frontiers | Frontiers in Neuroscience

TYPE Technology and Code
PUBLISHED 12 November 2024
DOI 10.3389/fnins.2024.1400412

Deriving comprehensive literature trends on multi-omics analysis studies in autism spectrum disorder using literature mining pipeline

Dattatray Mongad†, Indhupriya Subramanian† and Anamika Krishanpal*

Life Sciences Research, CTO Unit, Persistent Systems, Pune, India

OPEN ACCESS

EDITED BY
Jernej Kovac,
University Medical Center Ljubljanaju,
Slovenia

REVIEWED BY
Alan Packer,
Simons Foundation, United States
Mario Gorenjak,
University of Maribor, Slovenia
Matija Rijavec,
University Clinic of Pulmonary and Allergic
Diseases Golnik, Slovenia

*CORRESPONDENCE
Anamika Krishanpal
✉ anakrishanpal@gmail.com

†These authors have contributed equally to this work

RECEIVED 22 March 2024
ACCEPTED 26 July 2024
PUBLISHED 12 November 2024

CITATION
Mongad D, Subramanian I and Krishanpal A (2024) Deriving comprehensive literature trends on multi-omics analysis studies in autism spectrum disorder using literature mining pipeline.
Front. Neurosci. 18:1400412.
doi: 10.3389/fnins.2024.1400412

Autism spectrum disorder (ASD) is characterized by highly heterogenous abnormalities in functional brain connectivity affecting social behavior. There is a significant progress in understanding the molecular and genetic basis of ASD in the last decade using multi-omics approach. Mining this large volume of biomedical literature for insights requires considerable amount of manual intervention for curation. Machine learning and artificial intelligence fields are advancing toward simplifying data mining from unstructured text data. Here, we demonstrate our literature mining pipeline to accelerate data to insights. Using topic modeling and generative AI techniques, we present a pipeline that can classify scientific literature into thematic clusters and can help in a wide array of applications such as knowledgebase creation, conversational virtual assistant, and summarization. Employing our pipeline, we explored the ASD literature, specifically around multi-omics studies to understand the molecular interplay underlying autism brain.

KEYWORDS

classification, summarization, topic modeling, generative AI, NER

1 Introduction

Autism spectrum disorder (ASD) refers to a group of complex neurodevelopment disorders caused by differences in the brain that affect communication and behavior. ASD is often difficult to diagnose due to its complex and heterogenous etiology related to neurological development in interrelated systems. Individuals with ASD have an impact in their social, developmental, linguistic, and cognitive functions that evolve throughout their lifetime (Hus and Segal, 2021). It is crucial to characterize such a disorder and investigate the multiple layers of association to get the wholistic picture that can aid in treatment and betterment of affected individuals. Although there are standards such as DSM-5 scale and Developmental Diagnostic Dimensional Interview (3Di) that help in assessing and identifying the affected levels, there is still a lot of scope to further the assessment and treatment options (Kipkemoi et al., 2024). More genetic, social, and behavioral studies are in progress to improve early detection and intervention for ASD.

Multi-omics studies, integrating data from multiple high-throughput methods such as genomics, transcriptomics, proteomics, and epigenomics, have the potential to gain insights into complex molecular and genetic mechanisms that contribute to development of ASD (Higdon et al., 2015). Multi-omics studies have the potential to identify molecular subtypes and unravel novel targets and actionable biomolecules that can aid in the treatment and care for ASD patients (Higdon et al., 2015). Recent advances in sequencing technologies have enabled the identification of copy number variations (CNVs) and rare single nucleotide variants (SNVs) associated with ASD (Bacchelli et al., 2020; Fu et al., 2022; Wang et al., 2022; Zhou et al., 2022). As per latest release (release_16_01_2024) of ASD database, Simons Foundation Autism Research Initiative (SFARI), SNVs, and CNVs from 1,162 genes have been reported as genetic risk factors (Abrahams et al., 2013). As a result of increasing in multi-omics studies, a vast amount of data and literature has accumulated on platforms such as PubMed.

PubMed is the largest repository of published peer-reviewed scientific literature and acts as the gateway for directed search with its advanced query system. However, for a given query even with advanced and special filters, PubMed yields many results that involves huge amount of manual curation to identify relevant articles and further explore them for insights. Autism spectrum disorder (ASD), being an actively researched topic with multiple dimensions such as neuroscience, behavioral studies, diagnosis, and molecular mechanisms, had approximately 5,000 scientific publications in the last 1 year. Extracting meaningful clinical insights and knowledge from these large number publications is challenging due to data complexity, data volume, heterogeneity of ASD, and interdisciplinary nature of research.

Herein, we focus on analyzing the peer-reviewed scientific literature for identifying the trends and insights on multi-omics studies in ASD patients. We have designed a pipeline that can assist in mining large textual data such as biomedical literature for expediting extraction of relevant information. Leveraging topic modeling techniques and large language models (LLMs), the pipeline helps in simplifying text data by clustering them into thematic clusters. Topic modeling uses unsupervised methods to discover hidden patterns from a large collection of textual data (Barde and Bainwad, 2017). This machine learning technique analyzes textual data for similarity patterns and determines word groups that best represent a set of documents to create thematic clusters. We used topic modeling on PubMed abstracts to cluster articles with semantically related keywords and help in identifying the various keywords (topics) associated with the search query. Topic modeling can help in trend analysis and market survey and identify gap areas in research studies thus providing useful insights. Furthermore, we leverage LLMs to showcase possibilities of building an interactive Q&A and summarization model that can be of great value in scientific reporting.

2 Materials and methods

2.1 Data collection

In brief, a search was carried out with the query "(Autism Spectrum Disorder AND *Homo sapiens*) AND (('2013/01/01'[Date - Completion]: '3000'[Date - Completion]))" using *esearch.* A total of 28,304 abstracts published in last 10 years (as on 15 November 2023) were downloaded from PubMed. The abstracts were downloaded and extracted using a Biopython (Cock et al., 2009) implemented in a custom python script.

2.2 Topic modeling

Topic modeling using BERT embeddings and class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) was performed as implemented in BERTopic library (v0.15.0) (Grootendorst, 2022). Although there are multiple methods such as LDA, NMF, and Top2Vec, BERTopic was chosen for its flexibility and user friendliness. Based on the analysis presented by Egger and Yu (2022), BERTopic showed the potency to extract useful information from unstructured textual data.

The abstract text was subjected to lemmatization and filtration of pronouns, determiners, and conjunctions using WordNetLemmatizer implemented in NLTK (3.8.1). Filtered abstracts were fitted on BERTopic model with different combinations of UMAP and HDBSCAN parameters, and seed topics were provided for guided modeling (Table 1). Final model was selected based on the model evaluation metric, topic coherence (C_v and C_umass) (Mifrah, 2020). This evaluation method can be defined as the degree of significance between the words inside a topic and its ease of interpretation from human perspective. While C_v helps in measuring the coherence or similarity of the documents within a topic, C_umass considers the document co-occurrence counts. Higher C-v and lower C_umass (closer to 0) help in choosing a good topic model.

2.3 Named entity recognition

The biological entities within each abstract text were predicted using HunFlair model (Weber et al., 2021) implemented in Flair NLP framework (v0.13.0). It recognizes five important biomedical entity types with high accuracy, namely *Cell Lines*, *Chemicals*, *Diseases*, *Genes,* and *Species*. The gene names and gene symbols predicted by HunFlair were cleaned and compared with gene symbols available in org.Hs.eg.db (v3.16.0) (Carlson, 2022) and cleaned manually.

2.4 Conversational Q&A and summarization using generative AI

We used GPT3.5-turbo model from Azure OpenAI to create Retrieval-Augmented Generation (RAG)-based conversational chat assistant to perform Q&A on the articles (free full-text articles and abstracts). We also used Google's Gemini model from Google Cloud's VertexAI to generate summarized content for selected topics.

2.5 Code availability

A Jupyter notebook containing all codes for PubMed abstract download, processing, topic modeling, and creating a Q&A model is submitted to GitHub at https://github.com/pslomics/ASD_multiomics_analysis.

3 Results

3.1 Literature mining pipeline

The pipeline is designed to simplify deriving insights from large, voluminous, and unstructured biomedical literature that is usually resource and time intensive. The pipeline focuses on grouping the articles into thematic clusters based on the representative words occurring in the abstracts. This helps in providing an overview of the different topics studied for a given query, and the representative topics can help in narrowing down the search space. Thus, the articles are clustered into different topics, and specific topics of interest can be shortlisted for further study. Upon selection of topics, a wide array of applications can be automated to achieve meaningful insights from the data. We showcase that named entity recognition (NER) can be performed on the abstracts to extract entities such as genes, chemicals, drugs, and diseases, to create knowledgebase and knowledge graphs. With the advent of generative AI tools, it is possible to further this pipeline to enable a conversational virtual assistant to interact with the full-text articles to reach out to specific and summarized content. Figure 1 shows the schematic representation of the pipeline.

3.2 Literature cohort for topic model

Direct queries on PubMed such as "Autism Spectrum Disorder AND multi-omics" or "Autism Spectrum Disorder AND omics" yielded less than 100 articles and did not cover the entire space of different omics studies in ASD. We then created a corpus for topic model from PubMed using a more generic search query "Autism Spectrum Disorder AND *Homo sapiens*" in the last 10 years and used topic modeling to identify articles related to multi-omics studies. The resulting 28,304 abstracts were used for training the topic model.

3.3 Choosing the optimal model

Multiple techniques are available to build topic models such as unsupervised, guided, supervised, incremental, multi-modal, and dynamic. As our repertoire consists of a large literature repository,

TABLE 1 List of keywords used as seed list for guided topic modeling.

Seed list 1	Multi-omics	Pan-omics	Omics	Integrative omics	Multiple omics
Seed list 2	Genomics	Mutation	SNP	SNV	Genome
Seed list 3	Transcriptomics	RNA	Gene expression	Mirna	Transcriptome
Seed list 4	Epigenomics	Methylation	Methylome	Epigenetics	Epigenome
Seed list 5	Copy number alteration	Amplification	Deletion	Loss	Gain
Seed list 6	Metabolomics	Metabolome	Metabolite	Lipids	Metabolism
Seed list 7	Proteomics	Protein	Proteome	Biomarkers	Protein folding

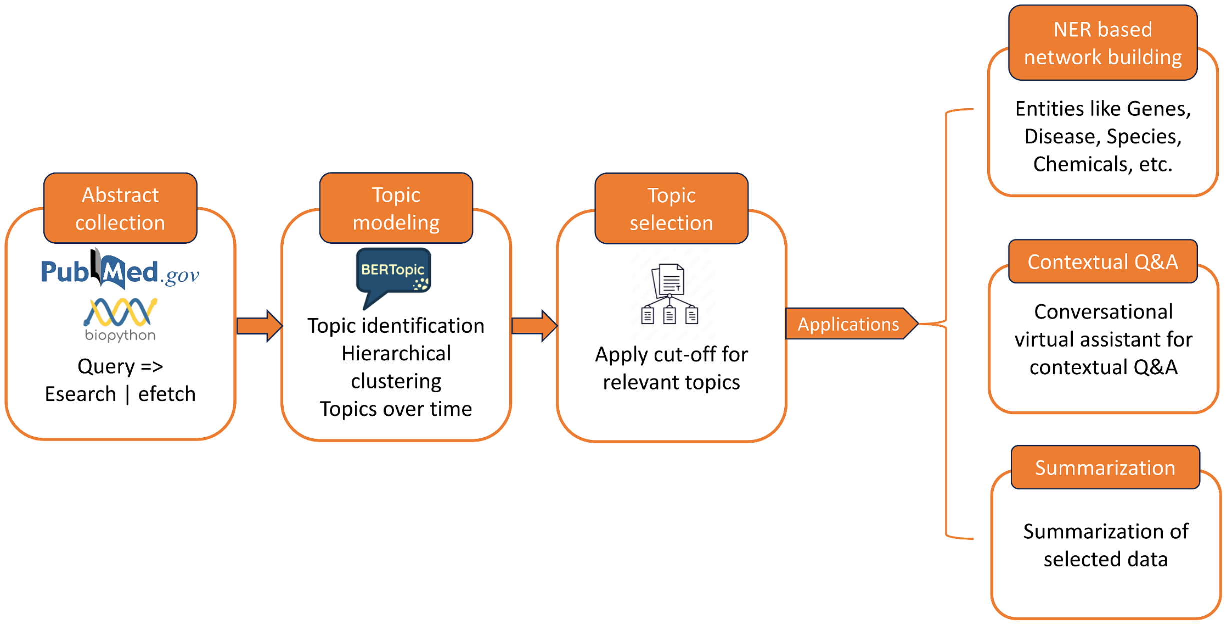


FIGURE 1
Overview of our literature mining pipeline and applications.

we chose unsupervised (library driven to identify the different research areas covered) and guided (to seed the model to look for specific keywords of our interest). Based on the model coherence metrics, we found that the guided model performed better than unsupervised (for unsupervised model, c_v: 0.61 and u_mass: −4.77). Topic model coherence measures the quality and interpretability of topics generated. We calculated two coherence measures: c_v: based on combination of word co-occurrence, and u_mass: rely solely on document co-occurrence statistics. Furthermore, we fine-tuned the guided model using different parameters (UMAP and HDBSCAN) to select the most optimal model for this cohort. Table 2 shows the performance metrics while fine tuning the guided topic model.

We observed that the guided model with 125 topics with "*min_samples*" as 40 showed good coherence (c_v: 0.669 and u_mass: −3.82), and hence, we will be using this model for further discussions. From the input abstracts of 28,304, 10,903 abstracts had a probability score of association to their respective topic greater than 0.8. (Supplementary Table S1). The topics are appended with a numerical value (serial number) for ease of identification.

3.4 Multi-omics research trends in ASD

As our focus is on multi-omics studies in ASD, we selected topics associated with multi-omics based on their representative words. Out of the 125 topics, 17 topics were found to be associated with multi-omics-related representative keywords with 1,283 research articles (Table 3). We present our results that show detailed overview of the multi-omics research studies studied in the context of ASD. As shown in Table 2, the multiple omics such as genomics (SNP, mutations and variants, and CNV), transcriptomics (RNA and miRNA), epigenomics (methylation), metabolomics, and microbiomics were classified as individual topics thus spanning the breadth of multi-omics research studies in ASD. Supplementary Figures S1, S2 show the word distribution score and probability distribution of association of abstracts to their multi-omics topics.

Furthermore, topic modeling helps in analyzing the trend of these topics in the last decade. As shown in Figure 2, the trend analysis plot shows that the topic "2_synaptic_genes_mice_protein," that represents studies associated with synaptic functions, especially mutation studies in mouse models with synaptic genes variants, shows steady increase and is most reported in 2022, followed by studies on impact of gut microbiome on ASD ("6_gut_microbiota_microbiome_gi"). We also observe that studies on *de novo* mutations ("15_genes_variants_mutations_novo") and 16p and 22q deletions ("16_deletion_duplication_16p11_22q11") are showing decline in the recent years, while studies associating ASD with PTEN mutations and macrocephaly ("120_pten_phts_macrocephaly_mutations") and copy number variations ("69_cnvs_cnv_copy_genomic") are less studied.

We explored the other topics that are closely associated with the above listed multi-omics topics using hierarchical clustering (Figure 3). This further clustering of topics based on distance helps in bringing closely related and studied topics together thus providing directions to analyze a group of topics together. The identified 125 topics are clustered into 21 different clusters out of which four clusters contained one or more shortlisted multi-omics topics shown in Table 3. Out of the 4 clusters, first cluster had only multi-omics-related topics associated with variants, mutations, SNPs, and CNVs. This

TABLE 2 Performance metrics of different guided topic models.

UMAP param			HDBscan param		Performance		No. of topics
N_neighbors	N_components	Min_dist	Min_cluster_size	Min_samples	C_V	U_mass	
15	2	0.1	100	100	0.438	−5.327	6
15	2	0.1	50	70	0.458	−5.39	6
15	2	0.1	5	50	0.637	−4.47	109
10	2	0.01	100	100	0.652	−3.834	57
10	2	0.01	50	70	0.66	−3.941	93
10	2	0.01	5	50	0.648	−4.328	142
3	2	0.001	100	100	0.6	−4.826	39
3	2	0.001	50	70	0.633	−4.238	89
3	2	0.001	5	50	0.605	−5.043	182
15	5	0	15	None	0.63	−4.57	100
10	2	0.01	5	100	0.61	−4.61	67
10	2	0.01	50	50	0.658	−3.95	112
10	2	0.01	100	50	0.645	−3.87	61
10	2	0.01	50	10	0.64	−4.29	157
10	2	0.01	50	20	0.65	−4.27	134
10	2	0.01	50	30	0.66	−4.2	125
10	2	0.01	50	40	0.669	−3.82	125

TABLE 3 Topics associated with multi-omics that are selected to study the multi-omics research study in ASD.

Name	Representative words	No. of articles
2_synaptic_genes_mice_protein	['synaptic', 'genes', 'mice', 'protein', 'mouse', 'gene', 'mutations', 'cell', 'proteins', 'expression']	160
6_gut_microbiota_microbiome_gi	['gut', 'microbiota', 'microbiome', 'gi', 'gastrointestinal', 'intestinal', 'microbial', 'axis', 'probiotics', 'fecal']	278
15_genes_variants_mutations_novo	['genes', 'variants', 'mutations', 'novo', 'sequencing', 'gene', 'genetic', 'genome', 'rare', 'exome']	53
16_deletion_duplication_16p11_22q11	['deletion', 'duplication', '16p11', '22q11', 'deletions', 'syndrome', 'region', 'microdeletion', 'carriers', 'duplications']	51
17_methylation_epigenetic_dna_genes	['methylation', 'epigenetic', 'dna', 'genes', 'gene', 'expression', 'chromatin', 'serotonin', 'histone', 'genome']	103
24_cacna1c_mutation_mutations_timothy	['cacna1c', 'mutation', 'mutations', 'timothy', 'gene', 'syndrome', 'qt', 'variant', 'variants', 'ts']	55
31_genetic_psychiatric_schizophrenia_polygenic	['genetic', 'psychiatric', 'schizophrenia', 'polygenic', 'adhd', 'genome', 'disorders', 'wide', 'bipolar', 'major']	51
38_mirnas_mir_mirna_expression	['mirnas', 'mir', 'mirna', 'expression', 'rna', 'genes', 'micrornas', 'gene', 'cell', 'cells']	56
41_variants_sequencing_exome_variant	['variants', 'sequencing', 'exome', 'variant', 'pathogenic', 'delay', 'intellectual', 'patients', 'disability', 'features']	53
50_gut_microbiota_metabolic_diseases	['gut', 'microbiota', 'metabolic', 'diseases', 'metabolites', 'metabolism', 'disease', 'microbiome', 'carnitine', 'mitochondrial']	57
52_metabolic_metabolism_amino_metabolites	['metabolic', 'metabolism', 'amino', 'metabolites', 'acid', 'acids', 'eacute', 'aacute', 'urine', 'plasma']	51
54_fxs_fragile_fmrp_fmr1	['fxs', 'fragile', 'fmrp', 'fmr1', 'cgg', 'syndrome', 'protein', 'retardation', 'translation', 'mrna']	61
57_snps_association_genetic_polymorphisms	['snps', 'association', 'genetic', 'polymorphisms', 'genome', 'allele', 'gene', 'haplotype', 'wide', 'genes']	50
69_cnvs_cnv_copy_genomic	['cnvs', 'cnv', 'copy', 'genomic', 'pathogenic', 'array', 'number', 'chromosomal', 'variants', 'microarray']	54
96_channel_scn2a_channels_voltage	['channel', 'scn2a', 'channels', 'voltage', 'variants', 'gated', 'scn8a', 'sodium', 'calcium', 'mutation']	50
117_biomarker_biomarkers_asd_mcnvs	['biomarker', 'biomarkers', 'asd', 'mcnvs', 'therapeutic', 'new', 'vgcc', 'hypothesis', 'potential', 'nachr']	50
120_pten_phts_macrocephaly_mutations	['pten', 'phts', 'macrocephaly', 'mutations', 'hamartoma', 'tumor', 'germline', 'cancer', 'mutation', 'phosphatase']	50

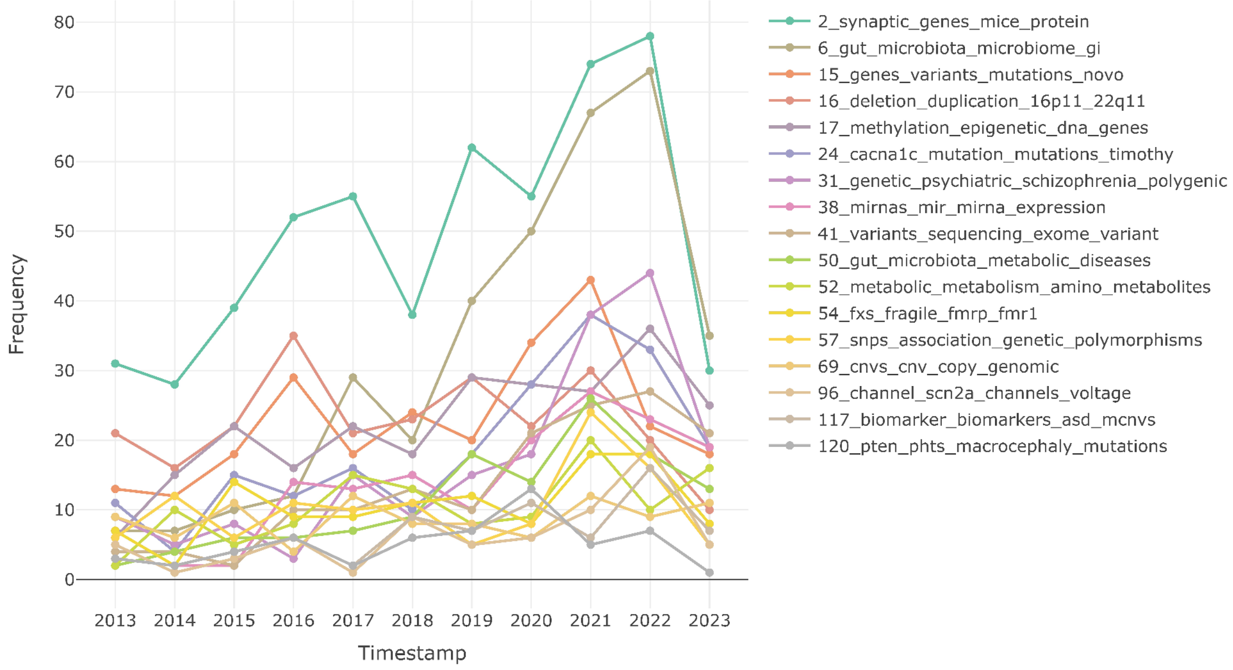


FIGURE 2
Research trends of different multi-omics topics over time. Timestamps are grouped by year.

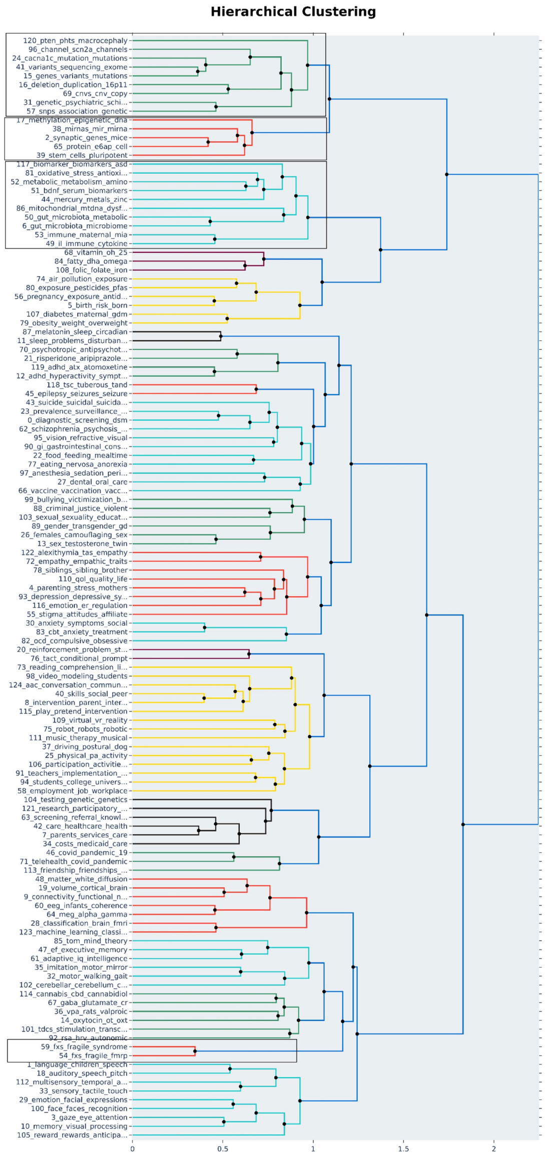


FIGURE 3
Hierarchical clustering: All topics identified by the topic model are clustered on the basis of distance (Ward's linkage) to identify the association among the topics. The 125 topics were grouped into 21 clusters out of which 4 clusters contained the multi-omics associated topics. The clusters containing multi-omics topics are highlighted.

cluster represents the variant analysis studies in ASD. Furthermore, the second cluster showed 4 multi-omics topics, namely, "50_gut_microbiota_metabolic_diseases," "6_gut_microbiota_microbiome_gi," "117_biomarker_biomarkers_asd_mcnvs," and "52_metabolic_metabolism_amino_metabolites" clustered along with mitochondrial dysfunction, oxidative stress, cytokine-driven immune responses, and metal toxicity related topics. Interestingly, there are publications that are exploring the relationship between gastrointestinal tract (GIT) microbiota and mitochondrial dysfunction in ASD (Hu et al., 2020). This validates the association picked up in hierarchical clustering between multiple topics. The third cluster focuses on synaptic gene studies in mouse models ("2_synaptic_genes_mice_protein"), methylation, and miRNA-associated topics ("17_methylation_epigenetic_dna_genes," "38_mirnas_mir_mirna_expression") that are clustered with ubiquitin proteins and stem cell pluripotency topics, highlighting that the mouse models and pluripotent stem cell studies are widely used in studying genetic modifications in ASD (Acab and Muotri, 2015; St. Clair and Johnstone, 2018; Silverman et al., 2022). The fourth cluster is a small, concentrated cluster on fragile X syndrome (FXS) caused due to modifications in the *FMRP* gene ("54_fxs_fragile_fmrp_fmr1"). Thus, clustering of topics can help in deciphering all the related topics in a single view and is useful for screening and selection of studies for further research.

3.5 Genes playing key role in ASD

We performed named entity recognition (NER) on all the abstracts falling under the 17 multi-omics topics, to identify the key molecular players in ASD (Supplementary Table S2). Of the 17 topics, 10 covered the gene mutations and variants in ASD. We grouped these into a broader category called variants, identified the genes mentioned in these abstracts, and found 1,310 unique gene mentions. For validation, this list of genes was compared with widely used ASD variant databases such as autism spectrum consortia, SFARI, and VariCarta (Abrahams et al., 2013; Belmadani et al., 2019; Satterstrom et al., 2019). Figure 4 shows the common genes across these databases and our gene list from NER. We observe that 11 genes are reported additionally in our list. Upon further manual curation of these 11 genes, we found out that 9 of these genes being explored for their role in ASD and other neurological disorders (Table 4). For instance, one of the genes, *GPRASP2*, variations in this gene are implicated to have a role in autism in females (Butler et al., 2015). There is another study reported in mouse models, to analyze the impact of *GPRASP2* mutations in neurological disorders through knockout experiment (Edfawy et al., 2019). Thus, we showcase that NER method can help in identifying biological entities and their relations and can accelerate the creation of knowledgebase and knowledge graphs. Furthermore, this can highlight research trends and potential gaps to direct future experiments and validate their potential role in ASD.

We also present the list of genes that have undergone methylation, genes that are reported in microbiome studies and metabolome data in Supplementary Table S2. These genes can potentially be useful in analyzing the molecular patterns in ASD for therapeutic targets and can help in advancing the knowledgebase of ASD.

3.6 Generative AI powered summarization

One of the wider used applications of generative AI is its ability to summarize large textual information. In this section, we showcase the ability of generative AI to help in generating concise and collated summaries of topics of interest. We chose topics related to fragile X syndrome (FXS) in autism and found two topics ("54_fxs_fragile_fmrp_fmr1," "59_fxs_fragile_syndrome_boys") associated with this syndrome with 116 articles. Using Google cloud services, we used the abstracts to generate summarized content on this topic. Supplementary Table S3 shows the summary of the topics associated with fragile X syndrome reported in the context of ASD. As there are multiple aspects studied in the context of this syndrome, we observe that the summary is divided into behavioral and social aspects in

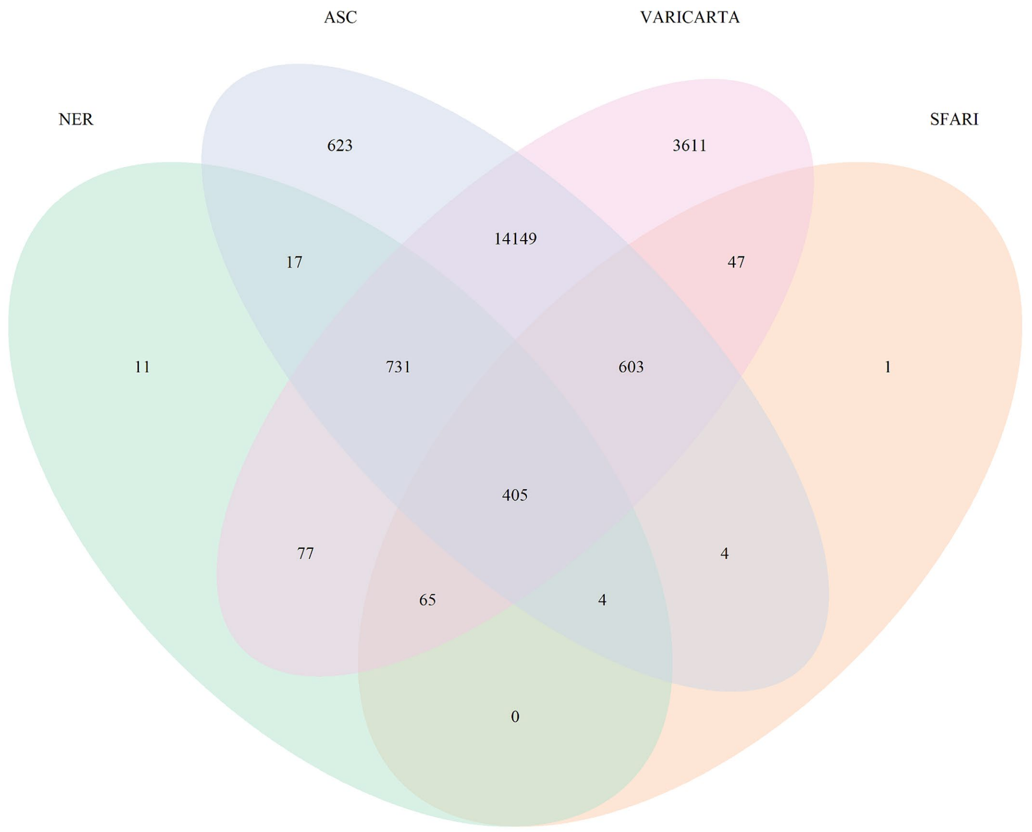
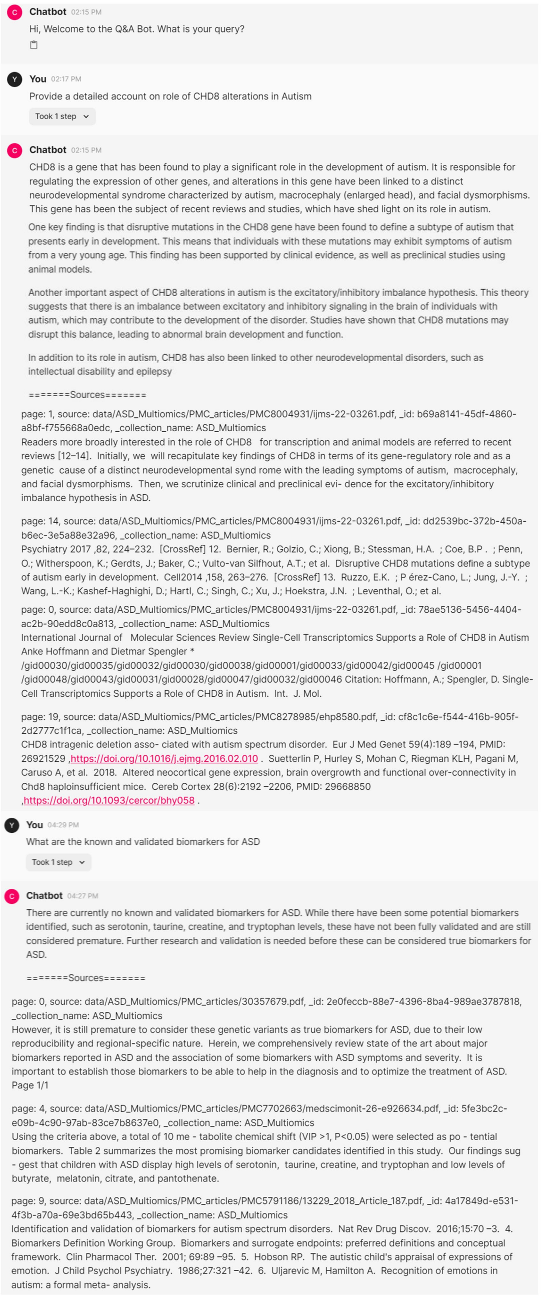


FIGURE 4
Venn diagram showing the common genes across ASC, SFARI, VariCarta, and our NER list.

TABLE 4 Genes identified using NER module of our pipeline that are not reported in ASD variant databases.

Gene Symbol	Gene name	PubMed link	Reported observations in ASD literature
UBE2A	Ubiquitin conjugating enzyme E2 A	https://pubmed.ncbi.nlm.nih.gov/23471985/ (Jahanshad et al., 2013)	This gene is related to other neurological disorder
GPRASP2	G protein-coupled receptor associated sorting protein 2	https://pubmed.ncbi.nlm.nih.gov/25574603/ (Butler et al., 2015)	The role of this gene is being explored in ASD
MSNP1	Moesin pseudogene 1	https://pubmed.ncbi.nlm.nih.gov/27417655/ (Torrico et al., 2017)	This gene has a previously reported GWAS risk variant (could not be replicated)
AMELY	Amelogenin Y-linked	https://pubmed.ncbi.nlm.nih.gov/31852540/ (Maxeiner et al., 2019)	This gene is not related to autism
NAT8B	N-acetyltransferase 8B	https://pubmed.ncbi.nlm.nih.gov/31945187/ (Rigby et al., 2020)	This gene might play a role dysfunctional ER acetylation which is associated with ASD
FOXA2	Forkhead box A2	https://pubmed.ncbi.nlm.nih.gov/32277595/ (Mohammed et al., 2020)	The deletion of this gene might be related to ASD
TCEAL3	Transcription elongation factor a like 3	https://pubmed.ncbi.nlm.nih.gov/36368327/ (Hijazi et al., 2022)	This gene is related to other neurological disorder
MORF4L2	Mortality factor 4 like 2	https://pubmed.ncbi.nlm.nih.gov/36368327/ (Hijazi et al., 2022)	This gene is related to other neurological disorder
PEG13	Paternally expressed 13	https://pubmed.ncbi.nlm.nih.gov/24980697/ (Delgado et al., 2014)	This gene is not yet proven to be associated with ASD
MTCO2P12	MT-CO2 pseudogene 12	https://pubmed.ncbi.nlm.nih.gov/25464930/ (Shen et al., 2015)	This gene is reported to play a role in the pathogenesis of autism



Chatbot 02:15 PM
Hi, Welcome to the Q&A Bot. What is your query?

You 02:17 PM
Provide a detailed account on role of CHD8 alterations in Autism

Chatbot 02:15 PM
CHD8 is a gene that has been found to play a significant role in the development of autism. It is responsible for regulating the expression of other genes, and alterations in this gene have been linked to a distinct neurodevelopmental syndrome characterized by autism, macrocephaly (enlarged head), and facial dysmorphisms. This gene has been the subject of recent reviews and studies, which have shed light on its role in autism.

One key finding is that disruptive mutations in the CHD8 gene have been found to define a subtype of autism that presents early in development. This means that individuals with these mutations may exhibit symptoms of autism from a very young age. This finding has been supported by clinical evidence, as well as preclinical studies using animal models.

Another important aspect of CHD8 alterations in autism is the excitatory/inhibitory imbalance hypothesis. This theory suggests that there is an imbalance between excitatory and inhibitory signaling in the brain of individuals with autism, which may contribute to the development of the disorder. Studies have shown that CHD8 mutations may disrupt this balance, leading to abnormal brain development and function.

In addition to its role in autism, CHD8 has also been linked to other neurodevelopmental disorders, such as intellectual disability and epilepsy

=======Sources=======

page: 1, source: data/ASD_Multiomics/PMC_articles/PMC8004931/ijms-22-03261.pdf, _id: b69a8141-45df-4860-a8bf-f755668a0edc, _collection_name: ASD_Multiomics
Readers more broadly interested in the role of CHD8 for transcription and animal models are referred to recent reviews [12–14]. Initially, we will recapitulate key findings of CHD8 in terms of its gene-regulatory role and as a genetic cause of a distinct neurodevelopmental synd rome with the leading symptoms of autism, macrocephaly, and facial dysmorphisms. Then, we scrutinize clinical and preclinical evi- dence for the excitatory/inhibitory imbalance hypothesis in ASD.

page: 14, source: data/ASD_Multiomics/PMC_articles/PMC8004931/ijms-22-03261.pdf, _id: dd2539bc-372b-450a-b6ec-3e5a88e32a96, _collection_name: ASD_Multiomics
Psychiatry 2017 ,82, 224–232. [CrossRef] 12. Bernier, R.; Golzio, C.; Xiong, B.; Stessman, H.A. ; Coe, B.P . ; Penn, O.; Witherspoon, K.; Gerdts, J.; Baker, C.; Vulto-van Silfhout, A.T.; et al. Disruptive CHD8 mutations define a subtype of autism early in development. Cell2014 ,158, 263–276. [CrossRef] 13. Ruzzo, E.K. ; P érez-Cano, L.; Jung, J.-Y. ; Wang, L.-K.; Kashef-Haghighi, D.; Hartl, C.; Singh, C.; Xu, J.; Hoekstra, J.N. ; Leventhal, O.; et al.

page: 0, source: data/ASD_Multiomics/PMC_articles/PMC8004931/ijms-22-03261.pdf, _id: 78ae5136-5456-4404-ac2b-90edd8c0a813, _collection_name: ASD_Multiomics
International Journal of Molecular Sciences Review Single-Cell Transcriptomics Supports a Role of CHD8 in Autism Anke Hoffmann and Dietmar Spengler * /gid00030/gid00035/gid00032/gid00030/gid00038/gid00001/gid00033/gid00042/gid00045 /gid00001 /gid00048/gid00043/gid00031/gid00028/gid00047/gid00032/gid00046 Citation: Hoffmann, A.; Spengler, D. Single-Cell Transcriptomics Supports a Role of CHD8 in Autism. Int. J. Mol.

page: 19, source: data/ASD_Multiomics/PMC_articles/PMC8278985/ehp8580.pdf, _id: cf8c1c6e-f544-416b-905f-2d2777c1f1ca, _collection_name: ASD_Multiomics
CHD8 intragenic deletion asso- ciated with autism spectrum disorder. Eur J Med Genet 59(4):189 –194, PMID: 26921529 ,https://doi.org/10.1016/j.ejmg.2016.02.010 . Suetterlin P, Hurley S, Mohan C, Riegman KLH, Pagani M, Caruso A, et al. 2018. Altered neocortical gene expression, brain overgrowth and functional over-connectivity in Chd8 haploinsufficient mice. Cereb Cortex 28(6):2192 –2206, PMID: 29668850 ,https://doi.org/10.1093/cercor/bhy058 .

You 04:29 PM
What are the known and validated biomarkers for ASD

Chatbot 04:27 PM
There are currently no known and validated biomarkers for ASD. While there have been some potential biomarkers identified, such as serotonin, taurine, creatine, and tryptophan levels, these have not been fully validated and are still considered premature. Further research and validation is needed before these can be considered true biomarkers for ASD.

=======Sources=======

page: 0, source: data/ASD_Multiomics/PMC_articles/30357679.pdf, _id: 2e0feccb-88e7-4396-8ba4-989ae3787818, _collection_name: ASD_Multiomics
However, it is still premature to consider these genetic variants as true biomarkers for ASD, due to their low reproducibility and regional-specific nature. Herein, we comprehensively review state of the art about major biomarkers reported in ASD and the association of some biomarkers with ASD symptoms and severity. It is important to establish those biomarkers to be able to help in the diagnosis and to optimize the treatment of ASD. Page 1/1

page: 4, source: data/ASD_Multiomics/PMC_articles/PMC7702663/medscimonit-26-e926634.pdf, _id: 5fe3bc2c-e09b-4c90-97ab-83ce7b8637e0, _collection_name: ASD_Multiomics
Using the criteria above, a total of 10 me - tabolite chemical shift (VIP >1, P<0.05) were selected as po - tential biomarkers. Table 2 summarizes the most promising biomarker candidates identified in this study. Our findings sug - gest that children with ASD display high levels of serotonin, taurine, creatine, and tryptophan and low levels of butyrate, melatonin, citrate, and pantothenate.

page: 9, source: data/ASD_Multiomics/PMC_articles/PMC5791186/13229_2018_Article_187.pdf, _id: 4a17849d-e531-4f3b-a70a-69e3bd65b443, _collection_name: ASD_Multiomics
Identification and validation of biomarkers for autism spectrum disorders. Nat Rev Drug Discov. 2016;15:70 –3. 4. Biomarkers Definition Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001; 69:89 –95. 5. Hobson RP. The autistic child's appraisal of expressions of emotion. J Child Psychol Psychiatry. 1986;27:321 –42. 6. Uljarevic M, Hamilton A. Recognition of emotions in autism: a formal meta- analysis.

FIGURE 5
Interactive virtual assistant application to create answers supported with references from scientific literature.

different groups, genetics, psychiatric, etc. Thus, this module helps in generating succinct summaries that can be highly beneficial for report generation.

3.7 Generative AI powered conversational chat assistant

In this section, we highlight the ability to leverage generative AI techniques to build a virtual assistant to quickly mine scientific literature for better insights. We had downloaded all 1,283 multi-omics topic-associated research articles (493 free full-text articles), to enable a chat application that can create an interface to search the articles based on simple linguistic queries. Figure 5 shows the conversation with virtual assistant to reach to specific answers to a given query. The answers are further supported with references for validation, thus, containing the hallucinations of generative AI models. We show that our pipeline significantly reduces the manual efforts required in curating specific information related to a topic.

4 Discussion

Literature mining plays a vital role in generating new hypothesis, validation of research, knowledge-gap analysis, and/or to study the market trends. Scientific literature-based evidence gathering has wide applications in pharmaceutical, clinical, and research communities. Tremendous amount of time and manpower is spent to curate and validate valuable information from these studies. We propose a pipeline that can expedite the scoping of data for actionable insights from clinical and biomedical literature. Though we have demonstrated the application using scientific literature, the pipeline can be extended to mine voluminous textual data such as clinical reports and health records.

ASD is a complex disorder which poses difficulty in diagnosis and treatment owing to its heterogenous symptoms. Although there are a few studies that have used multi-omics approaches for studying this disorder (Troisi et al., 2020; Nomura et al., 2021), there is a lack of holistic view of the research studies carried for different omic-profiles of ASD.

Leveraging our pipeline, we provide a bird's eye view to worm's eye view of the multi-omics studies in autism in the last 10 years. We have used topic modeling to provide an overview of the different research arenas in ASD. The guided modeling yielded 125 topics or thematic clusters that summarizes the different aspects of ASD studies. We narrowed down the topics by identifying omics-related representative keywords and identified 1,283 research articles that are associated with 17 different multi-omics topics. Gathering these articles directly from search engines would require multiple searches or multiple combinations of keywords. The topics identified showcased high probability of association, thus highlighting the efficiency of the model to mine, identify, and cluster similar articles into a thematic cluster. The identified topic encompassed the genomics, transcriptomics, epigenomics, and microbiomics thus spanning the length and breadth of multi-omics research in ASD. However, topic modeling may not be able to tag all the abstracts into a particular theme, and such abstracts are grouped under "miscellaneous." These abstracts are not available for further applications of the pipeline. Fine-tuning the modeling with different set of parameters can help in reducing the articles tagged as "miscellaneous."

Visualizing the results of topic modeling can aid in assessing closely associated topics based on distance, well represented topics, and topic trends over time. This helps in identifying specific topics of interest and their counterparts for downstream actions.

We have highlighted three possible applications in our pipeline. We have used NER to find the genes that have significant genetic modifications such as SNVs, CNVs, and epigenetic changes, which are reported to play a key role in ASD. The identified genes are validated by comparing with public databases on ASD variants, thus emphasizing that our pipeline can facilitate creation of curated knowledge bases.

We have employed the latest generative AI tools and techniques for faster data to insights in the form of a conversational virtual assistant (Q&A) and summarization. Our virtual assistant supports the results with appropriate references thus reducing the hallucinations in the answers and establishes a robust method to validate the answers.

Our literature mining pipeline significantly improves the efforts required in extracting meaningful details from literature. The semi-automated pipeline saves time, cost, and manual efforts required for curation and provides a perfect balance between speed and accuracy. The pipeline is agnostic to domain and can be extended to large cohorts of textual data such as reports, blogs, or any articles. It can address a wide range of applications such as market research, gap analysis, and trend analysis.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

DM: Writing – review & editing, Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization. IS: Data curation, Methodology, Writing – original draft, Visualization, Conceptualization, Investigation, Resources, Validation, Supervision, Writing – review & editing. AK: Resources, Validation, Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We would like to thank Srikant Verma and Shiva Kumar of Persistent Systems, for their valuable inputs in implementing the generative AI-based applications of the pipeline.

Conflict of interest

DM, IS, and AK were employed by company Persistent Systems. The authors declare that this study received funding from Persistent Systems. The funder had the following involvement in the study: Infrastructure and compute support for development of the pipeline and support for publication.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2024.1400412/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Plot showing the word scores of the top 5 words under each topic.

SUPPLEMENTARY FIGURE S2
Plot showing the probability distribution scores of abstracts under each topic.

SUPPLEMENTARY TABLE S1
Results of topic modeling.

SUPPLEMENTARY TABLE S2
Gene lists identified by the NER module from multi-omics abstracts.

SUPPLEMENTARY TABLE S3
Summary of the abstracts tagged under fragile X syndrome (FXS) topic using generative AI tools.

References

Abrahams, B. S., Arking, D. E., Campbell, D. B., Mefford, H. C., Morrow, E. M., Weiss, L. A., et al. (2013). SFARI gene 2.0: a community-driven knowledgebase for the autism Spectrum disorders (ASDs). *Mol. Autism.* 4:36. doi: 10.1186/2040-2392-4-36

Acab, A., and Muotri, A. R. (2015). The use of induced pluripotent stem cell technology to advance autism research and treatment. *Neurotherapeutics* 12, 534–545. doi: 10.1007/s13311-015-0354-x

Bacchelli, E., Cameli, C., Viggiano, M., Igliozzi, R., Mancini, A., Tancredi, R., et al. (2020). An integrated analysis of rare CNV and exome variation in autism Spectrum disorder using the Infinium PsychArray. *Sci. Rep.* 10:3198. doi: 10.1038/s41598-020-59922-3

Barde, B. V., and Bainwad, A. M. (2017). "An overview of topic modeling methods and tools" in In 2017 international conference on intelligent computing and control systems (ICICCS), (International conference in intelligent computing and control systems) 745–750.

Belmadani, M., Jacobson, M., Holmes, N., Phan, M., Nguyen, T., Pavlidis, P., et al. (2019). VariCarta: a comprehensive database of harmonized genomic variants found in autism Spectrum disorder sequencing studies. *Autism Res.* 12, 1728–1736. doi: 10.1002/aur.2236

Butler, M. G., Rafi, S. K., Hossain, W., Stephan, D. A., and Manzardo, A. M. (2015). Whole exome sequencing in females with autism implicates novel and candidate genes. *Int. J. Mol. Sci.* 16, 1312–1335. doi: 10.3390/ijms16011312

Carlson, Marc. (2022). Org.Hs.Eg.Db: Genome wide annotation for human.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Delgado, M. S., Camprubí, C., Tümer, Z., Martínez, F., Milà, M., and Monk, D. (2014). Screening individuals with intellectual disability, autism and Tourette's syndrome for KCNK9 mutations and aberrant DNA methylation within the 8q24 imprinted cluster. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 165, 472–478. doi: 10.1002/ajmg.b.32250

Edfawy, M., Guedes, J. R., Pereira, M. I., Laranjo, M., Carvalho, M. J., Gao, X., et al. (2019). Abnormal mGluR-mediated synaptic plasticity and autism-like Behaviours in Gprasp2 mutant mice. *Nat. Commun.* 10:1431. doi: 10.1038/s41467-019-09382-9

Egger, R., and Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Front. Sociol.* 7:6498. doi: 10.3389/fsoc.2022.886498

Fu, J. M., Kyle Satterstrom, F., Peng, M., Brand, H., Collins, R. L., Dong, S., et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* 54, 1320–1331. doi: 10.1038/s41588-022-01104-0

Grootendorst, M. (2022). BERTopic: neural topic modeling with a class-based TF-IDF procedure. *arXiv [Preprint].* doi: 10.48550/arXiv.2203.05794

Higdon, R., Earl, R. K., Stanberry, L., Hudac, C. M., Montague, E., Stewart, E., et al. (2015). The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism Spectrum disorders. *OMICS* 19, 197–208. doi: 10.1089/omi.2015.0020

Hijazi, H., Reis, L. M., Pehlivan, D., Bernstein, J. A., Muriello, M., Syverson, E., et al. (2022). TCEAL1 loss-of-function results in an X-linked dominant neurodevelopmental syndrome and drives the neurological disease trait in Xq22.2 deletions. *Am. J. Hum. Genet.* 109, 2270–2282. doi: 10.1016/j.ajhg.2022.10.007

Hu, T., Dong, Y., He, C., Zhao, M., and He, Q. (2020). The gut microbiota and oxidative stress in autism Spectrum disorders (ASD). *Oxidative Med. Cell. Longev.* 2020, 8396708–8396713. doi: 10.1155/2020/8396708

Hus, Y., and Segal, O. (2021). Challenges surrounding the diagnosis of autism in children. *Neuropsychiatr. Dis. Treat.* 17, 3509–3529. doi: 10.2147/NDT.S282569

Jahanshad, N., Rajagopalan, P., Hua, X., Hibar, D. P., Nir, T. M., Toga, A. W., et al. (2013). Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity. *Proc. Natl. Acad. Sci.* 110, 4768–4773. doi: 10.1073/pnas.1216206110

Kipkemoi, P., Kariuki, S. M., Gona, J., Murimi, P., Mandy, W., and Abubakar, A. (2024). Utility of the 3Di short version in the identification and diagnosis of autism in children at the Kenyan coast. *Front. Psych.* 15:4929. doi: 10.3389/fpsyt.2024.1234929

Maxeiner, S., Sester, M., and Krasteva-Christ, G. (2019). Novel human sex-typing strategies based on the autism candidate gene NLGN4X and its male-specific Gametologue NLGN4Y. *Biol. Sex Differ.* 10:62. doi: 10.1186/s13293-019-0279-x

Mifrah, S. (2020). Topic modeling coherence: a comparative study between LDA and NMF models using COVID'19 Corpus. *Int. J. Adv. Trends Comp. Sci. Eng.* 9, 5756–5761. doi: 10.30534/ijatcse/2020/231942020

Mohammed, I., Al-Khawaga, S., Bohanna, D., Shabani, A., Khan, F., Love, D. R., et al. (2020). Haploinsufficiency of the FOXA2 associated with a complex clinical phenotype. *Mol. Genet. Genomic Med.* 8:e1086. doi: 10.1002/mgg3.1086

Nomura, J., Mardo, M., and Takumi, T. (2021). Molecular signatures from multi-omics of autism Spectrum disorders and schizophrenia. *J. Neurochem.* 159, 647–659. doi: 10.1111/jnc.15514

Rigby, M. J., Ding, Y., Farrugia, M. A., Feig, M., Cortese, G. P., Mitchell, H., et al. (2020). The endoplasmic reticulum acetyltransferases ATase1/NAT8B and ATase2/NAT8 are differentially regulated to adjust engagement of the secretory pathway. *J. Neurochem.* 154, 404–423. doi: 10.1111/jnc.14958

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., et al. (2019). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *bioRxiv*. doi: 10.1101/484113

Shen, C., Huo, L.-r., Zhao, X.-l., Wang, P.-r., and Zhong, N. (2015). Novel interactive Partners of Neuroligin 3: new aspects for pathogenesis of autism. *J. Mol. Neurosci.* 56, 89–101. doi: 10.1007/s12031-014-0470-9

Silverman, J. L., Thurm, A., Ethridge, S. B., Soller, M. M., Petkova, S. P., Abel, T., et al. (2022). Reconsidering animal models used to study autism Spectrum disorder: current state and optimizing future. *Genes Brain Behav.* 21:e12803. doi: 10.1111/gbb.12803

St. Clair, D., and Johnstone, M. (2018). Using mouse transgenic and human stem cell technologies to model genetic mutations associated with schizophrenia and autism. *Phil. Trans. Royal Soc. B* 373:20170037. doi: 10.1098/rstb.2017.0037

Torrico, B., Chiocchetti, A. G., Bacchelli, E., Trabetti, E., Hervás, A., Franke, B., et al. (2017). Lack of replication of previous autism Spectrum disorder GWAS hits in European populations. *Autism Res.* 10, 202–211. doi: 10.1002/aur.1662

Troisi, J., Autio, R., Beopoulos, T., Bravaccio, C., Carraturo, F., Corrivetti, G., et al. (2020). Genome, environment, microbiome and metabolome in autism (GEMMA) study design: biomarkers identification for precision treatment and primary prevention of autism Spectrum disorders by an integrated multi-omics systems biology approach. *Brain Sci.* 10:743. doi: 10.3390/brainsci10100743

Wang, T., Kim, C. N., Bakken, T. E., Gillentine, M. A., Henning, B., Mao, Y., et al. (2022). Integrated gene analyses of de novo variants from 46,612 trios with autism and developmental disorders. *Proc. Natl. Acad. Sci. USA* 119:e2203491119. doi: 10.1073/pnas.2203491119

Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U., and Akbik, A. (2021). HunFlair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics* 37, 2792–2794. doi: 10.1093/bioinformatics/btab042

Zhou, X., Feliciano, P., Shu, C., Wang, T., Astrovskaya, I., Hall, J. B., et al. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat. Genet.* 54, 1305–1319. doi: 10.1038/s41588-022-01148-2