh 2025. Investigating HARs in Neurodevelopment via Host–Microbe–Metabolite Interactions: A Computational Protocol. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.ewov11dmkvr2/v1>

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: In development

We are still developing and optimizing this protocol



Created: September 01, 2025

Last Modified: September 02, 2025

Protocol Integer ID: 226169

Keywords: metabolomic pathway enrichment, gut microbiome, composition of the gut microbiome, hars in neurodevelopment, microbial data mining, human genome, evolved genomic loci, neurodevelopmental processes through host, metabolite interaction, including brain development, brain development, genomic loci, network biology, genomic analysis, evolutionary context mapping, neurodevelopment, neurodevelopmental process, associated genetic factor, commensal microbe, biology

Abstract

Human Accelerated Regions (HARs) are rapidly evolved genomic loci thought to contribute to human-specific traits, including brain development. Recent studies suggest that some HAR-associated genetic factors may influence the composition of the gut microbiome, hinting at co-evolution between human genomes and commensal microbes. The following protocol outlines a **multi-step computational workflow** to explore how HARs might impact neurodevelopmental processes through host–microbe–metabolite interactions. This integrative approach combines genomic analyses, microbial data mining, metabolomic pathway enrichment, pharmacokinetic predictions, network biology, and evolutionary context mapping.

Overview

- 1 Human Accelerated Regions (HARs) are rapidly evolved genomic loci thought to contribute to human-specific traits, including brain development. Recent studies suggest that some HAR-associated genetic factors may influence the composition of the gut microbiome, hinting at co-evolution between human genomes and commensal microbes. The following protocol outlines a **multi-step computational workflow** to explore how HARs might impact neurodevelopmental processes through host-microbe-metabolite interactions. This integrative approach combines genomic analyses, microbial data mining, metabolomic pathway enrichment, pharmacokinetic predictions, network biology, and evolutionary context mapping.
- 2 Before starting, ensure you have the following prerequisites and data prepared:
 - **HAR list and coordinates:** Obtain a list of HAR loci with genomic coordinates (e.g., from Doan *et al.* 2016 or other published HAR datasets). This will serve as the starting point for identifying HARs that may relate to the microbiome.
 - **Computational tools access:** The protocol relies on various public databases and web tools (listed in **Materials**). Ensure you have a stable internet connection and any required accounts or downloads for these resources. Most tools are web-based; some analyses (like genomic coordinate overlaps or custom network visualization) may require basic programming or the use of software like **bedtools** (for genomic overlaps) or **Cytoscape** (for network visualization).
- 3 By completing this protocol, you will identify candidate **microbiota-associated HAR loci**, the gut microbes linked to them, metabolites produced by those microbes, affected host pathways, and regulatory networks connecting these factors to neurodevelopment. Each step below includes the **biological rationale** for the analysis and references to relevant tools or databases.

Materials

- 4 **HAR Genomic Data:**
 - *HAR*
coordinates: Genomic locations of HARs from literature (e.g., Doan *et al.* 2016).
 - *Gene*
lists: Genes located near or within HAR loci (for downstream network analysis).
- 5 **Databases and Web Tools:**
 - **GIMICA (Gut Immune-Microbiota Genetic Catalog)** – Database of host genetic factors (SNPs, CNVs, non-coding RNAs) that shape human gut microbiota composition.
 - **NCBI Taxonomy** – Resource for taxonomic classification of microorganisms (to retrieve phylum, family, etc. for microbial species).

- **gutMGene**– Database of gut microbiome gene and metabolite targets; links gut microbes, the metabolites they produce, and host target genes.
- **MetOrigin 2.0** – Tool/database to determine the origin of metabolites (microbial vs host or co-metabolized).
- **MetaboAnalyst 5.0 /6.0** – Online suite for metabolomic data analysis; used here for pathway enrichment analysis of metabolite lists.
- **PubChem or HMDB** – Chemical databases to obtain metabolite identifiers or SMILES structures (for input into other tools).
- **ADMETlab 2.0/3.0** – Web server for ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) property prediction of small molecules.
- **Deep-BBB (Deep-B3)/other** – Deep learning-based web tool for predicting blood–brain barrier permeability of compounds.
- **enviPath**– Platform for predicting microbial biotransformation pathways of compounds (Eawag-BBD pathway prediction).
- **gutMDisorder 2.0** – Database of gut microbe–disease associations (e.g., links between specific microbes and human disorders, including neurological conditions).
- **NetworkAnalyst 3.0** – Web platform for network-based visual analytics of genes/proteins, including tissue-specific networks and gene regulatory networks.
- **DifferentialNet** – An atlas of differential protein–protein interactions (PPIs) across human tissues (integrated in NetworkAnalyst as a data source for tissue-specific PPIs).
- **iNetModels (TCSBN database)** – Tissue and Cancer Specific Biological Networks database providing gene co-expression networks for specific tissues (accessible via NetworkAnalyst).
- **STRING database** – Comprehensive database of known and predicted protein–protein interactions; used here via NetworkAnalyst for PPI confidence scoring.
- **footprintDB** – Database of transcription factor binding motifs and their associated transcription factors.
- **JASPAR 2022** – Database of curated transcription factor binding profiles (motifs), used in identifying regulatory networks in NetworkAnalyst.
- **TFCheckpoint** – Cross-referencing database for transcription factors across human, mouse, and rat, useful for checking conservation of TFs.
- **GRAND** – Database of gene regulatory network models with integrated GWAS(genome-wide association) data for various human conditions.
- **miRDB** – Database for predicted microRNA target genes (Human miRDB v2020).
- **DGIdb 4.0** – Drug–Gene Interaction Database for known druggable genes and drug compounds.
- **PathBank 2.0** – Pathway database that includes metabolic and disease pathways across model organisms (for cross-species pathway analysis).
- **HervD Atlas** – Database of human endogenous retrovirus (HERV) integrations with disease and gene association data (with a Knowledge Graph interface).
- **Cytoscape (version 3.10)** – Software for network visualization, used to construct and visualize networks (e.g., microbe–metabolite–gene networks, HERV networks).



- 6 Ensure you have access to all the above resources. It may be helpful to create an organized folder or spreadsheet to compile results from each step (lists of loci, microbes, metabolites, etc., along with their annotations and analysis results).

Software & web servers (access links)

7

Resource	Purpose	URL
GIMICA	Host genetic & immune factors shaping human microbiota	https://gimica.idrblab.net/ttd/
gutMGene	Microbe–metabolite–host gene links	https://bio-computing.hrbmu.edu.cn/gutmgene/
MetOrigin 2.0	Metabolite provenance (microbial vs co-metabolized)	https://metorigin.met-bioinformatics.cn/
MetaboAnalyst 5.x/6	Pathway enrichment/topology for metabolites	https://www.metaboanalyst.ca/
ADMETlab 2.0/3.0	Drug-likeness & ADMET prediction	https://admetmesh.scbdd.com/
Deep-B3	BBB permeability classifier	https://cbcb.cdutcm.edu.cn/deepb3/
enviPath	Microbial biotransformation prediction	https://envipath.org/
gutMDisorder v2.0	Microbe–disease associations	https://bio-computing.hrbmu.edu.cn/gutMDisorder/
NetworkAnalyst	Tissue-aware PPI/GCN layers & enrichment	https://www.networkanalyst.ca/
STRING	High-confidence PPIs	https://string-db.org/

	footprintDB	Motif/TF–DNA interfaces, motif similarity	https://floresta.eead.csic.es/footprintdb
	MEME Suite GOMo	Motif→GO term enrichment	https://meme-suite.org/meme/tools/gomo
	TFCheckpoint	Cross-refs for TFs (human/mouse/rat)	https://www.tfcheckpoint.org/
	GRAND	Context-specific GRNs + GWAS overlays	https://grand.networkmedicine.org/
	miRDB	miRNA target predictions	https://mirdb.org/
	DGIdb (≥4.x/5.x)	Drug–gene interactions	https://dgidb.org/
	PathBank 2.0	Cross-species pathway maps	https://pathbank.org/
	HervD Atlas (Knowledge Graph)	HERV–disease–gene links	https://ngdc.cncb.ac.cn/hervd/graph
	Cytoscape ≥3.10	Network visualization	https://cytoscape.org/download.html

Step-by-Step Procedure

8 1. Identify Microbiota-Associated HAR Loci

Rationale: The first step is to find HAR loci that might be functionally linked to the gut microbiome, hypothesizing that human genomic adaptations (HARs) could have co-evolved with the microbiome. By identifying HARs overlapping genetic variants known to affect microbiota composition, we pinpoint candidates for host–microbe co-evolution.

- 8.1 **Procedure:** Start with your list of HAR genomic coordinates. Using the **GIMICA** database, retrieve the catalog of host genetic variants (e.g., SNPs, copy number variations, non-coding RNAs) that are reported to significantly influence the abundance of gut microbial taxa. GIMICA provides a comprehensive list of such host genetic factors associated with

microbiota variation. Filter or query within GIMICA for variants linked to specific gut microbial species or families of interest (if any specific microbes are considered) or download the entire set of microbiota-associated variants. Next, perform a genomic overlap analysis between these variant coordinates and HAR coordinates. This can be done by converting both datasets into a common format (e.g., BED files) and using a tool like **bedtools intersect** or an equivalent script to find overlaps or near overlaps. Consider a variant as HAR-associated if it falls *within* a HAR or in close proximity (e.g., within 5–50 kb, depending on linkage disequilibrium considerations). Record all HAR loci that have one or more microbiota-associated genetic variants overlapping or nearby; these represent “**microbiota-associated HAR loci**”. For each such HAR locus, note the corresponding microbial taxa implicated (from the GIMICA entry) and any nearby genes that the HAR might regulate or reside in. This yields a list of candidate HARs potentially linking human genomic adaptation to microbiome composition.

- 8.2 **Expected output of this step:** A table or list of HAR loci that coincide with microbiome-associated variants, including HAR coordinates, the associated microbe(s) from GIMICA, and any nearby human genes. These candidates will be used for downstream analysis (coevolutionary analysis and network construction).

9 **2. Classify and Annotate Associated Microbial Species**

Rationale: After identifying microbes linked to HAR loci, it's important to understand these microorganisms' characteristics. Knowing their taxonomy and whether they are commensals, pathogens, or probiotics provides insight into **how** they might influence the host (e.g., pathogenic bacteria might drive immune-related HAR selection, whereas commensals might be involved in mutualistic co-evolution).

- 9.1 **Procedure:** Compile the list of microbial species obtained from Step 1 (via GIMICA or related sources). For each microbe, use **NCBI Taxonomy** (e.g., via the NCBI Taxonomy browser or E-utilities) to retrieve its taxonomic classification. Record the phylum, family, genus, and species for each organism. This can often be done by searching the species name on NCBI Taxonomy and noting the lineage. Next, gather information on the microbe's relationship with the host:
- Determine if the microbe is typically a **commensal (beneficial or neutral symbiont)**, a **pathogen (disease-causing)**, or a **probiotic (known to confer health benefits)**. This information may be found in the literature or databases like gutMGene or **gutMDisorder**, or by searching microbiome review articles for that species.
 - Note any known colonization sites and roles (e.g., a gut microbe influencing metabolism or an immune-modulating microbe).
- 9.2 **Expected output of this step:** An annotated list of microbial species, including their taxonomy (phylum, family) and physiological role (commensal/pathogen/probiotic). This ensures that downstream analyses focus on relevant human-associated microbes and

provides context (e.g., if many HAR-associated microbes are pathogens, it might indicate evolutionary pressure from infectious disease).

10 **3. Curate Microbial Metabolites and Perform Pathway Enrichment**

Rationale: Gut microbes can influence the host by producing metabolites that enter the host's circulation and affect physiology, including brain function. This step compiles metabolites produced or modified by the HAR-associated microbes and identifies which host metabolic pathways they impact. Enriching pathways highlights biological processes that might be under selective pressure due to microbe-derived metabolites

10.1 Procedure: Using your list of microbial species from Step 2, gather known metabolites associated with each microbe:

- Use **gutMGene** to find microbe–metabolite associations. gutMGene is a comprehensive database linking gut microbes to the metabolites they produce and any known host target genes of those metabolites. For each microbial species, query gutMGene for entries where the microbe is listed; extract the metabolites and any host gene targets involved. Focus on metabolites that have known interactions with host genes or pathways (gutMGene often provides a triplet: microbe, metabolite, host gene).
- Supplement this by consulting **MetOrigin 2.0**. MetOrigin can help determine the origin of a given metabolite – whether it is exclusively produced by microbiota, co-metabolized by both host and microbe, or solely of human origin. For each candidate metabolite from gutMGene, use MetOrigin (via its web interface or data) to confirm:
 1. Is the metabolite produced by gut bacteria? by human enzymes? or by both (co-metabolite)?
 2. Which bacterial taxa are known producers of this metabolite? Mark each metabolite accordingly (microbial, host, or co-metabolite) and remove any that are not relevant (e.g., metabolites purely of human origin can be de-prioritized if they are not actually contributed by the microbiome).
- The result should be a curated list of metabolites associated with the HAR-linked microbes.

Next, perform **metabolic pathway enrichment analysis** on the curated metabolite list:

- Prepare the metabolite identifiers in a format compatible with **MetaboAnalyst 5.0/6.0**. It is recommended to use unique identifiers like HMDB IDs or PubChem IDs for each metabolite to ensure accurate mapping. If you have metabolite names, MetaboAnalyst can often recognize common names, but HMDB or KEGG IDs improve accuracy.
- Open the MetaboAnalyst Pathway Analysis module (via MetaboAnalyst.ca). Choose the **Pathway Analysis** tool for metabolites. Select the **Homo sapiens** pathway library (e.g., KEGG Human metabolic pathways) since we want to see which *human* pathways these microbiome-derived metabolites impact.
- Upload or input the list of metabolites (HMDB or KEGG IDs). Choose **Over Representation Analysis (ORA)** as the enrichment method (hypergeometric test by

default) and **Pathway Topology Analysis** with the metric "betweenness centrality" (which considers a metabolite's position in a pathway).

- Run the analysis. Identify significantly enriched pathways – those with a False Discovery Rate (FDR) < 0.05, as these indicate pathways that are over-represented in your metabolite list compared to what would be expected by chance.
- Examine the output charts and tables. MetaboAnalyst will provide a list of pathways with p-values, FDR, and an impact score for each (impact is a measure of pathway centrality). Save these results (they may be needed for figure generation or further interpretation).

10.2 Expected output of this step: A finalized list of microbe-associated metabolites (with annotations of their origin) and a set of significantly enriched host metabolic pathways (with names, p-values/FDR, impact scores). The enriched pathways highlight which areas of host metabolism may be influenced by the microbiome under the influence of HAR-associated microbes.

11 4. Predict BBB Permeability and Drug-Likeness of Microbial Metabolites

Rationale: To understand how microbial metabolites could affect neurodevelopment, we must assess which metabolites can reach the brain. The blood–brain barrier (BBB) restricts many compounds from entering the central nervous system. Additionally, "drug-likeness" (physicochemical properties influencing absorption, distribution, etc.) can indicate if a metabolite might act like a signaling molecule or therapeutic. In this step, we predict which microbial metabolites are likely to cross the BBB and have favorable ADMET properties. This helps prioritize candidates that could directly influence brain development or function.

11.1 Procedure: For each metabolite in your curated list:

- **Obtain SMILES structures:** Using databases like **PubChem** or **HMDB**, retrieve the SMILES string (a text representation of chemical structure) for each metabolite. SMILES are required input for many predictive tools. Ensure the structures correspond to the correct metabolite (pay attention to isomers or specific forms if relevant).
- **ADMET property prediction:** Access **ADMETlab 2.0/3.0** (an integrated online platform for ADMET profiling). For each metabolite, input the SMILES string into the ADMETlab web server and run the prediction. Focus on the following output parameters:
 1. *Caco-2 permeability:* A prediction of intestinal epithelial permeability (helps gauge oral absorption potential).
 2. *P-glycoprotein (P-gp) substrate/inhibitor status:* Since P-gp can pump compounds out of the brain, a metabolite that is a P-gp substrate might have limited BBB penetration.
 3. *hERG liability:* Prediction of potential cardiotoxicity (hERG channel inhibition) – included as a safety/toxicity check.
 4. Other drug-likeness metrics, such as Lipinski's Rule-of-5 compliance, TPSA (topological polar surface area), etc., give a sense of overall bioavailability.

Record these properties for each metabolite. Particularly, identify metabolites with favorable permeability and no glaring toxicity flags, as these are more likely to reach and affect the brain.

- **BBB permeability prediction:** Next, use **Deep-BBB (Deep-B3)** or any other dedicated BBB permeability predictor based on deep learning models. Navigate to the Deep-B3 web interface. Input the SMILES structure for each metabolite and run the BBB classification. Deep-B3 will output a prediction such as “*BBB permeable*” or “*non-permeable*” for each compound, sometimes with probability scores or related descriptors. Note each metabolite’s predicted BBB status.
- ADMET-PrInt is another tool for interpreting ADMET profiles; if integrated, it might indicate if a metabolite is likely transported by common drug transporters. Note if any metabolite is identified as a likely substrate or inhibitor of key transporters relevant to brain uptake.
- After gathering ADMETlab and Deep-B3 results, highlight the metabolites that consistently show high permeability (Caco-2 positive, BBB permeable classification) and lack major issues (e.g., not a P-gp substrate or no toxicity alerts). These are candidates that could **traverse the BBB** and directly impact the central nervous system.

11.2 **Expected output of this step:** A profile for each metabolite, including key ADMET properties and a BBB permeability prediction. Several metabolites will emerge as strong candidates for crossing into the brain; these should be flagged for further discussion.

12 5. Construct Host–Microbe Co-metabolism Maps

Rationale: This step integrates host and microbial metabolic information to map out how microbial metabolites interface with host biology. By building a network of microbes, the metabolites they produce, and the host genes/pathways those metabolites affect, we can visualize the direct biochemical links between the microbiome and the host (including potential points of influence on neurodevelopment). Additionally, predicting microbial biotransformation pathways (metabolic reactions microbes can perform) and checking disease associations of the microbes provides a deeper understanding of functional interactions and their clinical relevance. This comprehensive map will reveal *co-metabolism* relationships (where host and microbe metabolism intersect) and any known disease ties of these interactions.

12.1 **Procedure:** This step can be divided into three parts:

5.A. Microbe–Metabolite–Host Gene Network:

Using the data gathered:

- Compile a list of triplet interactions (Microbial species – Metabolite – Host target gene). This information largely comes from the **gutMGene** database queries in Step 3. gutMGene entries that you collected often directly link a microbe to a metabolite and a

host gene that the metabolite influences. Aggregate all such interactions for your set of microbes and metabolites.

- Create an edge list or table where each row is one interaction: e.g., Microbe → Metabolite, and Metabolite → Host Gene. You can represent this as a network:
 1. Nodes: microbes, metabolites, and host genes.
 2. Edges: a microbe-produces-metabolite edge, and a metabolite-affects-gene edge.
- **Visualize the network:** Use **Cytoscape** (or an alternative network visualization tool) to create a network graph. Import the edge list (you may need to create two sets of edges and then merge). In Cytoscape, you can style nodes by type (e.g., squares for microbes, circles for metabolites, triangles for host genes) and use different colors. This network illustrates the biochemical dialogue between gut microbes and the host.
- Analyze the network for simple metrics: identify any microbes or metabolites that connect to multiple genes (hubs) or if certain host genes are convergence points for multiple microbial metabolites. These could be critical interaction nodes.

12.2 5.B. Microbial Biotransformation Pathway Prediction

- For each metabolite (especially key ones identified in previous steps), predict possible transformations by gut microbes using the **enviPath** platform. enviPath provides a rule-based prediction of biotransformation reactions based on known microbial metabolism rules (the Eawag-BBD database of biochemical transformations).
- On the enviPath website (or via its API if you prefer scripting), input the chemical structure or identifier of the metabolite. Use the *Pathway Prediction* tool with default settings (which usually apply the **Eawag-BBD** rules). Limit the prediction to **one reaction step** to find primary transformation products (since multiple steps can explode the search space).
- Examine the predicted transformation products for each metabolite. Ask: Does the microbe potentially convert the metabolite into another compound that is part of a known human metabolic pathway? Compare predicted microbial products with human metabolic intermediates (you may consult databases like KEGG or HMDB for known human metabolites). The aim is to find **“metabolic complementarity”** – cases where a microbial reaction produces an intermediate that the human body can further utilize, or vice versa.
- Record any such complementary pathways. These indicate points of host–microbe metabolic intersection and might highlight how microbes contribute to biochemical processes relevant to the host (potentially including brain chemistry).

12.3 5.C. Disease Associations of Microbes:

- Using the **gutMDisorder** database (or literature searches), check each microbial species for known associations with human diseases, particularly neurological or developmental disorders. gutMDisorder is a resource documenting microbial dysbiosis in various diseases. You can search by microbe name to see if there's an entry linking it to any disease.
- If a microbe is found, note the associated disorder and any provided metrics (e.g., an LDA score from LEfSe analysis, or p-values, depending on how gutMDisorder

presents data). Focus on diseases of the nervous system or childhood developmental conditions, as these are most pertinent to neurodevelopment.

- This information adds context: if a HAR-associated microbe is known to differ in autism, Parkinson's, or other neurodevelopmental/neuropsychiatric conditions, it strengthens the case that HARs could be tied to those phenotypes via microbiome differences.

12.4 **Expected output of this step:** An integrated **host–microbe–metabolite interaction network** (as a figure or Cytoscape session) and accompanying data:

- A list of microbe–metabolite–gene interactions summarizing which gut bacteria produce which metabolite and what host gene/pathway is affected.
- A summary of **predicted microbial biotransformations** for key metabolites, highlighting potential shared pathways between microbe and host metabolism. You might note “Microbe X can convert metabolite A to B, which is a precursor in the human pathway Y (linking microbial metabolism to a human neurotransmitter pathway).”
- A table of **microbe–disease associations** listing any diseases (with emphasis on neurodevelopmental or neurological disorders) associated with each microbe, along with the evidence metric (like an LDA score or reference). This provides an evolutionary or medical relevance layer to your findings.

13 **6. Analyze**

HAR-Linked Host Gene Networks and Transcription Factor Motifs

Rationale: Many HARs are non-coding and may function by regulating gene expression. To understand the impact of HARs on host biology, particularly in the context of neurodevelopment, it is crucial to examine the gene regulatory networks in which HAR-associated genes operate. This step identifies **tissue-specific interaction networks** of the host genes gathered (from HAR loci and microbe–target interactions) and examines transcription factor (TF) regulatory motifs that might be enriched in HAR sequences. By doing so, we uncover key signaling pathways, TFs, and motifs under potential selective pressure and relevant to brain development.

13.1 **Procedure:** This step has two sub-parts: building the host gene interaction network and analyzing regulatory motifs/TFs.

6.A. Build Tissue-Specific Host Gene Interaction Networks:

- **Compile the host gene list:** Merge relevant gene lists from earlier steps. This may include genes located near HAR variants (from Step 1) and host genes targeted by microbial metabolites (from Step 5A). The union of these forms a set of candidate genes through which HARs and microbes might influence neurodevelopment.
- **NetworkAnalyst setup:** Go to **NetworkAnalyst** and start a new network analysis. When prompted, upload or input your gene list (use official human gene symbols; NetworkAnalyst will auto-map aliases, and it will report if any genes are not recognized).

- **Tissue-specific PPI network (DifferentialNet):** In NetworkAnalyst's network construction options, select the protein–protein interaction (PPI) analysis. Choose the **DifferentialNet** option as the source for PPIs, which allows you to build a network specific to a tissue of interest. If neurodevelopment is the focus, you might select a brain-related tissue (if available in DifferentialNet, perhaps “brain” or specific CNS regions). DifferentialNet is a curated atlas of PPIs that differ across tissues; selecting a tissue will include edges (protein interactions) present or significantly changed in that tissue. Upload your gene list as “seed” or “query” genes for the PPI network.
- **Tissue-specific co-expression network (iNetModels TCSBN):** To incorporate transcriptional co-expression relationships, add another network layer. NetworkAnalyst allows integration of co-expression networks via the **TCSBN (Tissue and Cancer Specific Biological Networks)** data within **iNetModels**. From within NetworkAnalyst, choose to overlay a gene co-expression network corresponding to the same or a relevant tissue. Import a **tissue-specific gene co-expression network (GCN)** for a brain region from iNetModels. This will add edges between genes that are co-expressed (with a certain correlation threshold) in that tissue.
- **Integrate and refine the network:** Now you have a combined network of PPIs and co-expression links for your gene list, specific to the chosen tissue. Apply filters:
 1. Retain only high-confidence edges: e.g., for PPIs, you might set a confidence cutoff such as a STRING score > 0.7 (if NetworkAnalyst uses STRING scoring internally). For co-expression, filter for strong correlations (absolute Pearson $r \geq 0.6$) that are statistically significant (adjusted p -value < 0.05).
 2. Use the “Minimum Network” function (if available), which keeps all seed genes and connects them with the shortest possible paths. This prunes extraneous nodes while ensuring the network remains connected and all input genes are present.
- **Network analysis:** Once refined, examine the network:
 1. Identify hub genes by computing centrality measures like degree (number of connections) and betweenness (importance in connecting network paths). NetworkAnalyst can compute these statistics for the nodes.
 2. Detect modules or communities in the network. One approach is edge betweenness clustering or other community detection algorithms provided. This may reveal subnetworks (modules) of tightly connected genes that could represent specific pathways or complexes.
- Note which of your original genes are hubs or in key positions. Also, observe if any **neurodevelopment-related genes** (e.g., transcription factors, neural receptors) appear prominently; these could be crucial mediators of HAR effects on neurodevelopment.
- If NetworkAnalyst provides pathway analysis on the network. You may perform enrichment on the gene nodes to see which GO terms or pathways are overrepresented (though note, the gene list is already focused, so this is optional).
- (Optional) **Visualization:** You can view the network in NetworkAnalyst's interface or export it. For high-quality figures or further customization, exporting to **Cytoscape** is

an option. In Cytoscape, one could overlay additional information (like highlight HAR-associated genes vs. microbe-targeted genes with different colors, or node shapes).

13.2 6.B. HAR Motif and TF Analysis:

- **Motif discovery in HAR sequences:** If you have the DNA sequences of HAR loci (from a FASTA or extracted via genome coordinates), you might analyze them for transcription factor binding motifs, OR use the candidate TFs from the network analysis and use *footprintDB* for motif characterization. To replicate this concept:
 1. Identify candidate transcription factors of interest. One way is to look at your network, which TF genes (if any) are present, or which genes are known to be transcriptional regulators. Another way is to use multivariate analysis (e.g., PLS-DA – Partial Least Squares Discriminant Analysis) if you have some grouping of data. For simplicity, suppose we take the top ~20–25 TFs either present in the network or implicated by HARs.
 2. For each such transcription factor, use **footprintDB** to examine known DNA binding motifs. FootprintDB allows you to search by TF or by motif. It provides annotated **cis-regulatory elements and motifs** for many transcription factors. Retrieve the DNA sequence motifs associated with your TFs of interest. FootprintDB entries will list motif patterns, the transcription factors that bind them, and often similarity scores (e.g., STAMP e-value indicating how well a motif matches a known consensus).
 3. If you have HAR DNA sequences, you could scan them for these motifs (using tools like FIMO from the MEME Suite or other motif scanning programs). OR take known motifs and perform the functional annotation.
 4. Utilize the **Gene Ontology for Motifs (GOMo)** analysis, which is part of the MEME Suite or related to footprintDB's output. GOMo will link motifs to GO terms based on the genes that the TFs regulate. GOMo analysis connected motifs to biological processes, with statistical significance. Run a GOMo or similar enrichment for the motifs you identified to see if any neural development or neurophysiology-related GO terms are enriched. Record the GO terms and significance (q-values).
 5. The result should be a list of motifs (and their associated TFs) that appear relevant to your HAR-associated genes, along with potential functions. If you find that motifs for neural crest development factors are enriched, it suggests HARs could be affecting the binding sites of those TFs.
- **Cross-species TF comparison:** Take the top/all transcription factors you've pinpointed as important (from network hubs or motif analysis, e.g., 25 TFs). Use **TFCheckpoint** to verify their identities across species. TFCheckpoint is a database that cross-references TFs in human, mouse, and rat by gene symbol and other identifiers. This helps ensure that if you later compare animal model data (mouse/rat) with human, you have the correct TF orthologs. It also provides an update on TF nomenclature and synonyms, which can be useful if any gene symbols are outdated.
- **Trait/Disease relevance of TFs:** For each key TF, consult **GRAND** (Gene Regulatory Network Database) to see if they are implicated in any neurodevelopmental or neurological conditions via GWAS or regulatory network models. In GRAND, you can

search by gene or browse trait-associated networks. Alternatively, search each TF to see what traits it's linked to (GRAND integrates GWAS hits; it might tell you if a TF gene had SNPs associated with certain traits).

- Summarize how many of the top TFs intersect with neurological trait data and note any specific strong links (e.g., "TF X is linked to schizophrenia GWAS hits" or "TF Y appears in a regulatory network for neural progenitor cells"). This adds evidence that the TFs (hence the HARs influencing them) are relevant to neurodevelopment or disease.

13.3 **Expected output of this step:** A multi-layer view of HAR-associated host gene regulation:

- A **tissue-specific network** of HAR-linked genes, highlighting key hub genes and interactions.
- A list of **top transcription factors and motifs** associated with the HAR context. One might report "Top 25 motifs enriched in HAR sequences and their corresponding TFs," noting which TFs have roles in neurodevelopment or immunity (these were identified via footprintDB and GOMo).
- Cross-species TF reference table, listing those TFs in human/mouse/rat.
- A **TF-trait association matrix** or list indicating which of the key TFs overlap with GWAS traits and how strongly. Notably, if many HAR-associated TFs map to neurodevelopmental or psychiatric trait loci, that bolsters the link between HARs, gene regulation, and neurodevelopment.

14 **7. Integrate miRNA and Drug Target Interactions**

Rationale: To fully capture the regulatory network around HAR-associated genes, we must include post-transcriptional regulation by microRNAs (miRNAs) and potential pharmacological interactions. miRNAs can fine-tune gene expression and may themselves be influenced by microbial metabolites or HAR-related pathways. Meanwhile, identifying drugs targeting the key genes provides insight into whether these pathways are modifiable or already implicated in therapies. This step adds a layer of **miRNA-gene and drug-gene interactions** to the network, rounding out the host side of the host-microbe interaction landscape.

14.1 **Procedure:**

- **Identify candidate miRNAs:** Using your list of key host genes (the genes influenced by HARs/microbes), query **miRDB** (Human miRDB, latest version, e.g., 2020) for miRNAs that target these genes. You can do this by entering each gene into miRDB's target search to see which miRNAs are predicted to target it, or by entering each miRNA to see its gene targets. A practical approach:
 1. For each gene of interest, note the top predicted miRNA regulators with high confidence (miRDB provides a score 0–100; use a threshold such as **score $\geq$ 80** for high confidence).
 2. Alternatively, identify a set of top miRNAs that repeatedly appear targeting multiple genes in your list.

3. Gather the set of unique miRNAs that meet the criteria. For each selected miRNA, you may retrieve all its predicted target genes (to validate that your genes are included and to see what other genes it might co-regulate).

■ **Identify drug–gene interactions:** For the same set of host genes, use the **Drug–Gene Interaction Database (DGIdb 4.0)**. On the DGIdb web interface:

1. Input each gene symbol to find known drugs that interact with the gene or its protein product. Filter the results to focus on small-molecule drugs and well-characterized interactions (ignore broad categories like “inhibitor (predicted)” or nutraceuticals unless relevant). Look for drugs labeled as **inhibitors, agonists, antagonists, modulators** of the gene’s protein. DGIdb curates data from various sources, including FDA-approved drugs, clinical candidates, etc.

2. Record the drug name, interaction type, and perhaps the source database or reference if provided. If a gene has no known drug interactions in DGIdb, mark it as “no known drug” for now.

■ **Construct the miRNA–Gene–Drug network:** Now integrate these relationships, either conceptually or in a visualization:

1. You can expand the network from Step 6A by adding miRNA nodes and drug nodes. A miRNA node connects to a gene node if that miRNA targets the gene (edge could be directed from miRNA to gene to indicate regulation). A drug node connects to a gene node if the drug modulates the gene’s protein (edge labeled as inhibition, etc.).

2. If using Cytoscape, import these as additional edges: e.g., create a list of miRNA→gene interactions and drug–gene interactions and merge with the existing network. Use different shapes/colors for miRNAs (perhaps diamond shapes) and drugs (e.g., hexagons), to distinguish them from genes.

■ **Analysis:** Identify any interesting patterns:

1. Are multiple key host genes regulated by the same miRNA? This could indicate a miRNA that orchestrates a part of this network (and might be responsive to microbial metabolites, according to some studies).

2. Are there drugs that target multiple of the key genes? Such drugs (or combinations) could theoretically modulate the host-microbe interaction outcomes and might be of therapeutic interest.

3. Notably, if some of the identified drugs are neurologically active or used in neurodevelopmental disorders, that underscores the relevance of these pathways to brain function.

14.2 **Expected output of this step:** An expanded network that includes:

■ A list of top miRNAs and their target genes among the HAR/microbe-linked gene set (where each miRNA has genes with a miRDB score ≥ 80).

■ A list of drug–gene interactions, specifying which drugs target which of the key genes. You might find that several of the genes are kinase receptors with existing inhibitors.

■ The integrated miRNA–gene–drug network highlights additional regulatory control points and potential intervention points. These results provide a more complete

picture of how one might modulate the system experimentally or therapeutically – e.g., if a particular miRNA is critical, maybe it could be a biomarker or therapeutic target, and if certain drugs hit these pathways, they might influence the host–microbe interaction outcomes.

15 8. Cross-Species

Pathway Conservation Analysis

Rationale: To place the findings in an evolutionary context, we examine whether the metabolic pathways identified (from Step 3's enrichment) and the microbe-produced metabolites are conserved across different organisms. If a pathway is present in both humans and gut bacteria (and perhaps other animals), it suggests evolutionary conservation and possibly that these host–microbe metabolic interactions have ancient origins. This addresses the concept of *phylosymbiosis* (parallel evolution of hosts and their microbiomes) and highlights which metabolic capabilities are shared or co-evolved. Essentially, we're asking: do other species (especially model organisms) have analogous pathways for these metabolites, and is the host–microbe interplay a common theme across evolution?

15.1 Procedure:

- Take each metabolite from your curated list (especially those implicated in significant pathways or of high interest from prior steps). Using the **PathBank 2.0** database, find all metabolic pathways that include this metabolite. PathBank allows searching by metabolite name or ID and returns a list of pathways that the metabolite participates in.
- For each pathway returned, note the pathway name and the list of organisms that PathBank has entries for that pathway. PathBank covers various model organisms (human, mouse, rat, fruit fly, yeast, *E. coli*, etc.) and even some pathogen pathways. Each pathway entry will list the species in which that pathway (or a very similar pathway) exists, often by taxonomic category.
- Tabulate the presence/absence of each pathway across organisms. If metabolite X is involved in a pathway that appears in humans, mice, and bacteria, that's a **conserved pathway**. If a pathway is only in human and mouse but not in bacteria, it might be a host-specific pathway, whereas if it's in bacteria and human but not in mouse, that might indicate a pathway humans share with microbiota but not all mammals (could hint at diet or environment differences).
- Mark pathways that are present in **≥ 2 different taxonomic clades** (e.g., in both mammals and bacteria, or in mammals and plants, etc.) as evidence of conservation or convergent evolution. Pay special attention to pathways that include both a gut microbe and the human host (some might explicitly be host pathways that microbes contribute to, like certain vitamin biosynthesis).
- Also note any pathways unique to microbes that nonetheless affect the host (these wouldn't show up in human PathBank but are important microbe pathways for producing the metabolites; you might gather these from literature if needed).

- The results can be summarized as a comparative table that lists each metabolite or pathway, and indicates which organisms have it. Additionally, highlight if a pathway's presence correlates with the host's relationship with those microbes, pathways present in gut bacteria, and also in the human might be points of host-microbe synergy.

15.2 **Expected output of this step:** An **evolutionary conservation map** of pathways:

- A table or chart showing each metabolite's associated pathways and which organisms share those pathways. One might report, "Metabolite A is in the tryptophan metabolism pathway, which exists in humans, *E. coli*, and yeast, indicating conserved tryptophan utilization across domains." Meanwhile, another metabolite's pathway might be only in bacteria and not in humans, suggesting the metabolite is purely microbial but might mimic a host signaling molecule.
- An interpretation of the results: e.g., "We observed that several key pathways (X, Y, Z) involving our metabolites are conserved in both gut bacteria and mammals, consistent with co-evolutionary retention of these metabolic circuits. This supports the idea that host neurodevelopmental processes influenced by these pathways could be a product of long-term host-microbe coadaptation."
- These findings connect back to the concept of HARs: if HARs influenced those metabolic interactions, perhaps they did so in response to pressures that were present throughout evolution (e.g., a nutrient metabolism pathway critical for brain development that needed to integrate microbial contributions).

16 **9. Overlay Endogenous Retrovirus (ERV) Imprints on the Network**

Rationale: The final step incorporates an additional evolutionary layer: human endogenous retroviruses (HERVs). HERVs are remnants of ancient viral infections in the human genome that can influence gene regulation and have been implicated in various diseases (including neurological disorders). Investigating HERV–gene–disease associations provides insight into whether viral insertion events may intersect with our HAR-related network. This is somewhat tangential but adds to the evolutionary narrative: HARs, microbiome, and HERVs all represent forces in human evolution that could converge on neurodevelopmental gene networks.

16.1 **Procedure:**

- Use the **HervD Atlas** database, which catalogues HERV elements and their associations with diseases and genes. Access the **Knowledge Graph → Disease Network** interface on the HERVd Atlas website. This tool allows you to query and visualize connections between HERV families, specific HERV insertions, human genes, and diseases.
- Query the database for broad connections: Include all disease categories and all HERV families to get a comprehensive network (if the interface allows a global query). Alternatively, filter for diseases of interest (e.g., neurological disorders, developmental disorders) to focus on those.

- The HERVd Atlas will typically return data (possibly as a JSON or tabular output) that includes triplets or pairs like HERV element – gene – disease, along with some measure of evidence (often a literature-based score or count of supporting studies).
- Download or save the resulting network data. If provided in JSON, you might convert it to an edge list for easier handling. Standardize the nomenclature:
 1. Ensure disease names are mapped to a consistent ontology, EFO (Experimental Factor Ontology) and DOID (Disease Ontology) terms, to standardize disease names, which is good practice if you plan to integrate with other datasets.
 2. Ensure gene identifiers are official (map to HGNC symbols).
 3. HERV entities might be listed as families (e.g., HERV-W, HERV-K) or specific loci; keep them as given, but be aware of hierarchy (class vs specific element).
- **Visualize the HERV network:** Using **Cytoscape 3.10** (or the HERVd Atlas's own graph viewer), create a network where:
 1. Nodes represent **Diseases, HERV families/elements, and Human genes.**
 2. Edges represent documented associations (e.g., an edge between a HERV and a disease means that HERV activity is linked to that disease; an edge between a HERV and a gene suggests the HERV is near or affects that gene; an edge between gene and disease indicates gene is implicated in that disease, possibly via HERV or other mechanisms).
- Analyze the network for any overlaps with your HAR/microbiome gene list: Are any of the genes identified in earlier steps (HAR-associated or microbe-targeted genes) present in the HERV network? If yes, that could mean those genes are also influenced by viral insertions or are part of viral response pathways. Also, check if neurological diseases appear and which HERVs or genes connect to them.
- The presence of overlapping elements would suggest a complex interplay: e.g., a HAR might be near a gene that is regulated by a HERV, and that gene is involved in brain development – indicating multiple layers of evolutionary influence.

16.2 Expected output of this step: A HERV–Disease–Gene network showing how ancient viral elements intersect with modern human diseases and genes. Specifically:

- Lists of HERV classes/elements associated with neurological or developmental diseases, and the human genes through which they may exert influence (Table containing HERV–gene–disease associations).
- Observations on whether any HAR-linked genes from our analysis are also part of this HERV network. If a particular transcription factor gene is both near a HAR and has an embedded HERV element, that might be notable.
- This provides an additional evolutionary perspective: while not directly about the microbiome, it underscores how multiple genomic factors (HARs and HERVs) could be layering effects on key regulatory genes in neurodevelopment. It helps paint a more holistic picture of human evolutionary genomics in the context of brain development and environmental interactions.

Scientific Rationale

17 Notes on scientific rationale

- **GIMICA** uniquely aggregates host genetic & immune factors affecting microbiota, enabling the HAR-centric intersection in Module A.
- **gutMGene** and **MetOrigin 2.0** together separate microbe-derived compounds from co-metabolized compounds, strengthening causal interpretation of metabolite signals (Modules B–C/E).
- **MetaboAnalyst** provides well-established metabolite pathway enrichment/topology analysis.
- **ADMETlab** and **Deep-B3** complement each other for CNS relevance triage (pharmacokinetics/toxicity + BBB).
- **NetworkAnalyst** layers (DifferentialNet, STRING, iNetModels/TCSBN) offer tissue specificity critical for interpreting HAR-adjacent gene regulation.
- **HervD Atlas** adds an evolutionary/viral imprints dimension that can cohere with HAR-linked regulatory signals.

Expected Results

18 Expected Results

By following this protocol, you will assemble an integrative view of how HARs might influence neurodevelopment via host–microbe interactions. The expected results include:

- **HAR–Microbiome Associations:** A set of candidate HAR loci that overlap with genetic variants known to affect gut microbiota (microbiota-associated HARs). These loci come with a list of associated gut microbes, suggesting specific human genetic adaptations potentially tied to those microbes.
- **Microbial Taxonomy & Traits:** An annotated catalog of the gut microbes linked to those HARs, including their taxonomy (phylum, family) and whether they are commensals, pathogens, or probiotics. We expect many will be common gut commensals, but any pathogenic microbes present could imply past selective pressures (e.g., HARs that protect against certain infections).
- **Microbe-Derived Metabolites:** A curated list of metabolites produced or modified by these microbes, with notes on origin (microbial vs human vs co-metabolite). Anticipate a diverse array, including short-chain fatty acids, amino acid derivatives, and other small molecules. Each metabolite is mapped to whether gut bacteria produce it, the host can produce it, or both – highlighting potential unique microbial contributions to host biochemistry.
- **Pathway Enrichment:** A set of significantly enriched human metabolic pathways influenced by the microbial metabolites (FDR < 0.05). Likely outcomes might include pathways related to neurotransmitter synthesis, vitamin metabolism, or immune signaling. Each pathway comes with an impact score, suggesting which pathways are most central to the metabolite list.
- **BBB-Permeable Candidates:** Among the microbial metabolites, a subset will be predicted to cross the blood–brain barrier and possess drug-like properties. These

are prime candidates for mediating gut-brain communication. The expected result is a short list of compounds that are classified as BBB-permeable and that have favorable ADMET profiles (Caco-2 permeable, not P-gp substrates, etc.). These might include known neuroactive molecules (microbial metabolites reported to affect the brain) and possibly less characterized ones.

- **Host–Microbe Interaction Network:** A visual and analytical map of connections between gut microbes, the metabolites they produce, and host target genes. We expect to see a network where certain microbes (nodes) connect via metabolite links to clusters of host genes. This network illustrates the multi-dimensional interaction: how genetic factors (HARs influencing microbes) cascade to metabolic effects and then to gene regulation in the host.
- **Metabolic Complementarity:** Instances where microbial metabolism complements human metabolism. The enviPath predictions will reveal examples like a bacterium transforming compound A to B, and humans using B in a key pathway. These results underscore biochemical integration – e.g., a microbe might complete a step in a neurotransmitter precursor pathway that the human genome cannot, effectively extending the host's metabolic capabilities. Such findings align with theories that host and microbiota form a **“super-organism” metabolic network**.
- **Microbe–Disease Links:** Documentation of any known associations between the involved microbes and diseases, especially neurological or developmental disorders. If several HAR-linked microbes are reported to differ in autism spectrum disorder or in cognitive development measures, those results would be highlighted. We expect to see at least a few connections given growing literature on the gut–brain axis (e.g., certain gut bacteria reduced or enriched in autism or anxiety). These associations (from gutMDisorder or other sources) add clinical relevance to the findings and may point to specific microbe-metabolite-host pathways as therapeutic targets.
- **Tissue-Specific Gene Network Features:** A constructed network of HAR-related host genes (including those near HARs and those targeted by metabolites) in a relevant tissue context (likely brain). We anticipate identifying key **hub genes**, a developmental transcription factor, or a signaling receptor that has many connections (high degree) in the protein interaction/co-expression network. Community detection might reveal **modules** corresponding to biological processes (e.g., a module of immune genes vs. a module of neurodevelopmental genes), indicating that HARs and microbes might influence multiple facets of physiology. The network reduction ensures we focus on direct and relevant interactions, likely surfacing well-known pathways linking our gene set.
- **Enriched TF Motifs and Key Regulators:** Analysis of HAR sequences and network genes will yield a set of candidate transcription factors and their DNA motifs that are significant in this context. We expect to find certain motifs overrepresented – possibly those of TFs involved in brain development or immune response, reflecting HARs' roles.. The top 25 or so motifs/TFs (from PLS-DA or enrichment ranking) would be listed, giving insight into which regulatory factors HARs might be tweaking. These TFs, once cross-referenced in TFCheckpoint, will show if they are conserved in model

organisms. Moreover, through GRAND, some of these TFs will be linked to neurological traits, suggesting a possible mechanism by which variation in that TF's regulation (potentially due to HARs or HERVs) affects disease risk.

- **miRNA and Drug Network Insights:** The integrated miRNA–gene–drug network will reveal additional regulatory highlights. We anticipate identifying certain **miRNAs** that target multiple genes in the network (hence could coordinate the expression of that module). A miRNA might target several HAR-related neurodevelopmental genes – such a miRNA could be a key post-transcriptional regulator in the system. On the drug side, we might find that a number of our genes are targets of existing drugs (perhaps many are kinases or receptors). Some drugs could be psychotropic or metabolic drugs, indicating a known link to brain function. If a HAR-linked gene is targeted by an antiepileptic drug or a gut-brain axis drug, that is noteworthy. These results effectively provide a list of **regulatory miRNAs** and **potential small-molecule modulators** for the genes in our network. This paves the way for experimental validation – e.g., manipulating a miRNA or using a drug to see if it affects the host–microbe interaction outcome.
- **Conserved Pathways across Species:** Cross-species analysis should show that many metabolites and pathways under study are not unique to humans. We expect to see that several metabolite pathways (especially core metabolic ones) are present in both microbes and animals. This result supports the idea that those pathways have been maintained through evolution, possibly due to their importance in host-microbe symbiosis. If any pathways are found in both the gut bacteria and in evolutionarily distant organisms (like insects or nematodes), it suggests a very ancient origin. Conversely, finding a pathway that is present in human and microbiota but not in common model organisms could emphasize the need for caution when using models to study it (or indicate recent co-evolution in humans and their microbiome). The comparative pathway table will detail these findings, emphasizing conserved metabolic interactions as a hallmark of co-evolution.
- **HERV–Disease–Gene Network:** Finally, the overlay of HERV data will result in a network of endogenous retroviral elements connected to diseases and genes. We expect to identify certain HERV families that are known to be active or expressed in the brain and linked to diseases like multiple sclerosis or schizophrenia. The network will show those HERVs connecting to our gene set if applicable. Even if not, it provides a broader context that many genes in the genome (some possibly overlapping HAR regions) have regulatory inputs from HERV insertions. The Cytoscape visualization would illustrate clusters (perhaps diseases grouping with particular HERV families). While exploratory, this result underscores the complexity of the genomic regulatory landscape shaped by both co-evolution with microbes and viral elements.