scientific reports

Check for updates

OPEN

Machine learning of clinical phenotypes facilitates autism screening and identifies novel subgroups with distinct transcriptomic profiles

Wasana Yuwattana[1], Thanit Saeliw[2], Marlieke Lisanne van Erp[1], Chayanit Poolcharoen[3], Songphon Kanlayaprasit[2], Pon Trairatvorakul[4], Weerasak Chonchaiya[4], Valerie W. Hu[5] & Tewarit Sarachana[2]✉

Autism spectrum disorder (ASD) presents significant challenges in diagnosis and intervention due to its diverse clinical manifestations and underlying biological complexity. This study explored machine learning approaches to enhance ASD screening accuracy and identify meaningful subtypes using clinical assessments from AGRE database integrated with molecular data from GSE15402. Analysis of ADI-R scores from a large cohort of 2794 individuals demonstrated that deep learning models could achieve exceptional screening accuracy of 95.23% (CI 94.32–95.99%). Notably, comparable performance was maintained using a streamlined set of just 27 ADI-R sub-items, suggesting potential for more efficient diagnostic tools. Clustering analyses revealed three distinct subgroups identifiable through both clinical symptoms and gene expression patterns. When ASD were grouped based on clinical features, stronger associations emerged between symptoms and underlying molecular profiles compared to grouping based on gene expression alone. These findings suggest that starting with detailed clinical observations may be more effective for identifying biologically meaningful ASD subtypes than beginning with molecular data. This integrated approach combining clinical and molecular data through machine learning offers promising directions for developing more precise screening methods and personalized intervention strategies for individuals with ASD.

Keywords Autism spectrum disorder, Artificial intelligence, Machine learning, Screening, Subgrouping, Autism diagnostic interview-revised, Multi-omics, Transcriptomics, Phenome, Precision medicine

Autism spectrum disorder (ASD) is a neurodevelopmental condition that typically manifests in early childhood and displays considerable diversity in its presentation. It is marked by challenges in social interaction and communication, alongside repetitive behaviors, and restricted interests[1,2]. The global prevalence of ASD is estimated to be between 1–2%[3,4]. In the United States, the latest prevalence estimates indicate a notable rate of 1 in 36 children aged 8 years old affected by ASD, as reported by the Centers for Disease Control and Prevention (CDC)[5]. Moreover, ASD demonstrates a notable sex bias, with males showing a prevalence 3–4 times higher than females[3,5]. However, variations in how ASD individuals are defined and identified can lead to differences in estimated prevalence rates across countries[6]. Furthermore, the early diagnosis of ASD lead to earlier intervention which can potentially have a beneficial impact to help children develop skills for their living[7].

[1]The Ph.D. Program in Clinical Biochemistry and Molecular Medicine, Department of Clinical Chemistry, Faculty of Allied Health Sciences, Chulalongkorn University, Bangkok 10330, Thailand. [2]Chulalongkorn Autism Research and Innovation Center of Excellence (ChulaACE), Department of Clinical Chemistry, Faculty of Allied Health Sciences, Chulalongkorn University, Bangkok 10330, Thailand. [3]The M.Sc. Program in Clinical Biochemistry and Molecular Medicine, Department of Clinical Chemistry, Faculty of Allied Health Sciences, Chulalongkorn University, Bangkok 10330, Thailand. [4]Center of Excellence for Maximizing Children's Developmental Potential, Division of Growth and Development, Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand. [5]Department of Biochemistry and Molecular Medicine, School of Medicine and Health Sciences, The George Washington University, Washington, DC 20037, USA. ✉email: tewarit.sa@chula.ac.th

ASD displays diversity in symptom presentation and severity levels. This variation is believed to be influenced by molecular mechanisms[8], epigenetic factors[9–12] and environmental factors[13–16], contributing to the distinct clinical phenotypes observed within different subpopulations of ASD. Several studies suggested that categorizing individuals with ASD into subgroups can aid in pinpointing candidate genes and molecular mechanisms associated with ASD pathology within each subgroup[17,18]. This insight holds potential for advancing the development of ASD diagnosis and intervention strategies in the future. At present, the gold standard of ASD diagnosis is still based on behavior observation and parent/caregiver interviews. The Autism Diagnostic Interview-Revised (ADI-R)[19], Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)[20], and the Childhood Autism Rating Scale, Second Edition (CARS-2)[21] are the widely-used ASD assessment tools with contents addressing abnormal behaviors related to Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria[2]. However, the current diagnostic process is complex, time-intensive, and reliant on expert interpretation, potentially resulting in delayed intervention initiation.

In recent years, there has been a notable surge in research devoted to developing machine learning (ML) models for diagnosing or screening ASD. Wall et al. endeavored to streamline the questionnaires of ADI-R and ADOS assessments to reduce time consumption. Their investigation revealed that the optimal model utilized 7 out of the 93 items from the ADI-R and 8 out of the 29 items in Module 1 of the ADOS, achieving classification accuracies of 99.9% and 100%, respectively[22,23]. Bone et al. employed a combination of ADI-R and Social Responsiveness Scale (SRS) scores to enhance the accuracy of Support vector machines (SVM). Their research demonstrated that the SVM model, incorporating just five combined ADI-R and SRS items, effectively categorized ASD in individuals both below and above age 10, achieving sensitivities of 89.2% and 86.7%, and specificities of 59.0% and 53.4%, respectively[24]. These studies serve as inspiration for the development of ML tools to aid in overcoming limitations of conventional approaches. Although numerous studies over the past decade have endeavored to develop ML models utilizing ASD assessment scores, personal characteristics, or other abnormal clinical behaviors, many of these studies have employed small sample sizes. Therefore, there remains a need to assess the effectiveness of ML models using larger sample sizes. Additionally, the application of unsupervised ML holds promises for identifying latent ASD subgroups. Stratifying ASD individuals into subgroups based on their clinical symptoms or their gene expression could pave the way for precision medicine approaches in ASD intervention. Nevertheless, further investigation is needed to evaluate the effectiveness of different ML methods for diagnosing and subgrouping ASD.

This study aims to apply various ML algorithms to enhance ASD diagnosis/screening and subgrouping using a substantially larger sample size ($n = 2480$ ASD cases) compared to previous work. We analyzed two datasets: ADI-R scores and matched transcriptome profiling from GSE15402. The GSE15402 dataset was previously analyzed by Hu et al.[25,26], performing cluster analysis on ADI (-R) scores from 1954 individuals, combining data from both 1995 and 2003 versions of ADI (-R) questionnaire, to identify four distinct phenotypic subgroups then selected 87 representative ASD cases for gene expression analysis. The current study significantly expands upon this work by analyzing a much larger cohort of 2480 ASD cases using only the standardized 2003 ADI-R version, providing greater statistical power and consistency in current clinical assessment. The objective is to evaluate whether ML can improve ASD screening and to determine if subgrouping based on clinical symptoms or gene expression is more effective in linking gene expression with clinical features.

Results

Data preparation

In this study, the ADI-R data was obtained from the Autism Genetic Resource Exchange (AGRE) repository, which is a large database established to accelerate ASD research by providing genetic and phenotypic data from families affected by ASD to qualified researchers worldwide. From AGRE, we initially obtained ADI-R version 2003 data from 2800 individuals, consisting of 1482 ASD and 318 non-ASD individuals. According to the data modification and preparation, a total of 2794 individuals with ADI-R scores were included, comprising 2480 ASD individuals and 314 non-ASD individuals, as shown in Figs. 1 and 2a. The age of participants ranged from 1.69 to 47.68 years for ASD and 2.10 to 22.45 years for non-ASD cases, with the age distribution displayed in Fig. 2b,c. After preparation, the dataset was divided into three groups: Training_samples, Validate_samples for ASD screening, and ASD_samples for subgrouping, as illustrated in Fig. 2d.

ML models for ASD screening using ADI-R scores showed performance variation across supervised ML algorithms

In this study, we aimed to develop machine learning (ML) models for efficient ASD screening using a large-scale ADI-R dataset. Figure 1 illustrates our comprehensive workflow for developing and evaluating various ML approaches to ASD screening, The Training_samples ($n = 2514$) was utilized to train seven supervised ML algorithms—Naïve Bayes (NB), Decision Tree (DTree), Random Forest (RF), k-Nearest Neighbor (k-NN), Logistic Regression (LR), Support Vector Machine (SVM) and Deep Learning (DL)—using Altair AI Studio software to develop a ML model for ASD screening. Initially, each algorithm underwent optimization of parameter combinations, with the best combined parameters of each algorithm was used to train model (Supplementary Data 1). The results of the seven algorithms are presented in Table 1. DL provided the highest accuracy at 95.23% (CI 94.32–95.99%), although it did not achieve the highest sensitivity and specificity values, which were 97.94% (CI 97.26–98.45%) and 73.76% (CI 68.33–78.55%), respectively. Supervised predictive algorithms within the Tree family, such as DTree, RF, and SVM, demonstrated very high sensitivity, reaching up to 98.5–99.7%. However, these models had significantly lower specificity, ranging from 50.00 to 56.00%, which resulted in lower overall accuracy compared to DL. NB was the only supervised predictive algorithm that produced a relatively high specificity of 81.6% (CI 76.62–85.65%), but its sensitivity was lower compared to other supervised algorithms. Based on the ROC curve analysis, DL showed the highest True Positive Rate

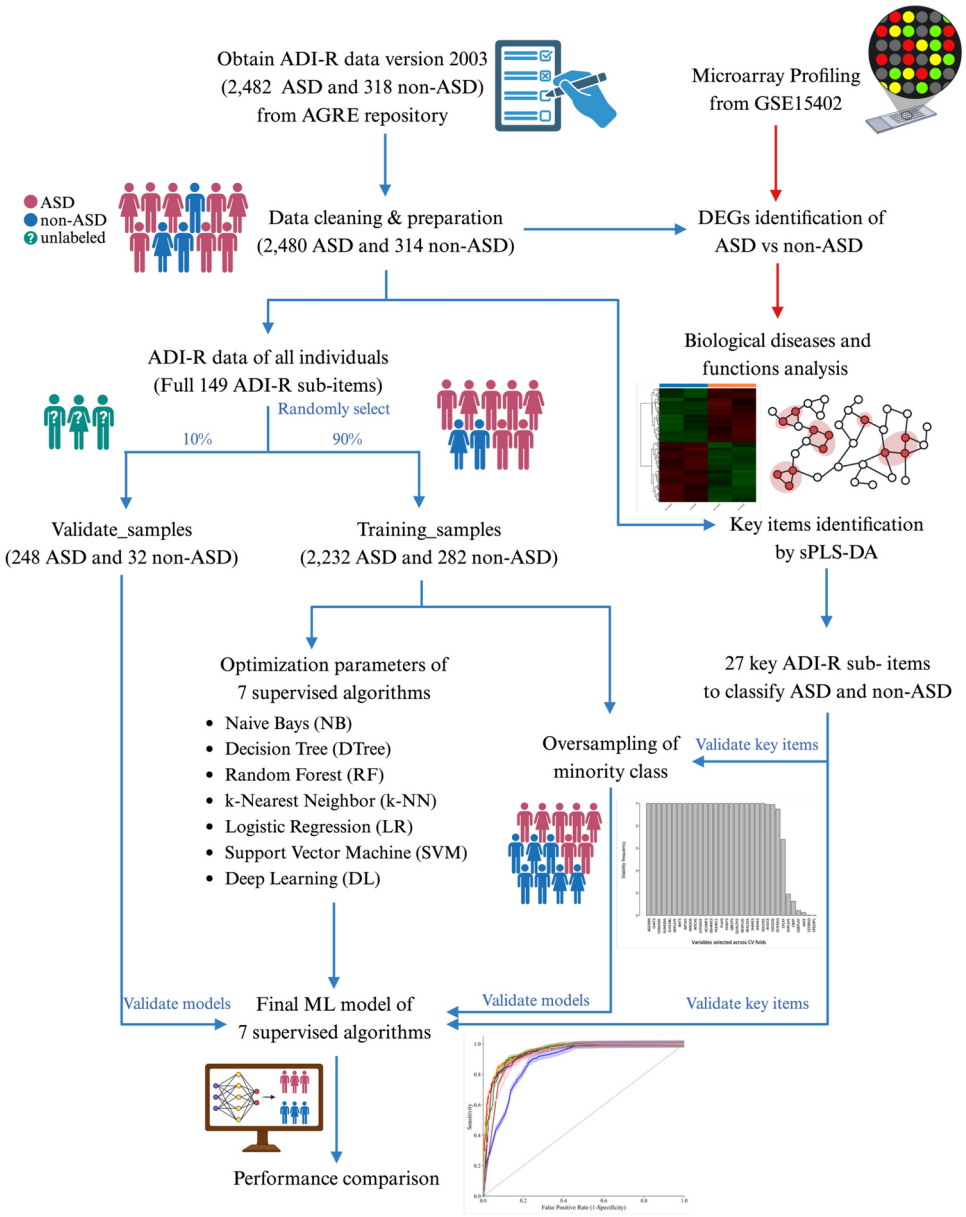

Fig. 1. Machine learning framework for ASD screening using ADI-R. *DEGs* Differentially expressed genes, *sPLS-DA* sparse partial least squares discriminant analysis, *ML* machine learning, *unlabeled* sample data without revealed diagnoses, *AGRE* Autism Genetic Resource Exchange Consortium. This Figure was created in BioRender.com.

and the lowest False Positive Rate as shown in Fig. 3a. Then the Validate_samples dataset ($n = 280$), consisting of ADI-R data not previously used for training the ML model, was used to validate the DL model, the results showed an accuracy of 92.50%, with a sensitivity of 95.56% and a specificity of 68.75% (Supplementary Data 2). This suggested that DL is the most suitable supervised predictive algorithm for developing a ML model to screen for ASD using ADI-R scores, particularly when working with a large sample size.

Since diagnosing ASD using the ADI-R is time-consuming due to its 93 items, many of which contain subquestions, we sought to reduce the number of ADI-R items by applying sparse Partial Least Squares Discriminant Analysis (sPLS-DA). sPLS-DA is a variation of sparse PLS in which the lasso penalty is specifically applied to the loading vector, allowing for the selection of the most important variables while reducing less relevant ones. After tuning the sPLS-DA model, it was found that only 27 ADI-R items from component 1 were sufficient for screening ASD from non-ASD individuals. The tenfold cross validation stability of these 27 selected items, 23 showed a stability frequency of 100%, indicating how consistently the same variables were selected (Fig. 3b and Table 2). These 27 ADI-R items were then used to train the ML models using the same parameter combinations applied to the full ADI-R items, with the results shown in Table 1. To further evaluate these 27 selected items, correlation analysis was performed to examine their relationships. Spearman correlation

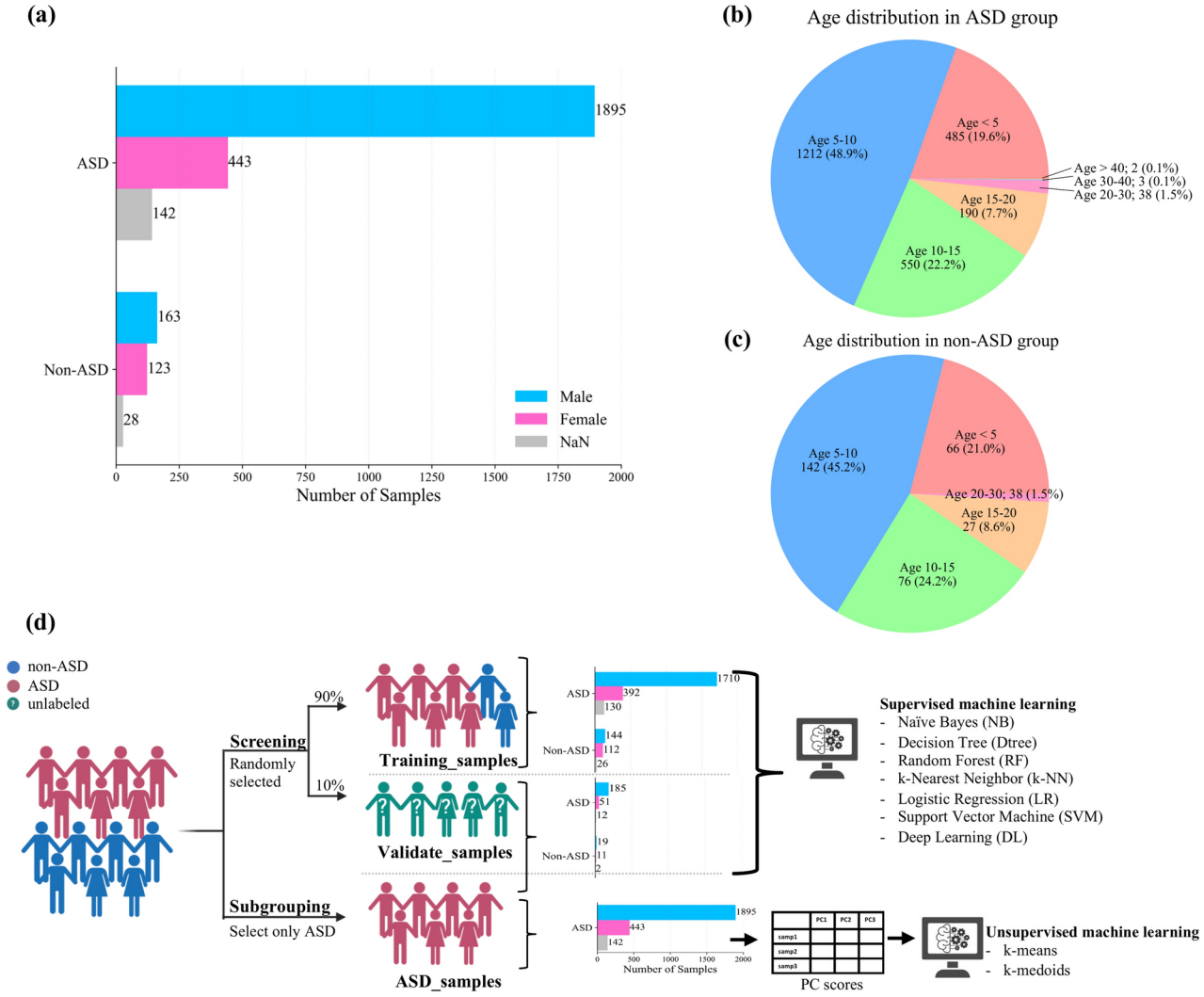

Fig. 2. Demographics of ADI-R data after data preparation. (**a**) Number of complete ADI-R samples, (**b**) age distribution in ASD group, (**c**) age distribution in non-ASD group, (**d**) experimental workflow for ADI-R data and machine learning development. *unlabeled* samples data without revealed diagnoses, *PC scores* principal component scores.

analysis revealed that all items exhibited low to moderate correlations (Supplementary Data 3), with the highest correlations observed between CSOCDIS-SOCDIS5 (r = 0.68) and AGEABN-JUDGAGE (r = 0.66). These moderate correlations, combined with the high stability in cross-validation, support that the sPLS-DA effectively identified both discriminative and non-redundant items for ASD screening (Fig. 3d). This was further supported by model performance, as the models trained with the reduced set of ADI-R items demonstrated comparable performance to those trained with the complete item set, with overall accuracy and sensitivity differing by only 1–2%. Three supervised predictive algorithms—NB, DTree, and RF—showed enhanced accuracy in discriminating between ASD and non-ASD individuals. Additionally, four algorithms (NB, DTree, k-NN, and LR) exhibited improved sensitivity. Notably, RF was the only model that achieved a 5% increase in specificity. Analysis of ROC curves (Fig. 3c) revealed that RF demonstrated superior performance in screening ASD from non-ASD using only the 27 selected ADI-R items. Based on these findings, we concluded that the reduced set of 27 ADI-R items, particularly when used with the RF algorithm, provides an effective and streamlined approach for ASD screening without compromising diagnostic performance (Fig. 3d).

Following our initial findings where DL achieved the highest accuracy but moderate specificity, we recognized a critical limitation in our imbalance training dataset with 2232 ASD and 282 non-ASD individuals. The imbalance in classification problems can lead to biased model performance, where the algorithm may become overly sensitive to the majority class (ASD) while performing poorly on the minority class (non-ASD), potentially resulting in high sensitivity but low specificity in real-world applications. To address these concerns and ensure model performance across both classes, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to upsample non-ASD class in the training dataset, creating an equally balanced representation of both classes. After SMOTE application (performance results shown in Supplementary Data 4), DL maintained its strong

Model	Accuracy (CI)		Sensitivity (CI)		Specificity (CI)		Precision (CI)	
	All items	27 items	All items	27 items	All items	27 items	All items	27 items
Deep learning (DL)	95.23% (94.32%-95.99%)	93.91% (92.91%-94.78%)	97.94% (97.26%-98.45%)	96.68% (95.86%-97.35%)	73.76% (68.33%-78.55%)	71.99% (66.48%-76.90%)	96.73% (95.91%-97.38%)	96.47% (95.62%-97.16%)
Random forest (RF)	94.75% (93.81%-95.56%)	94.99% (94.06%-95.77%)	99.69% (99.35%-99.85%)	99.24% (98.78%-99.52%)	55.67% (49.84%-61.36%)	61.35% (55.55%-66.84%)	94.68% (93.70%-95.52%)	95.31% (94.37%-96.10%)
k-nearest neighbors (k-NN)	94.67% (93.72%-95.48%)	94.55% (93.59%-95.37%)	98.07% (97.42%-98.57%)	99.51% (99.12%-99.72%)	67.73% (62.07%-72.92%)	55.32% (49.48%-61.01%)	96.01% (95.12%-96.74%)	94.63% (93.64%-95.47%)
Support vector machine (SVM)	94.23% (93.25%-95.08%)	94.15% (93.17%-95.00%)	99.82% (99.54%-99.93%)	99.64% (99.29%-99.82%)	50.00% (44.20%-55.80%)	50.71% (44.90%-56.50%)	94.05% (93.02%-94.93%)	94.12% (93.10%-95.00%)
Logistic regression (LR)	94.03% (93.04%-94.89%)	94.03% (93.04%-94.89%)	96.82% (96.01%-97.47%)	97.58% (96.86%-98.14%)	71.99% (66.48%-76.90%)	65.96% (60.25%-71.24%)	96.47% (95.63%-97.16%)	95.78% (94.87%-96.53%)
Decision tree (DTree)	93.56% (92.53%-94.45%)	94.00% (93.00%-94.86%)	98.48% (97.88%-98.91%)	98.97% (98.46%-99.31%)	54.61% (48.78%-60.32%)	54.61% (48.78%-60.32%)	94.50% (93.49%-95.35%)	94.52% (93.53%-95.37%)
Naïve Bayes (NB)	86.31% (84.92%-87.60%)	90.49% (89.28%-91.58%)	86.92% (85.45%-88.25%)	91.76% (90.54%-92.83%)	81.56% (76.62%-85.65%)	80.50% (75.47%-84.70%)	97.39% (96.59%-98.00%)	97.38% (96.61%-97.99%)

Table 1. Comparison of performance from seven supervised ML algorithms using all ADI-R items and 27 selected ADI-R items from sPLS-DA.

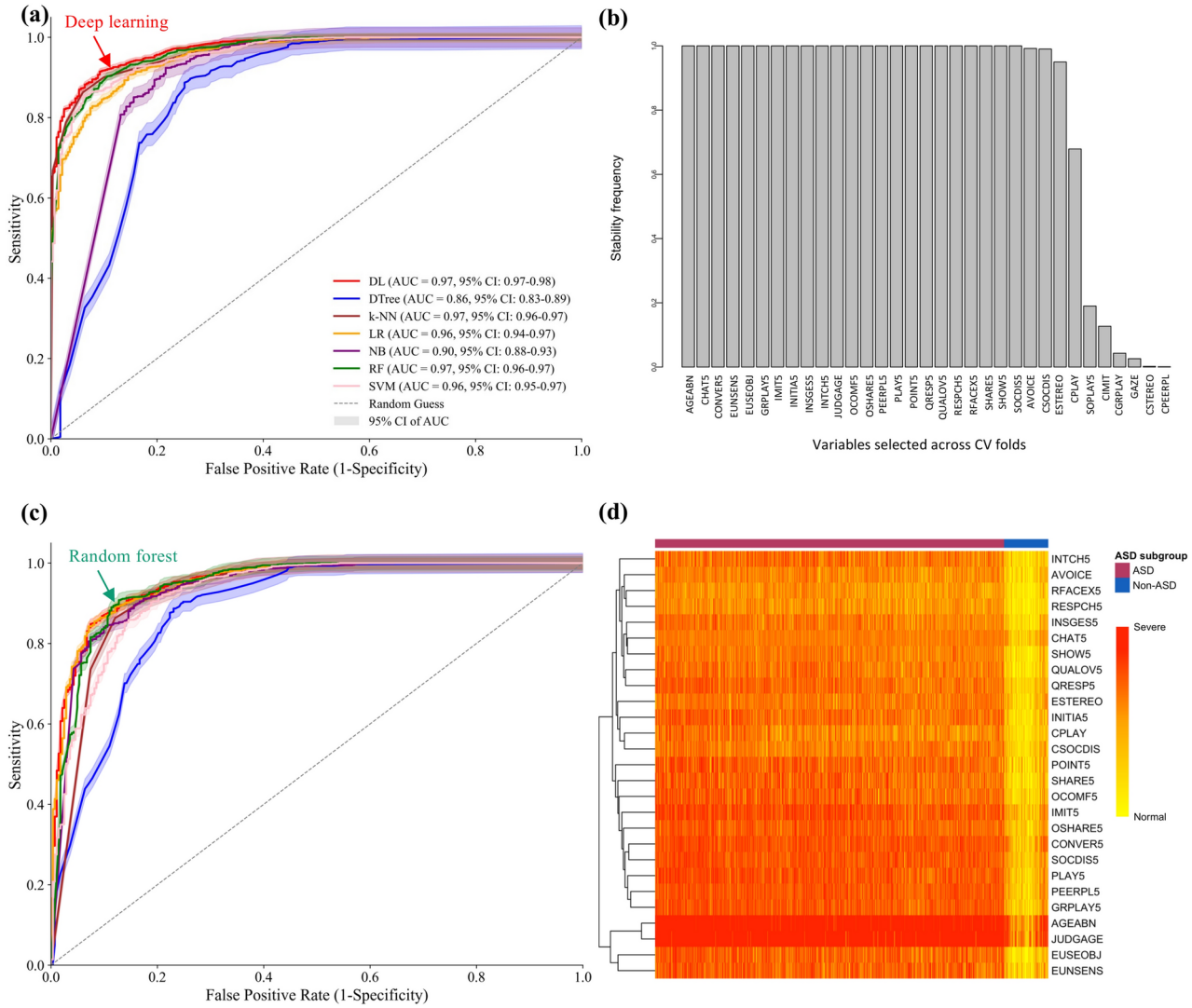

Fig. 3. ML model performance for ASD screening using ADI-R scores (**a**) ROC curve of ML model trained by all ADI-R scores, (**b**) stability of 27 variables selected from the sPLS-DA on component 1, (**c**) ROC curve of ML model trained by 27 selected ADI-R scores (**d**) cluster heatmap of severity symptoms on 27 selected ADI-R scores. *DL* Deep learning, *DTree* Decision tree, *k-NN* k-nearest neighbor, *LR* Logistic regression, *NB* Naïve Bayes, *RF* Random forest, *SVM* Support vector machine, *AUC* Area under ROC curve, *CV* Cross validation.

performance with comparable accuracy (95.19%, CI 94.28–95.96%) and sensitivity (96.91%, CI 96.11–97.55%), while showing improved specificity (81.56%, CI 76.62–85.65%) compared to the imbalanced dataset. Notably, when using the 27 ADI-R items selected by sPLS-DA, RF demonstrated the best overall performance with 91.85% accuracy (CI 90.71–92.85%), 93.06% sensitivity (CI 91.93–94.04%), and improved specificity of 82.27% (CI 77.39–86.28%). The strong and consistent performance of our models with both imbalanced and balanced datasets showed they are reliable and effective screening tools. However, for optimal clinical application, future model development would benefit from training with naturally balanced datasets containing equal numbers of ASD and non-ASD diagnosed through complete ADI-R assessments, rather than relying on synthetic data generation techniques.

Unsupervised ML model, k-means, using ADI-R scores successfully identified three distinct subgroups

ASD presents a broad spectrum of symptoms and severity levels, suggesting that identifying distinct subgroups could enhance personalized intervention approaches and improve clinical outcomes[17,18]. This clinical heterogeneity motivated a subgroup analysis using unsupervised ML approaches on ADI-R clinical symptoms, following the workflow shown in Fig. 4. By applying k-means and k-medoids clustering algorithms to Principal Component (PC) scores of the ASD_samples dataset ($n = 2480$). Both algorithms successfully identified three distinct ASD subgroups, as visualized in the PCA plots (Fig. 5a,b) and PLS-DA prediction areas (Supplementary Data 5). The k-means clustering (Fig. 5b) demonstrated clearer separation between subgroups compared to

ADI-R items	Description	ASD categorical symptoms	Stability frequency
AGEABN	Age when abnormality first evident	General behaviors	1.00
CHAT5	Social vocalization/ "chat"	Language and communication functioning	1.00
CONVER5	Reciprocal conversation (at whatever verbal level of complexity possible)	Language and communication functioning	1.00
EUNSENS	Unusual sensory interests	Interest and behaviors	1.00
EUSEOBJ	Repetitive use of objects or interest in parts of objects	Interest and behaviors	1.00
GRPLAY5	Group play with peers	Social development and play	1.00
IMIT5	Spontaneous imitation of actions	Language and communication functioning	1.00
INITIA5	Initiation of appropriate activities	Social development and play	1.00
INSGES5	Conventional/ instrumental gestures	Language and communication functioning	1.00
INTCH5	Interest in children	Social development and play	1.00
JUDGAGE	Interviewer's judgment on age when developmental abnormalities probably first manifest	General behaviors	1.00
OCOMF5	Offers comfort	Social development and play	1.00
OSHARE5	Offering to share	Social development and play	1.00
PEERPL5	Imaginative play with peers	Language and communication functioning	1.00
PLAY5	Imaginative play	Language and communication functioning	1.00
POINT5	Pointing to express interest	Language and communication functioning	1.00
QRESP5	Appropriateness of social responses	Social development and play	1.00
QUALOV5	Quality of social overtures	Social development and play	1.00
RESPCH5	Response to approaches of other children	Social development and play	1.00
RFACEX5	Range of facial expressions used to communicate	Social development and play	1.00
SHARE5	Seeking to share his/her enjoyment with others	Social development and play	1.00
SHOW5	Showing and directing attention	Social development and play	1.00
SOCDIS5	Social disinhibition	Social development and play	1.00
AVOICE	Attention to voice	Language and communication functioning	0.99
CSOCDIS	Social disinhibition	Social development and play	0.99
ESTEREO	Stereotyped utterances and delayed echolalia	Language and communication functioning	0.95
CPLAY	Imaginative play	Language and communication functioning	0.68

Table 2. Selected ADI-R items from sPLS-DA on component 1.

k-medoids (Fig. 5a), particularly along the first two principal components. Notably, the overall average distance metrics revealed that k-means clustering achieved a lower average distance (9.276) compared to k-medoids (10.553), suggesting more compact and well-defined clusters (Fig. 5c). The boxplot comparison further illustrates the distribution of distances to centroids across the three subgroups, with k-means consistently showing lower within-cluster variation than k-medoids across all subgroups. These findings suggested that k-means clustering provides a more robust and well-defined stratification of ASD_samples based on ADI-R scores, potentially offering a more reliable approach for identifying clinically meaningful ASD subgroups.

The k-means clustering analysis successfully identified three distinct ADI-R subgroups within the ASD population: ADIR_sub1 ($n=867$), ADIR_sub2 ($n=696$), and ADIR_sub3 ($n=917$), as demonstrated by PLS-DA prediction areas (Fig. 6a). Demographic analysis revealed comparable gender distributions across all three subgroups (Fig. 6b), while age distributions showed distinct patterns with mean ages of 9.90, 6.51, and 9.27 years for ADIR_sub1, ADIR_sub2, and ADIR_sub3, respectively (Fig. 6c). Each subgroup exhibited distinct clinical profiles based on ADI-R item severity scores (Fig. 7a). ADIR_sub1 was characterized by milder impairments overall, particularly in social interaction domains. In contrast, ADIR_sub2 demonstrated the most severe manifestations across all categories, with notably pronounced deficits in verbal communication. ADIR_sub3 presented an intermediate phenotype, with particular challenges in reciprocal social interaction and non-verbal communication. Furthermore, to identify the key ADI-R items that distinguish between the ADI-R subgroups, sPLS-DA was employed. The analysis revealed that a subset of 21 ADI-R items was sufficient to discriminate between ADIR_sub1, ADIR_sub2, and ADIR_sub3 (Fig. 7b). These distinct profiles underscored the heterogeneous nature of the ASD population, characterized by specific clinical behavioral presentations.

Significantly differentially expressed genes (DEGs) of ADI-R subgroups

The next step after identifying distinct ADI-R-based ASD subgroups was to investigate their underlying molecular mechanisms and biological signatures. To explore potential subgroup-specific molecular characteristics, transcriptomic analysis was performed using transcriptome profiling of lymphoblastoid cell lines (LCLs) from the Gene Expression Omnibus (GEO) dataset (GSE15402), a public repository for high-throughput gene expression data maintained by the National Center for Biotechnology Information (NCBI). This investigation utilized matched data from 42 ASD individuals whose GSM IDs corresponded to the individual IDs in the ADI-R dataset (Supplementary Data 6), enabling a direct connection between clinical subgroups and their molecular profiles. These 42 matched individuals were distributed across the three ADI-R subgroups: ADIR_sub1 ($n=17$), ADIR_sub2 ($n=2$), and ADIR_sub3 ($n=23$). It is important to note that the small sample size in ADIR_sub2

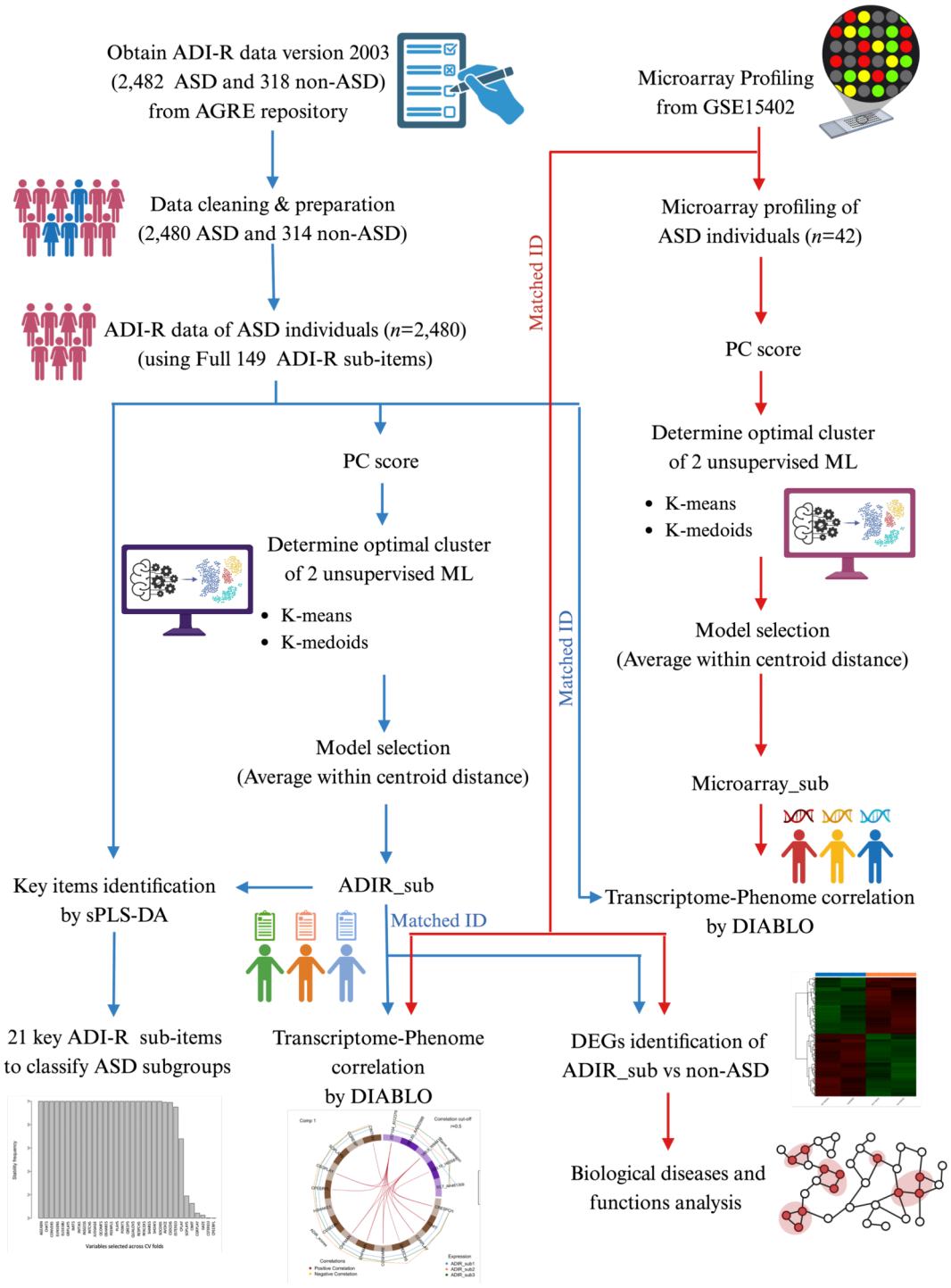

Fig. 4. Machine learning framework for ASD subgrouping using ADI-R or transcriptome profiling. *DEGs* Differentially expressed genes, *sPLS-DA* sparse partial least squares discriminant analysis, *ML* machine learning, *PC scores* principal component scores, *AGRE* autism genetic resource exchange consortium, *DIABLO* Data Integration Analysis for Biomarker Discovery using Latent Variable Approaches for Omics Studies. This Figure was created in BioRender.com.

that matched with the transcriptome profiling data was a limitation in this investigation, potentially affecting the reliability and generalizability of findings for this subgroup.

Differential gene expression analysis was performed using the limma package in R, with a significance threshold of p-value ≤ 0.01, comparing each subgroup against age- and sex-matched non-ASD controls. The analysis identified distinct sets of significantly DEGs: 643 DEGs in the ASD versus non-ASD comparison, 253 DEGs in ADIR_sub1, 221 DEGs in ADIR_sub2, and 426 DEGs in ADIR_sub3. The hierarchical clustering

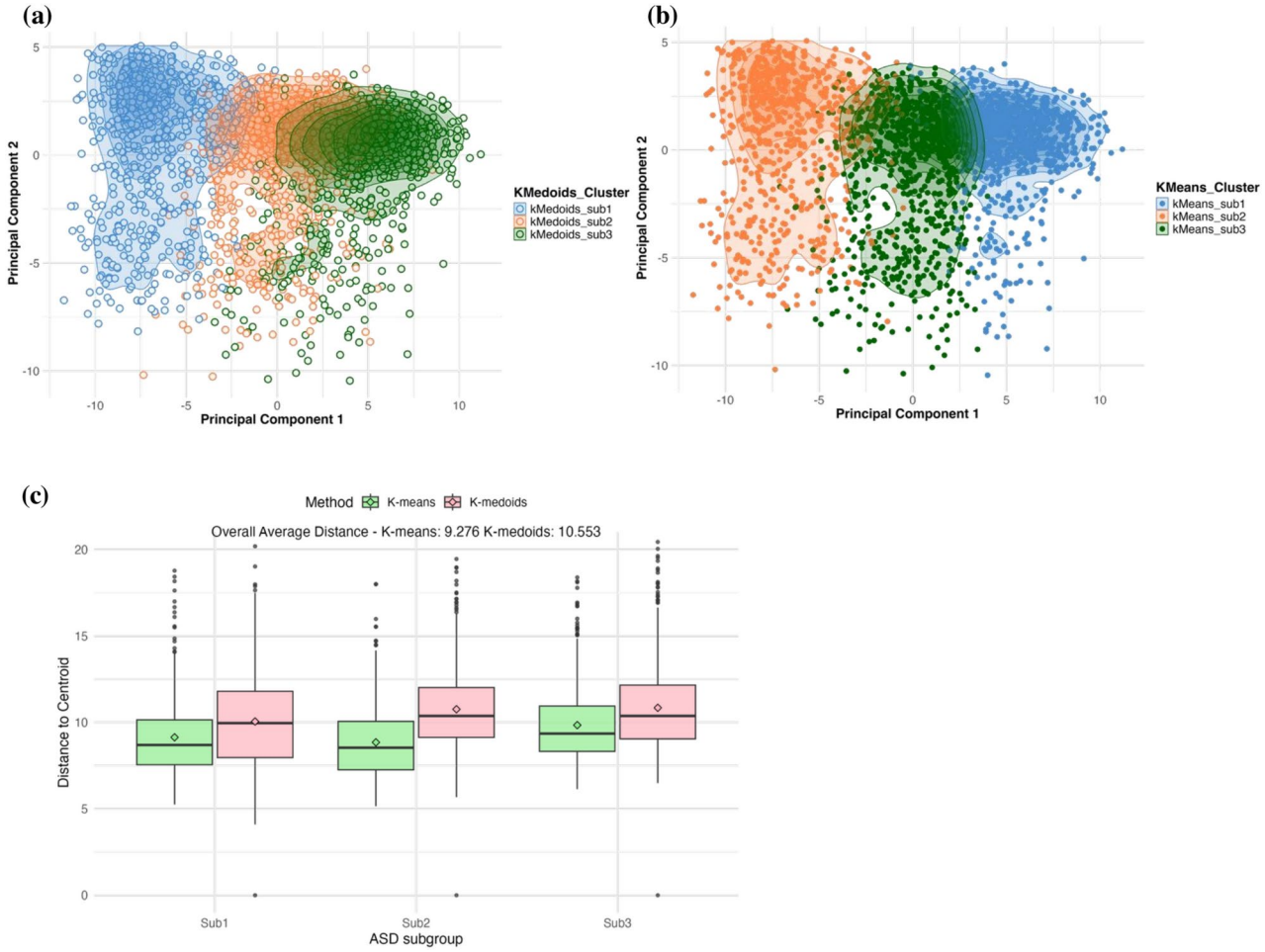

Fig. 5. Clustering of ASD subgroups based on ADI-R scores. (**a**) K-medoids cluster density of three ASD subgroups, (**b**) k-means cluster density of three ASD subgroups, (**c**) average within centroid distance comparison between ASD subgroups by k-means and k-medoids.

heatmaps (Fig. 8a–d) revealed distinct gene expression patterns characteristic of each comparison. Specifically, the heatmaps demonstrated clear separation between ASD and non-ASD samples (Fig. 8a), as well as unique expression signatures for each ADI-R subgroup compared to controls (Fig. 8b–d), suggesting distinct molecular profiles associated with each ADI-R subgroup.

Ingenuity Pathway Analysis (IPA) was performed to investigate the functional roles and associations of identified DEGs with ASD. The results revealed distinct molecular signatures corresponding to the clinical manifestations of each ADI-R subgroup (Table 3). In the overall ASD versus non-ASD comparison, the DEGs were significantly associated with broad neurological conditions, including familial central nervous system disease (p-value = 2.10E−05), progressive neurological disorder (p-value = 1.52E−04), and cognitive impairment (p-value = 3.36E−04). ADIR_sub1, which exhibited milder clinical impairments, showed molecular changes primarily related to neuronal structure and function, including neuritogenesis (p-value = 2.06E−03) and loss of dendritic spines (p-value = 4.62E−03). These molecular findings align with the subgroup's milder behavioral phenotype, particularly in social interaction domains. ADIR_sub2, characterized by the most severe clinical manifestations, demonstrated significant associations with cerebrovascular dysfunction (p-value = 1.85E−03) and broad neurodevelopmental disorders (p-value = 2.81E−03). The network analysis (Fig. 9a) revealed complex interactions among genes involved in multiple neurological pathways, consistent with this subgroup's severe clinical presentation, particularly in verbal communication. Furthermore, the network were associated with various ASD-related behavioral manifestations, including learning deficits, repetitive behaviors (grooming), mood disorders, and emotional disturbances, supporting the comprehensive nature of impairments observed in this subgroup. ADIR_sub3, which presented intermediate severity, showed strong associations with ASD (p-value = 1.90E−04) and mental retardation (p-value = 8.87E−03). The network analysis (Fig. 9b) highlighted interactions between genes involved in nervous system development and dendritic cell maturation, corresponding to this subgroup's notable deficits in reciprocal social interaction and non-verbal communication. These molecular findings provided biological support for the clinical heterogeneity observed among ADI-R subgroups and suggested distinct underlying pathophysiological mechanisms for each subgroup.

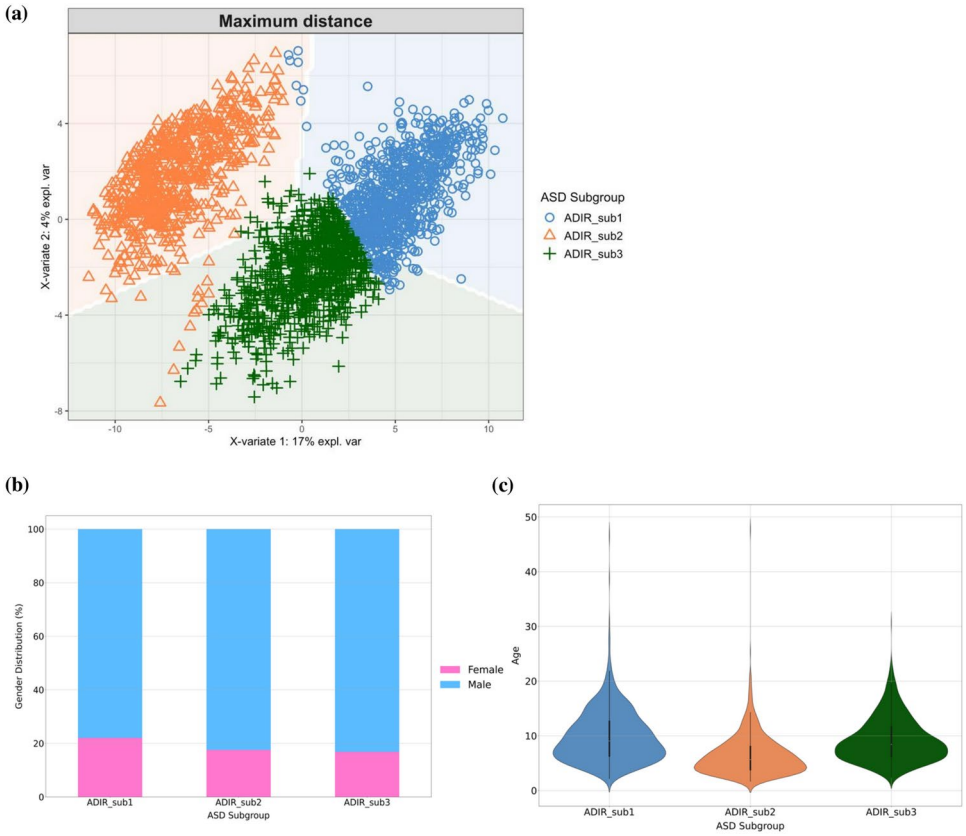

Fig. 6. Demographics of ADI-R subgroup using k-means. (**a**) Scatter plot of ASD individuals from PLS-DA prediction areas based on prediction distances, (**b**) bar chart of the sex distribution across ASD subgroups, (**c**) violin plot of the age distribution within ASD subgroups.

ASD subgroups based on ADI-R scores showed correlation with gene expression

To visualize the distribution of these individuals across the ADI-R subgroups, we performed PLS-DA. This analysis revealed distinct gene expression patterns across ADI-R subgroups, as visualized in Fig. 10a. This plot demonstrates clear separation between the 42 individuals, suggesting that each ADI-R subgroup possesses a unique transcriptomic fingerprint. The well-defined prediction areas for each subgroup provide compelling evidence that individuals with different ADI-R classifications exhibit fundamentally different patterns of gene expression. These distinct molecular signatures suggest that ADI-R subgroups may represent biologically meaningful subdivisions within ASD, each characterized by specific patterns of gene activity.

To further investigate the relationships between clinical symptoms and gene expression within each ADI-R subgroup, we employed DIABLO (Data Integration Analysis for Biomarker Discovery using Latent Variable Approaches for Omics Studies), also known as multiblock sPLS-DA. DIABLO extends sparse Generalised Canonical Correlation Analysis, facilitating the identification of covariances between pairs of omics data[27]. The DIABLO analysis of component 1 revealed a strong correlation (correlation coefficient = 0.8) between gene expression patterns and ADI-R clinical scores (Fig. 10b), identifying 5 key genes and 14 ADI-R items that best explained the relationship between molecular and clinical features. This strong correlation suggested an interplay between the clinical manifestations of ASD and the underlying gene expression patterns across the ADI-R subgroups.

To further elucidate the specific relationships between gene expression and clinical symptoms, the Circos plot (Fig. 10c) visualized correlations with a cut-off threshold of r = 0.5 (Supplementary Data 7). Three genes, GenBank number AI123790, GenBank number W94419 (gene symbol *DKFZp586H0623*), and GenBank number H60581 (gene symbol *BACE1*), showed higher expression in ADIR_sub2 and ADIR_sub3 but lower expression in ADIR_sub1, with positive correlations to social interaction symptoms. The gene AI123790 exhibited strong positive correlations with response to approaches of other children (CRESPCH), hand and finger mannerisms (CHFMAN), and group play with peers (CGRPLAY), while showing moderate positive correlations with spontaneous imitation of actions (CIMIT), imitative social play (CSOPLAY), offering to share (COSHARE), imaginative play with peers (CPEERPL), and current communicative speech (SPEECH5). Similarly, W94419 demonstrated moderate positive correlations with multiple social interaction symptoms, including CRESPCH, CHFMAN, CIMIT, CGRPLAY, EHFMAN, CSOPLAY, and COSHARE. H60581, which was differentially expressed in ADIR_sub3, showed a high correlation specifically with CRESPCH. The hierarchical clustering analysis (Fig. 10d) further supported these findings, revealing distinct patterns of gene-symptom relationships across ADI-R subgroups. These results suggested that differentially gene expression may

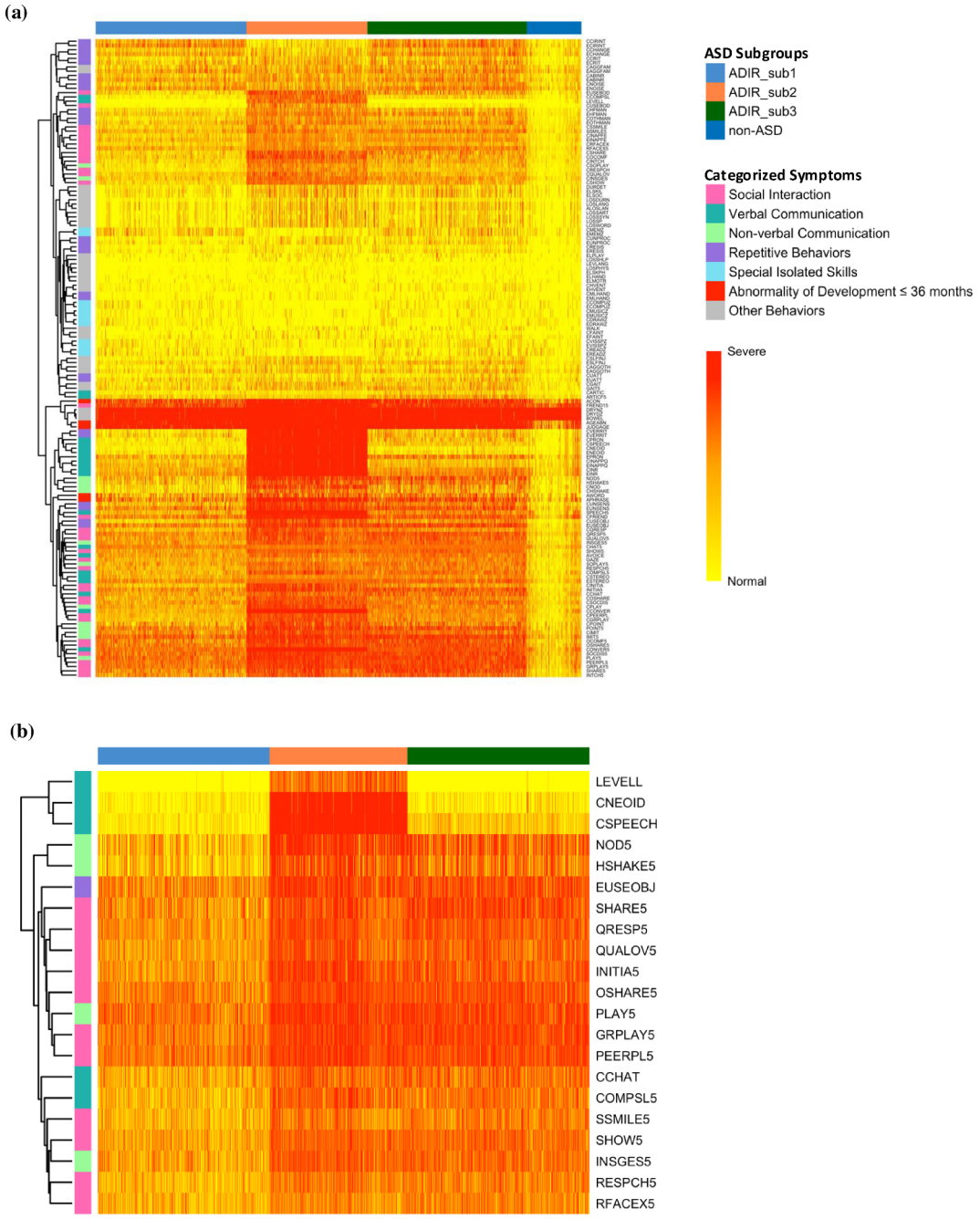

Fig. 7. Cluster heatmap of severity scores across ADI-R items. (**a**) Cluster heatmap of severity scores for all ADI-R items, (**b**) cluster heatmap of severity scores for 21 sPLS-DA selected ADI-R items.

underlie the heterogeneous clinical presentations observed in ASD individuals, particularly regarding social interaction deficits in ADIR_sub2 and ADIR_sub3.

ASD subgrouping based on transcriptome profiles revealed three distinct subgroups

To investigate the differences between ASD subgroups based on clinical symptoms versus gene expression, and to determine which ASD subgrouping method demonstrates a stronger correlation between clinical symptoms and gene expression, we conducted a subgroup analysis based on transcriptome profiling data. This approach aimed to compare the results with our previous ADI-R subgrouping and to explore potential biological underpinnings of ASD heterogeneity. We applied k-means and k-medoids clustering to the transcriptome profiling data of 85 ASD individuals from GSE15402, following the same procedural steps used in the ADI-R score-based subgrouping. This analysis revealed three distinct ASD subgroups through both k-means and k-medoids clustering approaches (Supplementary Data 5). The PCA visualization demonstrated clear separation between these subgroups in two-dimensional space (Fig. 11a,b). The k-medoids clustering showed well-defined

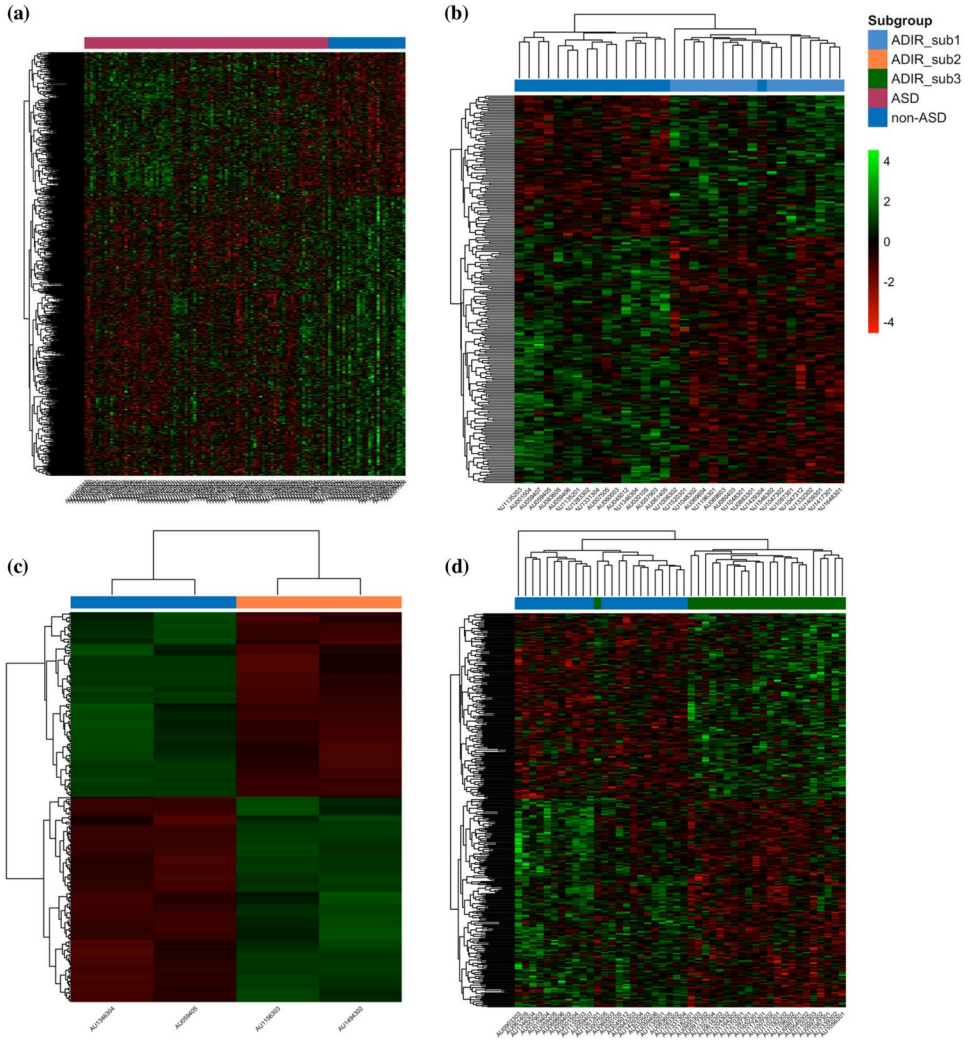

Fig. 8. Hierarchical clustering heatmaps of DEGs. (**a**) Hierarchical clustering heatmaps of DEGs in ASD versus non-ASD, (**b**) hierarchical clustering heatmaps of DEGs in ADIR_sub1, (**c**) hierarchical clustering heatmaps of DEGs in ADIR_sub2, (**d**) hierarchical clustering heatmaps of DEGs in ADIR_sub3.

boundaries between the three subgroups, with distinct density peaks for each cluster (Fig. 11a). Similarly, the k-means clustering exhibited comparable subgroup separation, though with slightly different cluster boundaries and density distributions (Fig. 11b). The overall average distance metrics indicated that k-means clustering (125.658) performed marginally better than k-medoids (136.797) in terms of cluster cohesion. Analysis of the average within-centroid distances across subgroups (Fig. 11c) revealed that both clustering methods successfully delineated the ASD cohort into three distinct subgroups with different transcriptome profiles. The superior performance of k-means clustering, as evidenced by its lower average within-centroid distance, suggested it may be the preferred method for ASD subgroup identification using gene expression data.

The k-means clustering revealed three distinct subgroups: Microarray_sub1 ($n=42$), Microarray_sub2 ($n=24$), and Microarray_sub3 ($n=21$). The separation of these subgroups was visually represented in Fig. 11d, which showed the PLS-DA prediction areas. To investigate potential confounding factors, we examined the age distribution across the three transcriptome-based subgroups. The mean ages for Microarray_sub1, Microarray_sub2, and Microarray_sub3 were 11.80, 12.88, and 12.76 years, respectively. As illustrated in Fig. 11e, the similar mean ages across subgroups suggested that age did not significantly influence the transcriptome-based subgrouping. This finding indicated that the observed differences in gene expression profiles were likely related to the underlying biological heterogeneity of ASD rather than age-related factors.

Transcriptome-based ASD subgroups showed lower correlation between gene expression and clinical symptoms

To further investigate the relationship between gene expression and clinical symptoms within each transcriptome-based subgroup, we analyzed the ADI-R dataset for 42 individuals with matching IDs in the transcriptome data. The distribution of these individuals across the transcriptome-based subgroups was visualized using a PLS-DA prediction plot (Fig. 12a), which clearly showed the separation of the three subgroups. However, when we

Diseases or functions annotation	p-value	# Genes
ASD vs non-ASD		
Familial central nervous system disease	2.10E−05	57
Progressive neurological disorder	1.52E−04	57
Seizure disorder	2.71E−04	25
Cognitive impairment	3.36E−04	34
Familial mental retardation	8.62E−03	16
ADIR_sub1 vs non-ASD		
Neuritogenesis of pheochromocytoma cell lines	2.06E−03	2
Neurodevelopmental disorder with hypotonia and cerebellar atrophy without seizures	3.76E−03	1
Loss of dendritic spines	4.62E−03	2
Working memory	6.85E−03	2
Movement Disorders	7.71E−03	19
ADIR_sub2 vs non-ASD		
Cerebrovascular dysfunction	1.85E−03	14
Neurodevelopmental disorder	2.81E−03	14
Congenital neurological disorder	3.04E−03	18
Cognitive impairment	1.32E−02	17
Developmental delay and hypotonia	1.42E−02	2
ADIR_sub3 vs non-ASD		
Autism spectrum disorder or intellectual disability	1.90E−04	28
Familial mental retardation	4.61E−03	14
Pervasive developmental disorder	5.10E−03	13
Dysmyelination	7.65E−03	15
Mental retardation	8.87E−03	18

Table 3. Significant diseases and functions of DEGs associated with ASD.

examined the average severity scores across ADI-R items for each ASD symptom category using a heatmap cluster plot (Supplementary Data 5), we found that the severity of symptoms in each ASD symptom category did not differ substantially among the transcriptome-based subgroups, although Microarray_sub3 showed a trend towards higher severity compared to the other groups. This intriguing result suggested that while gene expression profiles can distinguish clear subgroups within ASD, these subgroups may not directly correspond to dramatic differences in symptom severity as measured by ADI-R. Subsequently, we investigated the correlation between gene expression and clinical symptoms of transcriptome-based subgroups using DIABLO analysis. The DIABLO analysis at component 1 (Fig. 12b), which identified 30 key genes and 14 ADI-R items that best explained the relationship between molecular and clinical features, revealed a correlation coefficient of 0.61 between gene expression and ADI-R symptoms. This correlation was lower than that observed in the DIABLO component 1 analysis of ADI-R subgroups. To visualize these relationships in more details, we generated a Circos plot on component 1 and 2 (Fig. 12c) with a correlation cut-off of 0.5, which illustrated the gene expression patterns in each transcriptome-based subgroup and their positive and negative correlations with symptoms (Supplementary Data 8). Notably, when using the same correlation cut-off of 0.5 as applied in the ADI-R subgroup analysis, we observed a greater number of genes linked to symptoms in the transcriptome-based subgroups compared to the ADI-R-based subgroups. For example, the nodding symptom (NOD5), which showed severe abnormality in Microarray_sub3, moderate severity in Microarray_sub2, and mild severity in Microarray_sub1, demonstrated strong positive correlations ($r > 0.7$) with GenBank number AI359037 (gene symbol *FABP5*) and GenBank number H07920 (gene symbol *MAP2K6*), while exhibiting strong negative correlations ($r < -0.7$) with GenBank number AA497040 (gene symbol *STC2*), GenBank number AA015892 (gene symbol *DDIT3*), and GenBank number H56147 (gene symbol *FLJ22060*). Furthermore, Fig. 12d displayed the patterns of gene expression and clinical symptoms from ADI-R that are top covariants from component 1 and 2. This visualization revealed distinct patterns for each subgroup, particularly in gene expression profiles. These findings collectively indicated that while transcriptome-based subgroups showed clear distinctions in gene expression patterns, these molecular differences might translate to more subtle variations in clinical presentation. This underscored the complex relationship between genetic factors and observable symptoms in ASD, highlighting the potential for gene expression profiles to reveal biological heterogeneity that may not be immediately apparent in clinical assessments.

Evaluation of ADI-R-based and transcriptome-based ASD subgrouping approaches

To comprehensively understand the strengths and limitations of our subgrouping approaches, we conducted a comparative analysis of the ADI-R-based and transcriptome-based subgrouping methods. Both methods successfully identified three distinct subgroups within our ASD cohort, each offering unique insights into ASD heterogeneity. The ADI-R-based subgroups demonstrated clear distinctions in symptom profiles across the

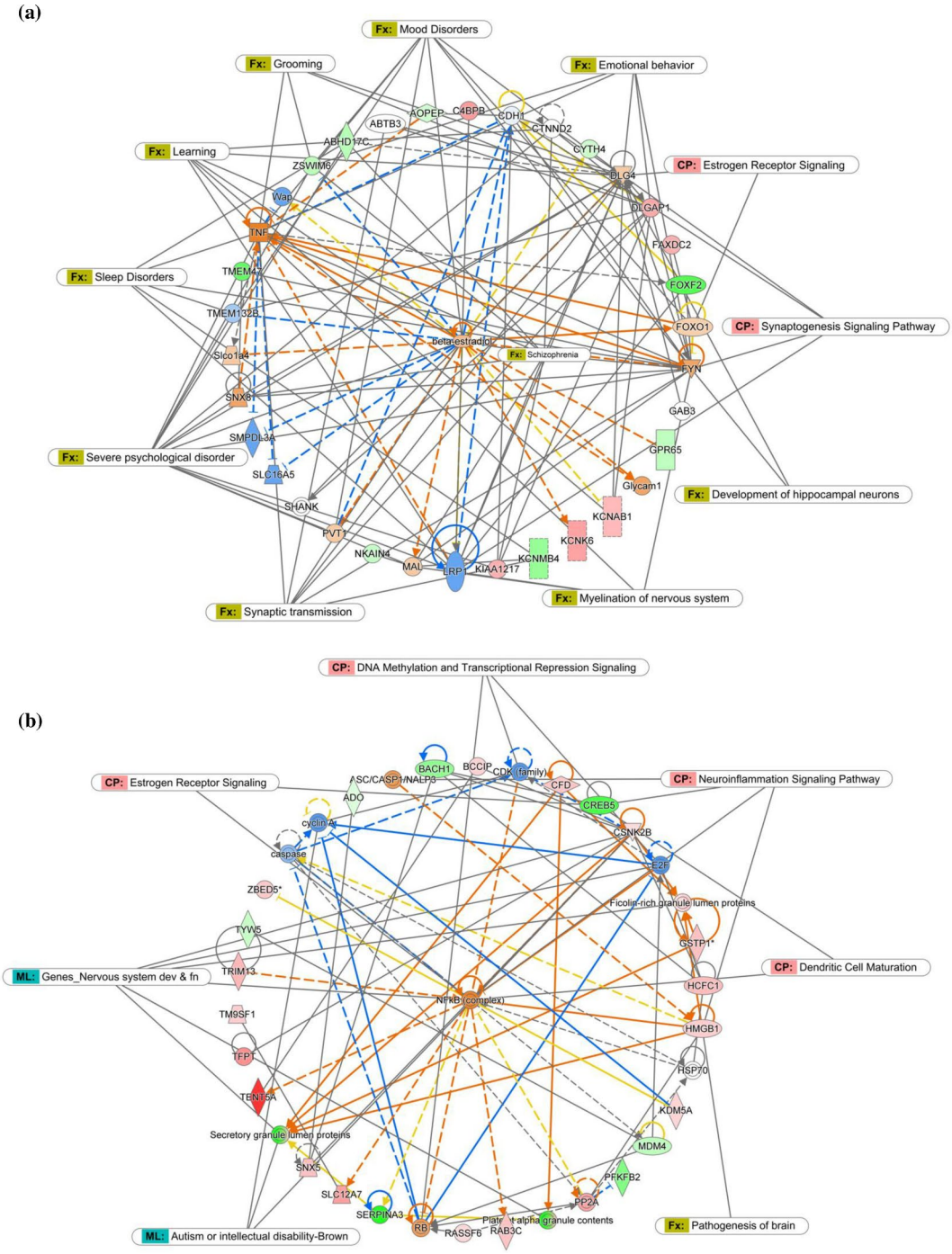

Fig. 9. Network analysis of DEGs. (**a**) Network analysis of DEGs of ADIR_sub2, (**b**) network analysis of DEGs of ADIR_sub3.

three main ASD categories (Fig. 7a,b): social interaction deficits, communication impairments, and repetitive behaviors. Each subgroup exhibited distinct patterns of severity within these core domains, providing a clinically intuitive categorization of ASD heterogeneity. In contrast, the transcriptome-based subgroups showed less pronounced differences in symptom severity across ASD categories. The relative consistency of abnormalities across ASD symptom domains in these subgroups may be attributed to molecular signatures that relate to multiple interconnected abnormalities, making it challenging to distinctly separate severity according to specific ASD categories. Despite these differences, both methods demonstrated high correlations between clinical symptoms and gene expression profiles. The ADI-R-based subgroups showed a strong correlation coefficient of 0.80 between gene expression and ADI-R symptoms, while the transcriptome-based subgroups showed a lower correlation coefficient of 0.61. The high correlation in ADI-R-based subgroups indicated a stronger link between molecular patterns and clinical presentation than transcriptome-based subgroups, underscoring their biological

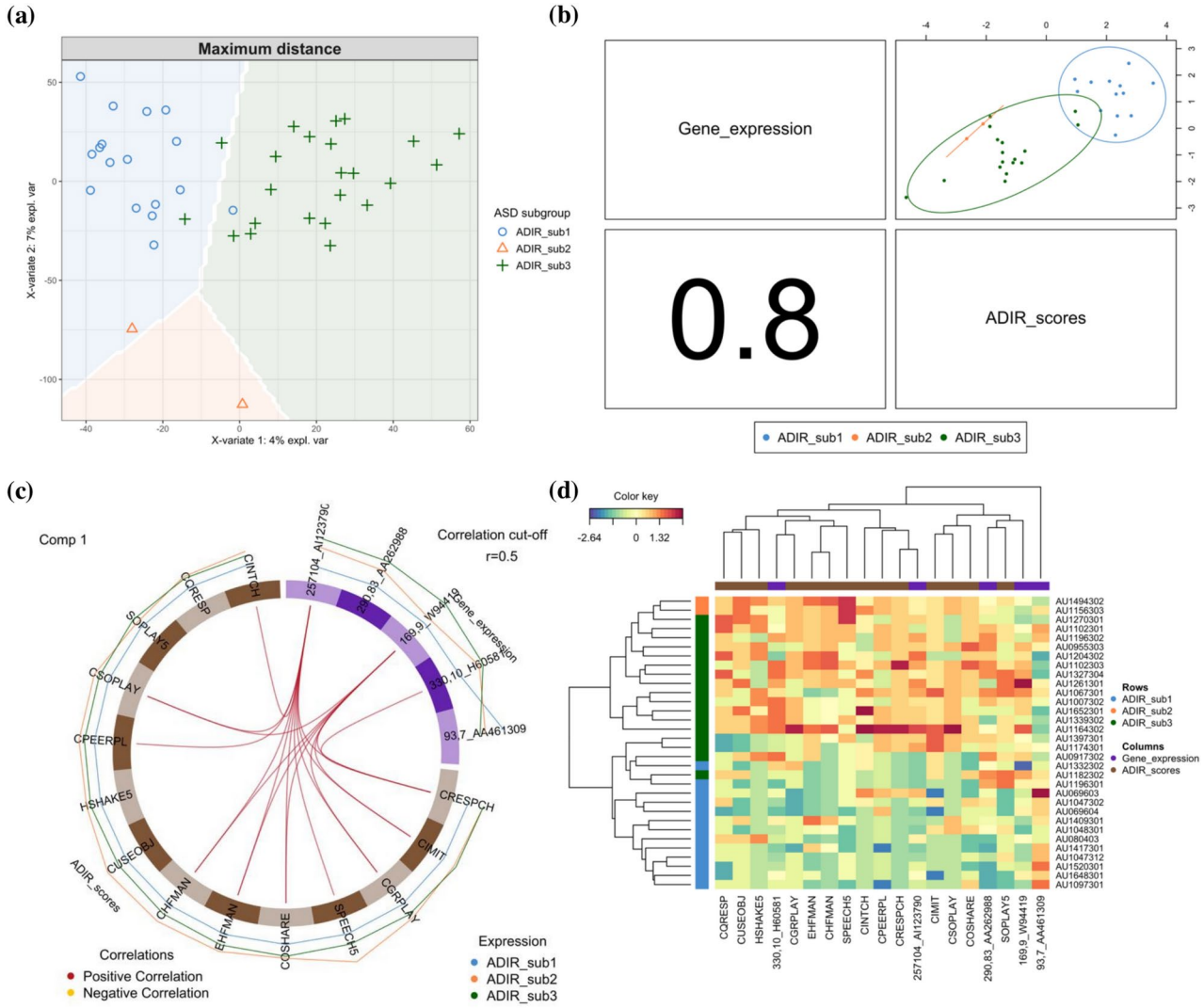

Fig. 10. Integrative analysis of gene expression and clinical symptoms in ADI-R subgroups. (**a**) PLS-DA prediction areas of transcriptome profiling across ADI-R subgroups, (**b**) DIABLO plot of clinical symptoms (ADI-R) and gene expression on component 1, (**c**) Circos plot of clinical symptoms (ADI-R) and gene expression on component 1, (**d**) Clustered Image Map showing the expression of severity symptoms and gene expression on component 1.

relevance. The ADI-R-based method offers immediate clinical relevance, potentially informing personalized intervention strategies. It benefits from the use of widely available clinical data but may be influenced by the subjective nature of clinical assessments. In contrast, the transcriptome-based approach provides valuable insights for future biomarker development and targeted therapeutic approaches. It offers an objective, molecular-level analysis but requires specialized equipment. In conclusion, both subgrouping approaches offer valuable and complementary insights into ASD heterogeneity. The ADI-R-based method provides a clinically intuitive categorization, while the transcriptome-based approach uncovers underlying biological differences that may not be immediately apparent in clinical presentations. Integrating these approaches could lead to a more comprehensive understanding of ASD subtypes, potentially guiding both clinical practice and future research directions.

Discussion

In this study, we investigated both ASD screening using ADI-R, and ASD subgrouping through ADI-R and gene expression profiling data. For the screening aspect, our DL model applied to ADI-R scores achieved remarkable performance with 95.23% accuracy (CI 94.32–95.99%), 97.94% sensitivity, and 73.76% specificity using a comprehensive dataset of 2794 individuals. Several previous studies have investigated ML approaches for ASD screening using various diagnostic instruments. In ADI-R-based studies, Wall et al. examined a dataset of 966 individuals (891 ASD, 75 non-ASD), identifying seven key items that achieved 99.9% accuracy[22]. Bone et al. integrated ADI-R with SRS using SVM on 1726 individuals (1264 ASD, 462 non-ASD), achieving moderate performance (sensitivity: 86.7–89.2%, specificity: 53.4–59.0%) with five behavioral codes[24]. Studies utilizing

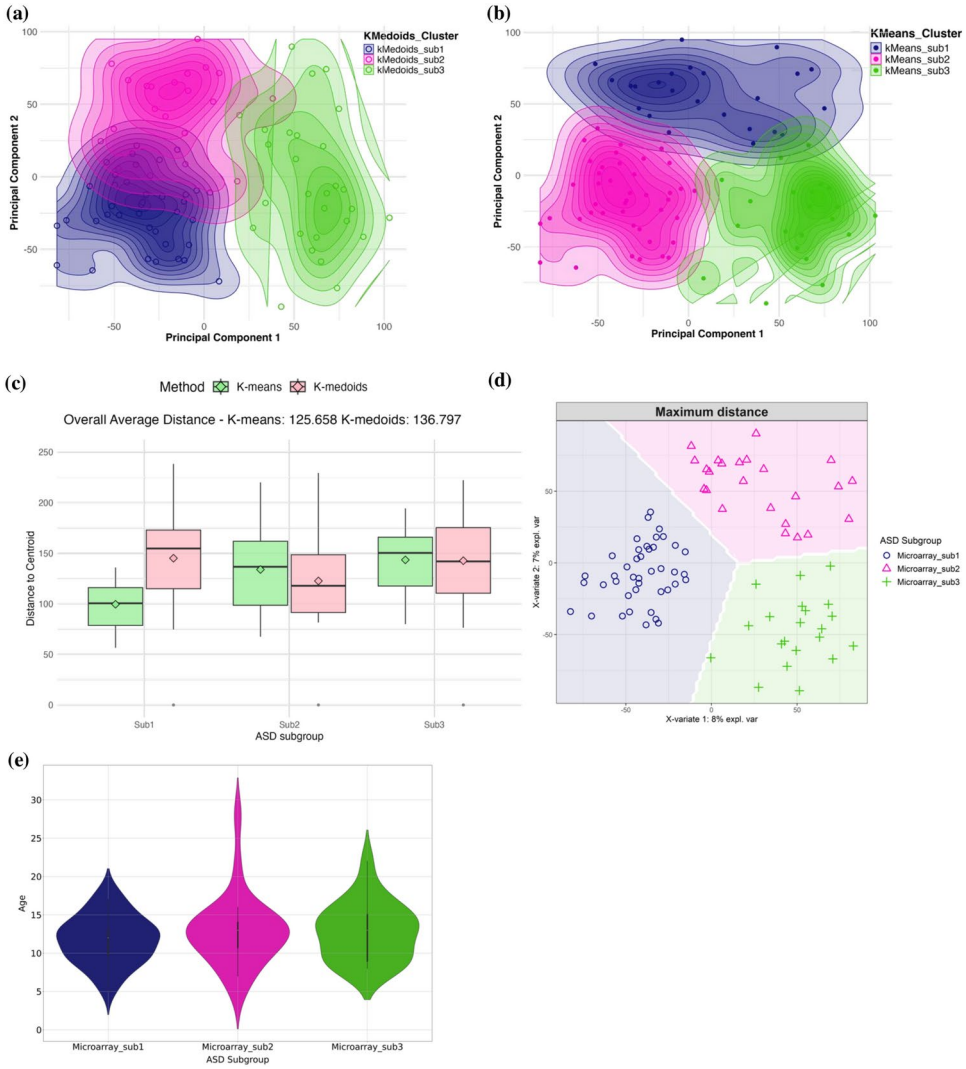

Fig. 11. Clustering of ASD subgroups based on transcriptome profiling data. (**a**) K-medoids cluster density of three ASD subgroups, (**b**) k-means cluster density of three ASD subgroups, (**c**) average within centroid distance comparison between ASD subgroups by k-means and k-medoids, (**d**) PLS-DA prediction plot showing k-means subgroup classification based on microarray data, (**e**) age distribution across transcriptome-based subgroups using k-means clustering.

ADOS showed a similar pattern. Wall et al. analyzed a limited dataset of 627 individuals (612 ASD, 15 non-ASD) with Module 1, reporting high accuracy using eight crucial items[23]. Subsequently, Duda et al. analyzed a limited dataset of 627 individuals (612 ASD, 15 non-ASD) with Module 1, reporting high accuracy using eight crucial items[28]. Notably, this progression from smaller to larger sample sizes across both ADI-R and ADOS studies revealed a consistent pattern: while smaller studies often report exceptionally high accuracies, larger sample sizes tend to yield more realistic and generalizable performance metrics. Our study, representing one of the largest ADI-R-based screening investigations to date, produced performance metrics that align more closely with Duda et al.'s large-scale ADOS study rather than Wall et al.'s smaller ADI-R study, underscoring the importance of large, diverse datasets in developing reliable and clinically applicable screening tools. A critical consideration in our study was addressing the inherent dataset imbalance (2232 ASD vs. 282 non-ASD), which initially resulted in high accuracy but moderate specificity in our DL model. This imbalance is a common challenge in ASD screening studies, potentially leading to models that excel at identifying ASD but perform less optimally with non-ASD. By applying SMOTE to create a balanced training dataset, we significantly improved model performance, particularly in specificity. The DL model maintained its strong overall performance (accuracy: 95.19%, CI 94.28–95.96%; sensitivity: 96.91%, CI 96.11–97.55%) while achieving notably improved specificity (81.56%, CI 76.62–85.65%). These balanced results align more closely with clinical requirements for reliable screening tools.

Additionally, we successfully reduced the ADI-R assessment from 93 questionaires (149 sub-items) to 27 sub-items using sPLS-DA, a powerful feature selection method that applies a lasso penalty to the loading vector to systematically identify the most discriminative items while eliminating less informative ones. The

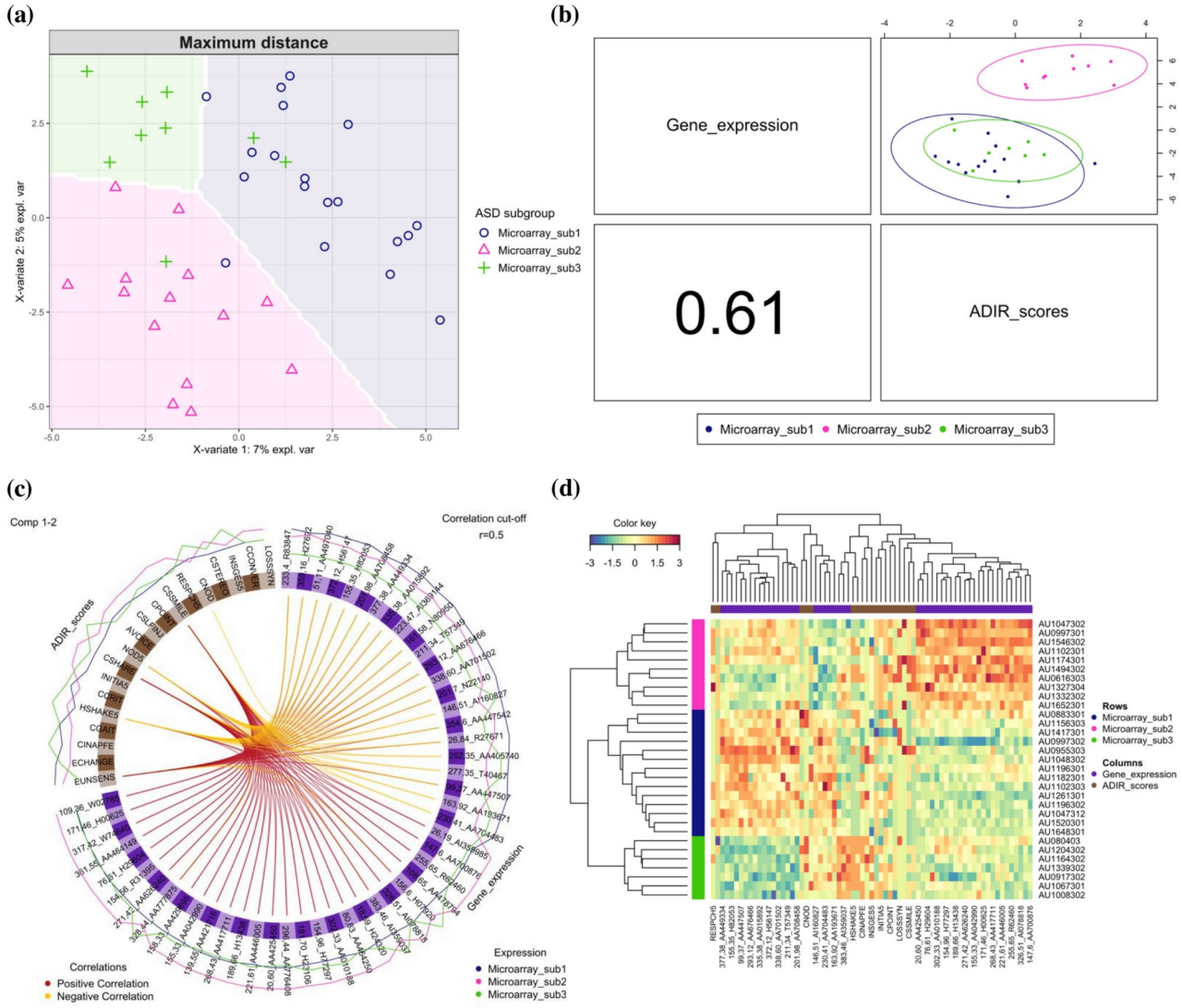

Fig. 12. Integrative analysis of gene expression and clinical symptoms according to transcriptome-based ASD subgroups by k-means clustering. (**a**) PLS-DA prediction plot of transcriptome-based subgroups on ADI-R data, (**b**) DIABLO plot linking ADI-R scores to gene expression across transcriptome-based subgroups, (**c**) Circos plot integrating ADI-R scores and gene expression of transcriptome-based subgroups, (**d**) Clustered heatmap correlating symptom severity with gene expression across transcriptome-based subgroups.

robustness of our feature selection approach was demonstrated by the high stability of the selected items, with 23 out of 27 items showing 100% selection consistency in cross-validation. Correlation analysis of these selected items revealed minimal redundancy, with all correlations below 0.7 and the highest correlations observed between CSOCDIS-SOCDIS5 (r = 0.68) and AGEABN-JUDGAGE (r = 0.66). This low correlation pattern, combined with the high stability of selected items, validates sPLS-DA's effectiveness in identifying a concise set of independent diagnostic indicators for ASD screening. The effectiveness of sPLS-DA in capturing fundamental diagnostic indicators was evidenced by the versatility of our reduced item set, which maintained high performance across multiple ML algorithms with minimal impact on accuracy (1–2% difference from full ADI-R). This consistency across algorithms suggested that sPLS-DA successfully identified items that represent core diagnostic features of ASD rather than algorithm-specific patterns. Among the supervised algorithms, RF emerged as the superior performer with our 27-item set, demonstrating enhanced accuracy and achieving a notable 5% increase in specificity compared to its performance with the full item set, as confirmed through ROC curve analysis. When applied to the balanced dataset, these 27 items maintained strong performance with RF, achieving 91.85% accuracy (CI 90.71–92.85%), 93.06% sensitivity (CI 91.93–94.04%), and improved specificity of 82.27% (CI 77.39–86.28%). This finding suggested that the combination of our sPLS-DA-selected items with RF classification provided an optimal approach for efficient and accurate ASD screening while maintaining diagnostic comprehensiveness. Although our models demonstrated strong performance with both imbalanced and balanced datasets, future development of screening tools would benefit from naturally balanced datasets containing equal numbers of ASD and non-ASD diagnosed through complete ADI-R assessments, rather than

relying on synthetic data generation techniques. This approach would further enhance the clinical applicability and reliability of ML-based screening tools.

The phenotypic heterogeneity of ASD presented significant challenges for both clinical management and research. Our study employed ML approaches to analyze ADI-R scores from a large cohort of 2480 individuals with ASD, revealing three distinct subgroups with specific behavioral profiles. These findings contributed to the evidence supporting the existence of clinically meaningful ASD subgroups and offered potential pathways for more targeted therapeutic approaches. Our findings demonstrated that unsupervised ML using k-means clustering successfully identified distinct ASD subgroups based on ADI-R scores. The clear separation observed in both PCA plots and PLS-DA prediction areas suggested that these subgroups represented genuinely distinct phenotypic patterns rather than arbitrary divisions of a continuous spectrum. The identification of three distinct subgroups (ADIR_sub1, ADIR_sub2, and ADIR_sub3) differed from previous findings by Hu et al. who identified four subgroups using a smaller sample size ($n = 1954$)[17]. Our analysis revealed patterns of stratification based on overall symptom severity and specific behavioral challenges. ADIR_sub1 demonstrated milder impairments across domains, particularly in social interaction. ADIR_sub2 exhibited the most severe manifestations, with pronounced deficits in verbal communication. ADIR_sub3 presented an intermediate phenotype with particular challenges in reciprocal social interaction and non-verbal communication. When comparing our subgroups with Hu's study, we observed distinct differences in subgroup characterization and distribution. While Hu's study identified specific subgroups including a savant skills group, our clustering approach revealed a different pattern based primarily on symptom severity. Analysis of the sample overlap between studies showed that Hu's language-impaired group distributed across our ADIR_sub2 (severe) and ADIR_sub3 (intermediate) categories, while their mild groups showed considerable overlap with our ADIR_sub1 (mild) category. These differences in subgroup identification likely stemmed from several factors: our larger sample size (2480 versus 1954), the version of ADI-R questionaires, and different methodological approaches to clustering. These methodological differences highlighted the complexity of ASD subtyping and suggested that different analytical approaches could reveal different aspects of ASD heterogeneity.

The integration of gene expression data with our ADI-R-based subgroups provided valuable insights into the biological underpinnings of these clinically defined phenotypes. Through DIABLO analysis, we identified strong correlations (correlation coefficient = 0.80) between gene expression patterns and clinical symptoms, particularly in social interaction domains. Three key genes—GenBank number AI123790, GenBank number W94419 (gene symbol *DKFZp586H0623*), and GenBank number H60581 (gene symbol *BACE1*)—showed differential expression patterns that aligned with the severity of social interaction deficits in ADIR_sub2 and ADIR_sub3. The molecular profiles of each subgroup had corresponded well with their clinical presentations: ADIR_sub2's severe phenotype was reflected in molecular changes associated with cerebrovascular dysfunction and broad neurodevelopmental disorders, while ADIR_sub1's milder presentation showed changes primarily in neuronal structure and function. However, it is important to acknowledge significant limitations in our gene expression analysis, particularly regarding sample size disparities across subgroups. After matching ADI-R-based subgroups with available transcriptome data, subgroup 2 included only two individuals, which substantially limited the statistical power and generalizability of findings for this subgroup. While our DIABLO analysis provided preliminary insights into potential molecular signatures associated with clinical phenotypes, these results should be interpreted with caution due to this sample size limitation. The minimal explained variance in the gene expression clustering, especially for ADIR_sub2, suggests that larger, more balanced samples are needed to validate these molecular associations. Despite these limitations, our findings provide a valuable framework for future investigations into the biological basis of ASD subgroups and highlight the importance of considering sample size requirements when designing integrated clinical-molecular studies.

To investigate whether a molecular-first approach might better capture ASD heterogeneity, we performed k-means clustering analysis on transcriptome profiling data from GSE15402. This analysis revealed three distinct subgroups (Microarray_sub1, Microarray_sub2, and Microarray_sub3) with clear separation in both PCA visualization and PLS-DA prediction areas, suggesting that these molecular subgroups represented genuinely distinct biological entities. To further explore this complexity and understand which analytical approach might be more effective for capturing the biological basis of ASD heterogeneity, we compared our ADI-R-based and transcriptome-based subgrouping methods. Our comparative analysis of these two approaches revealed striking differences in their ability to integrate molecular and clinical features. While both methods successfully identified distinct subgroups, the ADI-R-based approach demonstrated superior performance in correlating molecular and clinical features, as evidenced by a stronger correlation coefficient (0.80) in the DIABLO analysis compared to the transcriptome-based approach (0.61). This finding suggested that beginning with clinical phenotypes might better guide the identification of biologically meaningful subgroups in ASD. The transcriptome-based subgroups, although molecularly distinct, showed less pronounced differences in symptom severity across ADI-R categories, with only Microarray_sub3 displaying a slight trend toward higher severity. This observation indicated that molecular differences alone might not directly translate to distinct clinical phenotypes, highlighting the complexity of the relationship between gene expression patterns and observable symptoms in ASD. Interestingly, while the transcriptome-based subgroups showed a greater number of gene-symptom correlations at the 0.5 threshold, these correlations appeared to be less functionally relevant compared to those identified in the ADI-R-based approach. For example, specific symptoms like nodding (NOD5) demonstrated correlations with several genes (*FABP5*, *MAP2K6*, *STC2*, *DDIT3*, and *FLJ22060*) across transcriptome-based subgroups, but these correlations did not align with clear patterns of broader symptom presentation. This finding suggested that while gene expression profiling could reveal molecular subtypes within ASD, these subtypes did not necessarily correspond to clinically meaningful differences in symptom presentation and severity.

Based on these observations, we recommend a stratification strategy that primarily utilizes clinical measures such as ADI-R for initial subgroup identification, followed by molecular profiling as a secondary characterization

tool. This approach would better capture clinically relevant heterogeneity while still providing insights into the biological underpinnings of different ASD presentations. Future research should focus on developing more sophisticated integrated approaches that can effectively combine clinical and molecular data, potentially incorporating ML techniques that can optimally weight different data types. Additionally, larger sample sizes for molecular studies and standardized protocols for data integration will be crucial for validating and refining these subgrouping strategies. Such integrated approaches could ultimately lead to more personalized therapeutic strategies that address both the clinical presentations and underlying biological mechanisms of ASD.

Methods

Autism diagnostic interview-revised (ADI-R) data collection

This research study was conducted under ethical approval from the research ethics review committee for research involving human research participants, group 1, Chulalongkorn University (exception COA NO. 109/2021 for ADI-R data), and the Institutional Review Board of the Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand (COA No.1238/2022). Access to ADI (Autism Diagnostic Interview) scores was granted by the Autism Genetic Resource Exchange (AGRE) Consortium under AGRE Data Access Application (DACO-2021-15). The clinical data obtained from AGRE included ADI-R assessment scores along with information about family configuration, age at time of testing, sex, psychopathology, diagnosis, cognitive functioning, family and medical history, and other clinically relevant information. All the participants are informed about the study, and they have all signed the informed consent form. In accordance with AGRE's data protection policies, all personally identifying information about the subjects and their family members was excluded from the dataset. This manuscript does not contain any information or images that could lead to identification of study participants. The AGRE dataset comprised ADI scores from 3926 individuals, including both the 1995 version (ADI) and the 2003 revised version (ADI-R). For this study, we analyzed data from 2800 individuals who were assessed using the 2003 revised version, consisting of 2482 individuals with ASD and 318 without ASD. All methods were performed in accordance with the relevant guidelines and regulations.

Transcriptome profiling data collection

Gene expression data for this study was obtained from transcriptome profiles published by Hu et al.[25]. The dataset comprised 116 individuals, including both ASD and non-ASD participants. Of these, 112 individuals had corresponding data in the AGRE database, consisting of 27 non-ASD and 85 individuals with ASD. Among the ASD group, 42 individuals had complete ADI-R scores available for analysis (Supplementary Data 6).

ADI-R modification

The complete ADI-R scores for 2800 individuals were obtained from AGRE. Within this dataset, individuals were categorized into the ASD group ($n = 2482$) and the non-ASD group ($n = 318$) using ADI-R cut-off scores. These scores were adjusted according to the methodology outlined in a prior publication by Hu and Steinberg[17] (Supplementary Data 9). Specifically, only ADI-R items scored on a scale from 0 to 3 were considered, where 0 signified normal and 3 indicated the most severe symptoms. Sub-questionnaires for each item, distinguishing between "current" and "ever" responses, were also incorporated. Scores of 8 or 9 in most items (except item 31) indicated "not asked" or "not applicable," while a score of -1 signified missing data and was replaced with a blank. For item 31 in the spoken language subgroup, a score of 8 was replaced with a rating of 3 to reflect insufficient language ability. If item 30 was scored as 1 or 2, items 31–41 were scored as 3. Regarding sarvant skill items, a score of 4 was substituted with a rating of 3 to maintain consistency on a 0–3 scale. Furthermore, a score of 7 was adjusted to either 0 or 3 based on the context of the answer for each question.

ADI-R preparation

After ADI-R modification, duplicate IDs of modified ADI-R scores were removed, and IDs without ASD cut-off predictions were excluded. This resulted in a final dataset of 2794 individuals, comprising 2480 with ASD and 314 non-ASD. Missing values were imputed using the k-nearest neighbors (k-NN) algorithm with $k = 5$. For the dataset utilized to develop machine learning (ML) screening models, the modified 2794 individuals were randomly divided into 2 groups, 10% of both the ASD and non-ASD groups were selected to create the "Validate_samples," while the remaining data was designated as "Training_samples" (Fig. 1 and Fig. 2d). The Training_samples were utilized to train the ML model, while the Validate_samples were reserved for assessing the performance of the trained ML model as new data. After preparation step, the Training_samples consisted of 2,514 individuals (n of ASD = 2232, n of non-ASD = 282), whereas the Validate_samples comprised 280 individuals (n of ASD = 248, n of non-ASD = 32). For the dataset utilized for developing ML subgrouping models, all 2480 ASD individual data (ASD_samples) were used (Fig. 4).

ML for ASD screening

ML models for ASD screening were developed through a systematic two-phase approach using Altair AI Studio software[29]. The initial phase focused on developing models using the complete ADI-R questionnaire, while the second phase evaluated a streamlined version using key diagnostic items. The development process began with training seven supervised ML algorithms on the full 149 sub-items from the ADI-R questionnaire. These algorithms included Naïve Bayes (NB), Decision Tree (DTree), Random Forest (RF), k-Nearest Neighbor (k-NN), Logistic Regression (LR), Support Vector Machines (SVM), and Deep Learning (DL). The Training_samples dataset underwent tenfold cross-validation to ensure model robustness. Each algorithm's parameters were optimized using the 'Optimize Parameters (Grid)' function to maximize accuracy, resulting in the identification of optimal parameter settings for each algorithm. The optimized models were then compared based on accuracy, specificity, sensitivity, and precision, with the best model selected using ROC (Receiver Operating Characteristic)

chart analysis and AUC (Area Under Curve). The Validate_samples dataset was used for final model validation to test each model's performance on previously unseen data (Fig. 1). To address the class imbalance in the training dataset (2232 ASD vs 282 non-ASD), the SMOTE (Synthetic Minority Over-sampling Technique) upsampling operator in Altair AI Studio was applied to the minority class (non-ASD). This technique generated synthetic samples of the minority class to create a balanced dataset with equal representation of ASD and non-ASD. The previously developed models were retrained on this balanced dataset to evaluate whether the class balance would affect model performance.

The second phase focused on identifying the most discriminative ADI-R items to potentially streamline the screening process. Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) from the mixOmics R package[27] analyzed the full ADI-R questionnaire to identify key items that optimally differentiated between ASD and non-ASD. The effectiveness of these sPLS-DA-selected items was then evaluated against the previously developed models, maintaining identical parameters and evaluation metrics (Fig. 1). Additionally, correlation analysis was performed to evaluate potential multicollinearity among the sPLS-DA-selected items. Spearman correlation coefficients were calculated between all ADI-R items to assess their pairwise relationships. Correlations with absolute values greater than 0.7 were considered strong indicators of potential redundancy.

ML for ASD subgrouping

The subgroup analysis was conducted through parallel clustering analyses of two distinct datasets: the ADI-R questionnaire data and gene expression data from transcriptome profiling. The first dataset comprised complete ADI-R questionnaire responses (149 sub-items) from 2480 ASD individuals, while the second dataset consisted of transcriptome profiling data from GSE15402 ($n = 42$)[25]. Principal Component (PC) scores were calculated for both datasets to reduce dimensionality and capture key variations in the data prior to clustering analysis. For both datasets, PC scores were calculated as a preprocessing step to reduce dimensionality and capture key variations in the data. Two unsupervised ML algorithms, k-means[30] and k-medoids[31] clustering, were then applied independently to each dataset's PC scores. The optimal number of clusters (k) was determined separately for each dataset through background prediction analysis using Partial Least Squares Discriminant Analysis (PLS-DA) from the mixOmics R package[27]. The selection criteria for optimal k focused on achieving clear separation between subgroups with minimal overlap. For each dataset, model selection between k-means and k-medoids was based on the averaged within-centroid distance.

The clustering analysis of ADI-R data resulted in the identification of ADI-R based ASD subgroups, representing distinct symptom-based ASD phenotypes. To characterize these ADI-R subgroups, sparse Partial Least Squares Discriminant Analysis (sPLS-DA) was applied to identify the key ADI-R items that most effectively discriminated between the identified subgroups. Independently, the clustering analysis of transcriptome profiling data produced transcriptome-based ASD subgroups, representing molecularly distinct ASD subtypes based on gene expression patterns.

Correlation analysis between clinical symptoms and gene expression in ASD subgroups

Following the identification of ADI-R-based and transcriptome-based ASD subgroups, a correlation analysis was performed to evaluate which subgrouping approach better captured the relationship between clinical symptom severity and gene expression patterns. This analysis utilized DIABLO from the mixOmics package, which enables comprehensive integration and analysis of multi-omics data. The correlation analysis focused on matched individuals who had both ADI-R scores and gene expression data available, allowing for direct comparison between clinical phenotypes and molecular profiles. The analysis was conducted separately for both subgrouping approaches. For ADI-R-based subgroups, the correlation between subgroup-specific ADI-R patterns and their corresponding gene expression profiles was examined. Similarly, for transcriptome-based subgroups, the relationship between gene expression patterns and the severity of clinical symptoms within each molecular subtype was analyzed. This parallel analysis structure enabled direct comparison of how effectively each subgrouping method aligned molecular and clinical characteristics of ASD.

Differentially expressed genes (DEGs) analysis and gene ontology analysis

Differentially gene expression (DEGs) analysis was performed using the limma package in R[32]. Gene expression data were filtered and preprocessed prior to analysis. Four separate DEGs analyses were performed: ASD versus non-ASD, and each of the three ADI-R subgroups versus their age- and sex-matched non-ASD controls. The analysis pipeline employed a linear modeling approach using the lmFit function, followed by empirical Bayes moderation using eBayes to improve the reliability of variance estimates. Contrasts were defined using makeContrasts and fitted to the linear model. Multiple testing correction was performed using the Bonferroni method to control for false positives. Genes were considered differentially expressed at a significance threshold of $p\text{-value} < 0.01$. Subsequently, significantly DEGs from each analysis were subjected to functional annotation, disease association analysis, and pathway analysis using Ingenuity Pathway Analysis (IPA) software.

Statistical analysis

Statistical analyses were performed using R software. A two-tailed Student's t-test with a significance threshold of $p\text{-value} < 0.05$ was used to test for differences between groups. Differentially gene expression analysis was conducted using the limma package with a significance threshold of $p\text{-value} \leq 0.01$. Spearman correlation analysis was performed to examine relationships between ADI-R items, with strong indicators of potential redundancy as $|r| > 0.7$. The relationships between clinical symptoms and gene expression patterns were analyzed using DIABLO from the mixOmics R package, with correlations filtered using a cut-off threshold of $r \geq 0.5$.

Data availability

The transcriptome data were deposited into the Gene Expression Omnibus database under accession number GSE15402 and are available at the following URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15402. Example from: https://doi.org/https://doi.org/10.1002/aur.73. The clinical data used in this study are available through the Autism Genetic Resource Exchange (AGRE) database. Researchers can request access to the AGRE dataset by registering at https://research.agre.org/agree/login.cfm. Upon approval, the ADI-R scores can be obtained through the AGRE data portal.

Received: 22 November 2024; Accepted: 20 March 2025
Published online: 05 April 2025

References

1. Boat, T. F. & Wu, J. T. (eds) *Mental Disorders and Disabilities Among Low-Income Children* (National Academies Press, 2015).
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5* (American Psychiatric Association, 2013).
3. Zeidan, J. et al. Global prevalence of autism: a systematic review update. *Autism Res.* **15**(5), 778–790 (2022).
4. Mandavilli, A., et al. Global map of autism prevalence by Spectrum. https://autismprevalence.thetransmitter.org/ (2025) [access 27 Jan, 2025].
5. Maenner, M. J. et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2020. *MMWR Surveill. Summ.* **72**(2), 1 (2023).
6. Chiarotti, F. & Venerosi, A. Epidemiology of autism spectrum disorders: a review of worldwide prevalence estimates since 2014. *Brain Sci.* **10**(5), 274 (2020).
7. Okoye, C., Obialo-Ibeawuchi, C. M., Obajeun, O. A., Sarwar, S., Tawfik, C., Waleed, M. S., *et al.* Early diagnosis of autism spectrum disorder: A review and analysis of the risks and benefits. *Cureus*. **15**(8) (2023).
8. Hodges, H., Fealko, C. & Soares, N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. *Transl. Pedia.* **9**, S55–S65. https://doi.org/10.21037/tp.2019.09.09 (2020).
9. Saeliw, T. et al. Integrated genome-wide Alu methylation and transcriptome profiling analyses reveal novel epigenetic regulatory networks associated with autism spectrum disorder. *Molecular Autism* **9**, 1–19 (2018).
10. Tangsuwansri, C. et al. Investigation of epigenetic regulatory networks associated with autism spectrum disorder (ASD) by integrated global LINE-1 methylation and gene expression profiling analyses. *PLoS ONE* **13**, e0201071 (2018).
11. Saeliw, T. et al. Epigenetic gene-regulatory loci in Alu elements associated with autism susceptibility in the prefrontal cortex of ASD. *Int. J. Mol. Sci.* **24**, 7518 (2023).
12. Saeliw, T. et al. LINE-1 and Alu methylation signatures in autism spectrum disorder and their associations with the expression of autism-related genes. *Sci. Rep.* **12**, 13970 (2022).
13. Kanlayaprasit, S. et al. Sex-specific impacts of prenatal bisphenol A exposure on genes associated with cortical development, social behaviors, and autism in the offspring's prefrontal cortex. *Biol. Sex Differ.* **15**, 40 (2024).
14. Thongkorn, S. et al. Sex differences in the effects of prenatal bisphenol A exposure on autism-related genes and their relationships with the hippocampus functions. *Sci. Rep.* **11**, 1241 (2021).
15. Kasitipradit, K. et al. Sex-specific effects of prenatal bisphenol A exposure on transcriptome-interactome profiles of autism candidate genes in neural stem cells from offspring hippocampus. *Sci. Rep.* **15**, 2882 (2025).
16. Saechua, C. et al. Impact of gene polymorphisms involved in the vitamin D metabolic pathway on the susceptibility to and severity of autism spectrum disorder. *Sci. Rep.* **14**, 28333 (2024).
17. Hu, V. W. & Steinberg, M. E. Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders. *Autism Res.* **2**(2), 67–77 (2009).
18. Tammimies, K. et al. Molecular diagnostic yield of chromosomal microarray analysis and whole-exome sequencing in children with autism spectrum disorder. *JAMA* **314**(9), 895–903 (2015).
19. Rutter, M., Le Couteur, A., Lord, C. *Autism Diagnostic Interview-Revised*, vol. 29, 30 (Western Psychological Services, 2003).
20. Lord C, Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2).* (Western Psychological Services, 2012).
21. Schopler, E., Reichler, R. J., Renner, B. R. *The childhood autism rating scale (CARS)* (WPS Los Angeles, 2010).
22. Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., DeLuca, T. F. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS ONE.* **7**(8), e43855 (2012).
23. Wall, D. P., Kosmicki, J., Deluca, T., Harstad, E. & Fusaro, V. A. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Transl. Psychiatry.* **2**(4), e100-e (2012).
24. Bone, D. et al. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. *J. Child Psychol. Psychiatry* **57**(8), 927–937 (2016).
25. Hu, V. W. et al. Gene expression profiling differentiates autism case–controls and phenotypic variants of autism spectrum disorders: Evidence for circadian rhythm dysfunction in severe autism. *Autism Res.* **2**(2), 78–97 (2009).
26. Hu, V. W., Lai, Y. Developing a predictive gene classifier for autism spectrum disorders based upon differential gene expression profiles of phenotypic subgroups. *N. Am. J. Med. Sci.* **6**(3) (2013).
27. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**(11), e1005752 (2017).
28. Duda, M., Kosmicki, J., Wall, D. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Transl. Psychiatry.* **4**(8), e424-e (2014).
29. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T., editors. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006).
30. Lüdecke, D. et al. easystats: Streamline model interpretation, visualization, and reporting. *R package*. https://easystats.github.io/easystats/ (2023).
31. Maechler, M. et al. cluster: Cluster Analysis Basics and Extensions. *R package*. https://cran.r-project.org/package=cluster (2011).
32. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47. https://doi.org/10.1093/nar/gkv007 (2015).

Acknowledgements

We gratefully acknowledge the Autism Genetic Resource Exchange (AGRE) Consortium and the participating AGRE families for providing valuable resources for this research. This research was supported by the Program Management Unit for Human Resources and Institutional Development, Research and Innovation (PMU-B) (grant number B36G660008, the Thailand Science Research and Innovation Fund, Chulalongkorn Universi-

ty (HEAF67370092 and HEA_FF_68_083_3700_006), and the Ratchadapisek Somphot Fund for Supporting Chulalongkorn Autism Research and Innovation Center of Excellence (ChulaACE), Chulalongkorn University (CE66_046_3700_003), awarded to TS. WY was supported by the Second Century Fund (C2F), Chulalongkorn University, the Royal Golden Jubilee Ph.D (sub-code NRCT5-RGJ63001-018). Program, and the 90th Anniversary of Chulalongkorn University Scholarship (grant number GCUGR1125662069D). TSae and SK were supported by the Second Century Fund (C2F), Chulalongkorn University. MLE received funding from the 90th Anniversary of Chulalongkorn University Scholarship and the Graduate Scholarship Programme for ASEAN or Non-ASEAN Countries. CP was funded by the 90th Anniversary of Chulalongkorn University Scholarship and H.M. King Bhumibol Adulyadej's 72nd Birthday Anniversary Scholarship. We would like to thank Prof. Kim-Anh Lê Cao from Melbourne Integrative Genomics and School of Mathematics and Statistics, University of Melbourne for her guidance and advice on the use of mixOmics software. We are deeply grateful to all funding organizations for their contributions.

Author contributions

T.S. conceptualized the study and acquired funding. W.Y., V.W.H., and T.S. performed data curation. W.Y. and T.S. conducted the investigation. W.Y., T.Sae., M.L.E., C.P., S.K. and T.S. developed the methodology. T.S. administered the project. W.Y. and V.W.H. provided resources. T.S., W.C., P.T., and V.W.H. supervised the research. W.Y., T.Sae. and T.S. performed validation. W.Y., T.Sae., and T.S. wrote the original draft. W.Y., T.Sae., and T.S. reviewed and edited the manuscript. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethics declarations

This research study has undergone ethical considerations by the research ethics review committee for research involving human research participants, group 1, Chulalongkorn University, and was considered an exception (COA NO. 109/2021) for complete ADI-R data and human ethical approval (COA No.1238/2022) has been approved by The Institutional Review Board of the Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. All methods were performed in accordance with the relevant guidelines and regulations.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-95291-5.

Correspondence and requests for materials should be addressed to T.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.