·论著·一次研究·

小样本多组学数据下利用注意力和元学习进行孤独症诊断的方法研究

王 琦[1]，谢 堃[1]，梁学致[1]，罗向阳[2]，刘 瑛[3]，陈 雯[1]

1. 中山大学公共卫生学院医学统计学系（广州 510080）
2. 中山大学孙逸仙纪念医院儿科（广州 510000）
3. 广东省妇幼保健院儿童保健科（广州 511400）

【摘要】目的 利用注意力和元学习开发针对小样本多组学数据的深度学习方法，用于建立孤独症诊断模型。**方法** 设计由组学特征预降维模块、多组学数据融合学习模块和参数优化模块组成的深度学习模型—基于元学习的注意力网络（MLAN），对高维多组学数据进行差异表达分析，初步筛选出重要特征用于后续建模；采用多通道注意力机制进行多组学数据的组间重要性学习和融合，构建神经网络模型进行融合特征的进一步学习并实现诊断任务；使用 Reptile 算法进行参数优化，得到最佳模型。为验证模型效果，采集 58 例儿童的唾液样本，包括诊断孤独症患儿 21 例，单纯社交障碍患儿 12 例，以及健康对照儿童 25 例，采用质谱分析方法检测样本的蛋白质和代谢组学数据。按照 4∶1 的比例将样本分层随机划分为训练集和测试集，训练集按同样方法划分出训练数据和验证数据，分别用于模型的训练和验证，测试集用于模型效果的最终评估。构建 5 种基线模型和 3 种消融模型，采用准确率、宏 F1 分数和加权 F1 分数进行所有模型的评价。**结果** MLAN 在多分类孤独症诊断任务中取得的准确率、宏 F1 和加权 F1 分数分别为 0.850 ± 0.066、0.817 ± 0.103 和 0.834 ± 0.087，3 个指标均优于所有基线模型和消融模型。**结论** 本研究提出的基于注意力和元学习的 MLAN 模型能有效应对异质性小样本多组学数据，并在孤独症的诊断问题上取得良好效果，有望为孤独症的临床诊断提供帮助。

【关键词】孤独症诊断；注意力机制；元学习；小样本；多组学数据整合

【中图分类号】R 749.94 **【文献标识码】**A

Study on the method of using attention mechanism and meta-learning to diagnose autism under small sample multi-omics condition

WANG Qi[1], XIE Kun[1], LIANG Xuezhi[1], LUO Xiangyang[2], LIU Ying[3], CHEN Wen[1]

1. Department of Medical Statistics, School of Public Health, Sun Yat-sen University, Guangzhou 510080, China

2. Department of Pediatrics, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China

3. Department of Child Health, Guangdong Provincial Maternal and Child Health Hospital, Guangzhou 511400, China

Corresponding author: CHEN Wen, Email: chenw43@mail.sysu.edu.cn

DOI: 10.12173/j.issn.1005-0698.202412113

基金项目：广州市科技计划项目（2024A04J6551）；中山大学公共卫生学院融创人才支持计划

通信作者：陈雯，博士，教授，博士研究生导师，Email：chenw43@mail.sysu.edu.cn

【Abstract】Objective To develop a deep learning method for small sample multi-omics data using attention mechanism and Meta-learning for the establishment of autism diagnosis model. Methods MLAN (Meta-learning based attentive network) consisting of the omics feature pre-reduction module, the multi-omics data fusion and feature learning module, and the parameter optimization module was designed. Firstly, differential expression analysis was performed on high-dimensional multi-omics data to preliminarily screen out unimportant features. Secondly, a multi-channel attention mechanism was used to learn the importances of every set of omics data and to realize data fusion, and a two-layer fully connected network was constructed to further extract latent features and realize the diagnosis task. Finally, the Meta-learning algorithm Reptile was used to optimize the initial parameters of the above model to obtain the optimal parameters. A total of 58 children's saliva samples were collected, including 21 children diagnosed with autism, 12 children with social disorders, and 25 healthy controls, and the protein and metabolomics data were detected by mass spectrometry. All data were randomly divided into training set and test set by 4∶1, and the training set was divided into training data and validation data in the same way for model training and validation. The test set was used for the final evaluation of the model effect. Five baseline models and three ablated models were constructed and evaluated along with MLAN based on metrics including multi-classification accuracy, F1-macro and F1-weighted scores. Results The constructed multi-classification autism diagnosis model MLAN achieved multi-classification accuracy, F1-macro and F1-weighted scores of 0.850±0.066, 0.817±0.103 and 0.834±0.087. The values of all three indicators were better than those of baseline models and the ablated models. Conclusion The proposed MLAN can effectively deal with heterogeneous multi-omics data with small samples and achieve good results, which is expected to provide assistance for the clinical diagnosis of autism.

【Keywords】Autism diagnosis; Attention mechanism; Meta learning; Small sample; Multi-omics data integration

孤独症谱系障碍（autism spectrum disorders，ASD），通称孤独症或自闭症，指的是一类以个人不同程度的社交能力缺陷以及重复、受限的行为模式、兴趣或活动为主要特征的神经发育障碍[1]。2020年的一项研究[2]结果显示，在中国，儿童的孤独症患病率约为0.29%，略低于世界平均水平。孤独症具有广泛异质性，在不同年龄和发育状态群体中具有不同的表现，其症状与注意缺陷多动障碍等疾病存在相似性，疾病机制和生物标志物尚不明确，缺乏客观的临床诊断方法和特效治疗药物[3]。目前，孤独症的诊断由临床医生根据患者症状和标准化工具做出，具有一定主观性，存在误诊、漏诊风险[4–5]。准确的诊断是后续制定有效药物治疗方案的前提，对于已确诊的孤独症患者，根据其核心症状可以使用利培酮、阿立哌唑等药物进行治疗，但以上药物的使用还缺乏高质量证据基础[6]。为提高孤独症的早期诊断效果并促进药物治疗研究，目前已开展了基于影像和脑电图等技术的相关研究[7–9]，但仍缺乏基于生物特征的孤独症临床诊断工具。

多组学技术有望为孤独症机制研究、早期诊断和药物治疗提供新的机会。组学数据蕴含了与疾病、机体相关的丰富信息。一项基于肠道菌群的研究[10]表明孤独症儿童肠道菌群结构与其行为症状密切相关，基于家系间外显子测序数据的研究[11]也揭示了孤独症相关的关键遗传变异和生物通路，这些研究已证明组学特征在疾病机制中的重要作用，提示高通量测序数据在孤独症诊断中的潜在价值。然而，由于不同分子水平物质含量、测量技术和实验批次均不同，多组学数据具有较高异质性。由于孤独症患病率较低，且患儿存在的社交障碍和刻板行为等临床症状会限制生物样本采集，故孤独症多组学数据还普遍存在小样本的问题[12]。

因此，本研究以基于小样本多组学数据进行孤独症诊断为目的，提出一种新的深度学习模型，即基于元学习的注意力网络（Meta-learning based attentive network，MLAN），通过在端到端神经网络中引入多通道注意力机制（multi-channel attention mechanism，MCA）实现异质多组学数

据的融合和特征提取，进而通过元学习（Meta learning，ML）优化神经网络参数以应对数据的小样本问题，最终实现疾病诊断任务。

1 资料与方法

1.1 MLAN模型构建方法

MLAN 模型由 3 个模块组成，分别是组学特征预降维模块、多组学数据融合学习模块和参数优化模块，整体流程见图 1。

1.1.1 组学特征预降维模块

由于组学数据数量众多，为更好地应对小样本问题、降低模型复杂度并提高模型效率，本研究使用基于生物信息的差异表达分析法（differential expression analysis，DEA）对所有组学数据进行特征筛选和预降维[13]。采用线性模型对组学的表达矩阵进行拟合并对比不同类别下物质的表达水平，使用贝叶斯方法对估计结果进行调整以提高统计稳定性。使用基于 Benjamini–Hochberg（BH）法的错误发现率（false discovery rate，FDR）校正多重比较的 P 值，最终基于校正 P 值和倍数变化（fold change）筛选出不同类别中表达量具有显著变化的物质，即差异表达物。为适应多通道自注意力机制对每个通道维度相同的数据要求、避免不同组学特征数量不均衡带来的偏差并简化模型设计，在 MLAN 中，按校正 P 值由小到大在不同组学中保留了相同数量的特征用于后续高维特征向量的构建。

如图 1 所示，假设 $n \times P_m$ 维矩阵 $\boldsymbol{X}^m = (X_1^m, X_2^m, ..., X_n^m, ..., X_N^m)'$ 是第 m 类组学的特征表达矩阵，其中 $1 \times P_m$ 维向量 $X_n^m = (x_{n,1}^m, x_{n,2}^m, ..., x_{n,p}^m, ..., x_{n,P}^m)$ 表示第 n 个样本的第 m 类组学所有特征的表达值，$x_{n,p}^m$ 表示第 p 个物质的表达水平。则经过差异表达分析后，所有的 P_m 均降维至相同尺度，即不同组学的特征表达矩阵 $\boldsymbol{X}^m$ 维度由 $N \times P_m$ 化为相同的 $N \times P$，其中 P 为待设置的超参数。

1.1.2 多组学数据融合学习模块

参考 SENet（squeeze–and–excitation networks）这种多通道注意力机制的结构设计[14]，将多组学数据视为高维特征的不同通道，进而通过自注意力（self–attention）算法学习到通道重要性，也即不同组学特征的重要性，最后使用加权求和的方法得到与原特征维度一致的多组学融合特征。如图 1 所示，首先将 M 个 $N \times P$ 维特征表达矩阵 $\boldsymbol{X}^m$ 映射到 $N \times M \times P$ 维空间中，使用平均池化函数 $F_{sq}(\cdot)$ 提取每个样本在每组的特征表达向量 $\mathcal{P}_m = (\mathcal{p}_1^m, \mathcal{p}_2^m, ..., \mathcal{p}_n^m, ..., \mathcal{p}_N^m)'$ 其中 $\mathcal{p}_n^m$ 表示第 n 个样本的第 m 类组学指标，计算如下：

$$\mathcal{p}_n^m = F_{sq}(X_n^m) = \frac{1}{P}\sum_{p=1}^{P} x_{n,p}^m, m = 1,2, ..., M \quad \text{公式（1）}$$

进一步使用激励函数 $F_{ex}(\cdot)$ 学习不同组学数据的组间关联并输出每个组学的多通道注意力得分 $S(m)$，计算如下：

$$S(m) = F_{ex}(\mathcal{P}_m, \boldsymbol{W}^{MCA}) = sigmoid\left[W_{(2)}^{MCA}\left(Relu\left(W_{(1)}^{MCA}\mathcal{P}_m\right)\right)\right] \quad \text{公式（2）}$$

其中 $\boldsymbol{W}^{MCA}$ 为待优化参数。最后多组学融合特征矩阵计算为不同组学特征表达的加权和，即

$$\boldsymbol{X}^0 = \sum_{m=1}^{M} S(m)\boldsymbol{X}^m \quad \text{公式（3）}$$

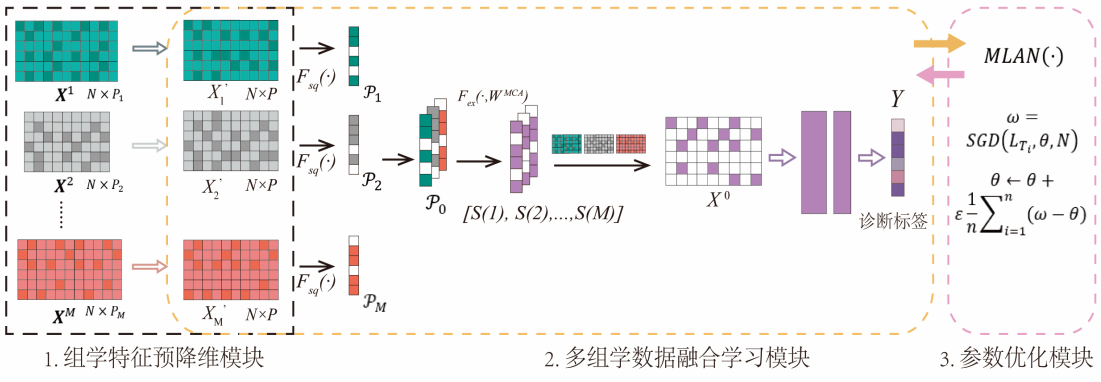

图1 MLAN模型的整体流程

Figure 1. Overall process of MLAN

其中 $\boldsymbol{X}^0$ 的维度与 $\boldsymbol{X}^m$ 一致，均为 $N\times P$。为适应小样本问题、减小过拟合风险，本研究使用简易的两层线性全连接函数进一步学习融合特征的潜在表示并实现样本分类。使用 *classifier*(·) 函数表示该部分学习过程，输入的特征维度是 $N\times P$，隐藏层维度是待设置的超参数，函数的输出维度为样本的类别数。在本研究中，基于数据特征和经验确定隐藏层维度为 128。$\boldsymbol{W}^{classifier}$ 表示该部分的待优化参数，则有样本预测标签向量 Y 计算如下：

$$Y = classifier\left(\boldsymbol{X}^0, \boldsymbol{W}^{classifier}\right) = W_{(2)}^{classifier}\left(LeakyRelu\left(W_{(1)}^{classifier}\boldsymbol{X}^0\right)\right)$$

公式（4）

综上，若使用 *MLAN* (·) 表示以上多组学数据融合学习模型的数据传递过程，则有：

$$Y = MLAN\left(\boldsymbol{X}^1, \boldsymbol{X}^2, ..., \boldsymbol{X}^M, \boldsymbol{W}^{MCA}, \boldsymbol{W}^{classifier}\right)$$

公式（5）

1.1.3 参数优化模块

基于以上构建的模型 *MLAN*(·)，进一步使用批次更新的 Reptile 算法学习模型的最优初始化参数，获得能迅速适应新数据和小样本情况的元模型 [15]。元模型的参数学习通过构建内循环和外循环实现，其中内循环用于特定样本上的基模型优化，外循环用于所有样本上的元模型全局参数更新。假设基模型 *MLAN*(·) 和元模型初始参数向量均为 θ，首先在训练数据中完全随机采样一定数量的样本组成 n 个不同任务 T_i，其中 i=1, 2, ..., n, n 表示训练的任务数即元模型参数更新的轮次数。其次在每个任务 T_i 上使用随机梯度下降（stochastic gradient descent，SGD）法迭代 N 次得到基模型参数估计值 θ_i^N，其中超参数 N 表示 Reptile 算法的内循环次数。设置外循环的学习率为超参数 ε，使用 2 个参数估计值的差值更新元模型参数，即：

$$\theta_i^N = SGD\left(L_{T_i}, \theta, N\right)$$

公式（6）

$$\theta \leftarrow \theta + \varepsilon\frac{1}{n}\sum_{i=1}^{n}\left(\theta_i^N - \theta\right)$$

公式（7）

进一步可通过调整内循环和外循环的超参数优化元模型训练结果，最终得到基模型 *MLAN*(·) 的最优参数 θ，即可在新数据上快速泛化的 MLAN 元模型。

1.2 实证数据及来源

本研究采用病例－对照研究设计，课题组于 2022 年 9 月—2023 年 10 月在中山大学孙逸仙纪念医院、广州维儿康专科门诊部和广东省妇幼保健院采集新诊断孤独症患儿、单纯社交障碍患儿和健康对照儿童各 25 例的唾液样本。纳入标准：①儿童年龄 2~6 岁；②孤独症和单纯社交障碍儿童的诊断由临床医生依据《精神障碍诊断与统计手册（第 5 版）》（*Diagnostic and Statistical Manual of Mental Disorders Fifth Edition*，DSM-5）[16] 的诊断标准及孤独症儿童行为评定量表（Autism Behavior Checklist，ABC）[17] 和儿童期孤独症评定量表（Childhood Autism Rating Scale，CARS）[18] 的评分做出，健康对照儿童来自于健康体检人群。排除标准：①有遗传综合征；②有智力障碍；③患有慢性自身免疫性或炎症性疾病；④有龋齿、牙周疾病或炎症性水肿。由于单纯社交障碍患儿的临床样本有限，本研究最终采集到 25 例新诊断孤独症患儿（其中 4 例患儿的样本因未通过质量控制而被排除，最终纳入分析样本 21 例）、12 例单纯社交障碍患儿对照和 25 例健康对照儿童的唾液样本。3 组儿童在采集唾液样本前均禁食不少于 30 min，清洁口腔后采用项目组研制的唾液采集器 [19] 分别采集唾液样本 150 μL 和 50 μL 用于蛋白质和代谢组学检测，样本储存于 –80 °C 环境。本研究基于规范的数据采集流程，已获得中山大学公共卫生学院生物医学研究伦理审查委员会审查批准（编号：中大公卫医伦〔2022〕第 048 号），收集的信息严格保密，儿童家长均签署知情同意书。

在蛋白组学分析中，首先使用添加了抑制剂的裂解缓冲液提取样本组织并在 4 °C 环境中 12 000 × *g* 离心 10 min，提取上清液，使用二喹啉甲酸（bicinchoninic acid，BCA）试剂盒（BCA Protein Assay Kit）测定蛋白浓度。其次使用 5 mmol · L^{-1} 的二硫苏糖醇在 56 °C 环境中还原蛋白质溶液 30 min，在暗室中用 11 mmol · L^{-1} 的碘乙酰胺（iodoacetamide，IAA）烷基化溶液 15 min，使用四乙基溴化铵（tetraethylammonium bromide，TEAB）稀释溶液，并使用胰蛋白酶分次消化蛋白质至肽段。最后使用液相色谱－串联质谱（liquid chromatography-tandem mass spectrometry，LC-MS/MS）法获得肽段信号，

并使用 Spectronaut 17.0 软件进行数据搜索和鉴定[20-21]。对于代谢组学分析，首先将冷冻样本在冰上解冻并涡旋震荡 10 s，取 50 μL 样本，加入 150 μL 含有内标的提取液，涡旋 3 min 并在 4 ℃ 环境中 13 523×*g* 离心 10 min。其次收集 150 μL 上清液于 −20 ℃ 环境中静置 30 min，后在 4 ℃ 下 13 523×*g* 离心 3 min，收集 120 μL 上清液用于液相色谱－质谱（liquid chromatography-mass spectrometry，LC-MS）分析。使用 Analyst TF 1.7.1 软件的信息依赖性采集（information dependent acquisition，IDA）模式采集数据，使用 ProteoWizard 软件将数据转化为 mzXML 格式，XCMS 程序用于代谢信号处理，实验室自建数据库、公共数据库、AI 数据库和 metDNA 数据库用于识别代谢信息[22-23]。

基于得到的蛋白质和代谢物表达数据，参考相关文献，删除在所有样本中缺失值比例高于 80% 的特征，使用 K- 近邻（K-nearest neighbor，KNN）法填充余下的缺失值，其中超参数 K 的取值为 5，使用 z-score 法对组学数据中每个样本的所有特征进行标准化[24-25]。最终得到预处理后的多组学数据表达矩阵，包括 2 283 个蛋白质特征和 10 279 个代谢物特征。

1.3 实验设计与模型评价指标

对于本研究使用的数据，按 4∶1 的比例分层随机划分出训练集和测试集分别用于模型的训练和效果测试。在 MLAN 模型的训练中，按同样方法划分训练集得到训练和验证数据。设置外循环轮次，使用早停（early stopping）技术监控模型的训练损失，通过观察损失函数曲线和模型在验证数据上的表现优化超参数。内循环次数 N 的初始化范围为 1~10，内循环和外循环的学习率范围为 1.0×10^{-5}~1.0×10^{-2}，外循环轮次数 n 的范围为 100~500，早停的阈值范围为 10~50，每个任务的样本数取值为 4~16。为提高稳健性，基于 10 个不同的随机数重复实验 10 次，模型的平均表现被用于最终效果评估。

选择逻辑斯谛回归（logistic regression，LR）、KNN、支持向量机（support vector machine，SVM）、朴素贝叶斯（naive Bayes，NB）和多层感知器（multi-layer perceptron，MLP）模型作为基线，其中 LR 为线性模型，可处理简单医药数据、建立疾病风险或药物不良反应预测模型以及分析生物标志物关联；KNN 常用于小样本数据评估，实现疾病分类或药物结构分类任务；SVM 适用于高维小样本数据，在医学影像分类、癌症诊断、药物预测等方面应用广泛；NB 常用于电子病历文本分析和基因突变数据分析；MLP 表达能力强，可用于多组学和多变量非线性数据分析，并用于疾病预测、药物敏感性预测建模[26]。以上方法覆盖统计和机器学习策略，使用范围和特点不一、各具代表型，能提供强有力的基线参考，在本研究中用于与 MLAN 进行对比，以验证 MLAN 模型的有效性和优越性。基线模型的多组学融合特征是通过直接串联特征表达矩阵获得的，模型参数设置参考了相关文献[25, 27-28]，采用与 MLAN 相同的数据划分方式进行模型的训练、验证和测试。

此外，构建消融模型 MLAN-ML、MLAN-MCA 和 MLAN-ML-MCA，其中 MLAN-ML 在 MLAN 中删去了参数优化模块，即为“1.1.3”中的基模型，用于检验 MLAN 中元学习部分对模型效果的提升能力；MLAN-MCA 删去了多通道注意力机制，使用串联法融合多组学数据，用于检验多通道注意力机制对模型效果的提升能力；MLAN-ML-MCA 删去了元学习和多通道注意力机制，直接使用神经网络对串联多组学数据进行分析，用于进一步验证两个模块的效果。MLAN-ML 和 MLAN-ML-MCA 使用小批量梯度下降（mini-batch gradient descent，MBGD）、自适应矩估计（adaptive moment estimation，Adam）、K 折交叉验证（K-fold cross validation，K-fold CV）和早停技术并分迭代轮数（epoch）进行参数优化[29]。MLAN-MCA 使用与 MLAN 相同的方法进行训练、验证和测试。

采用准确率、宏 F1 分数和加权 F1 分数评价以上构建模型效果，进一步绘制受试者工作特征（receiver operating characteristic，ROC）曲线并计算曲线下面积（area under curve，AUC），观察模型整体效果[29-30]。研究基于 R 4.2.3 软件和 Python 3.11.3 软件进行。

2 结果

2.1 模型训练和测试

经训练最终确定 MLAN 模型的超参数取值，包括内循环次数 N 和学习率为 3 和 1.0×10^{-2}，外循环的轮次数 n 为 100，学习率 ε 为 1.0×10^{-5}，早停的阈值为 10，每个任务中样本数为 8 个。消

融模型 MLAN–ML 和 MLAN–ML–MCA 采用 5 折交叉验证，每批次样本量为 8，学习率为 1.0×10^{-3}，迭代轮数为 100，早停的阈值为 10。MLAN–MCA 使用与 MLAN 相同的训练超参数。

MLAN 模型在第 1 对划分的训练和测试数据上的结果见图 2，MLAN 的损失函数曲线在训练轮次数到达 40 时已经下降并维持在较低的水平（图 2–A），在测试数据上的表现也显示了 MLAN 良好的疾病诊断效果，其多分类 ROC 曲线均靠近左上角，宏平均和加权平均 AUC 值分别为 0.915 和 0.879。进一步发现 MLAN 在孤独症类、健康对照类和单纯社交障碍类样本的 AUC 值分别为 0.829、0.906 和 0.926，多分类 AUC 值最小为 0.829，属于孤独症类样本（图2–B）。

2.2 与基线模型对比

与 MLAN 的验证流程相同，最终得到每个基线模型在孤独症多组学数据上的表现如表1 所示。MLAN 在测试集上的平均准确率达到 0.850 ± 0.066，显示了良好的疾病诊断效果，其宏 F1 分数和加权 F1 分数分别为 0.817 ± 0.103 和 0.834 ± 0.087，提示模型良好的疾病诊断综合性能。

基线模型中表现最好的是 LR，其准确率、宏 F1 分数和加权 F1 分数分别为 0.825 ± 0.062、0.784 ± 0.099 和 0.805 ± 0.081。尽管 LR 的 3 个指标取值高于其他基线模型，但仍低于 MLAN，说明 MLAN 基于小样本多组学数据进行孤独症诊断具有优越性。

2.3 与消融模型对比

消融模型 MLAN–ML、MLAN–MCA、MLAN–ML–MCA 和 MLAN 在测试数据上的平均表现见图 3。仅去除参数优化模块的 MLAN–ML 模型的准确率、宏 F1 分数和加权 F1 分数分别为 0.833 ± 0.056、0.800 ± 0.097 和 0.820 ± 0.079，略低于 MLAN 模型，高于表 1 中所有基线模型。仅去除多通道注意力机制的 MLAN–MCA 模型的准确率、宏 F1 分数和加权 F1 分数分别为 0.775 ± 0.056、0.693 ± 0.101 和 0.729 ± 0.084，均低于 MLAN 和 MLAN–ML，其宏 F1 分数和加权 F1 分数与 MLAN 的对应指标至少相差 0.100，但 MLAN–MCA 的综合效果优于表 1 中除 LR 之外的其他基线模型。

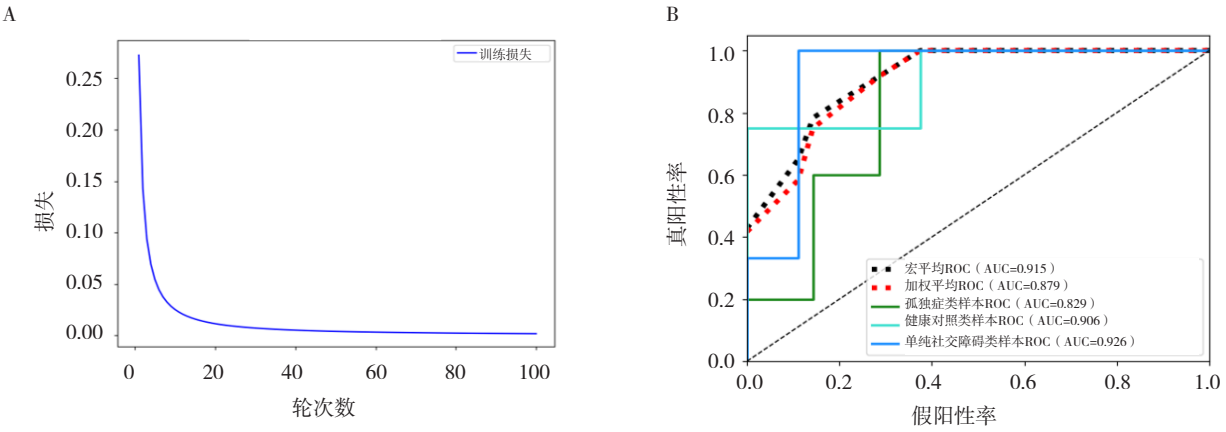

图2 MLAN在训练集上的损失函数曲线和在测试集上的多分类ROC曲线

Figure 2. The loss function curve of MLAN on the training set, and the multi–classification ROC curves on the test set

注：A. 训练集上的损失函数曲线；B. 在测试集上的多分类ROC曲线。

表1 MLAN和基线模型在测试集上的准确率、宏F1分数和加权F1分数（$\bar{x} \pm s$）

Table 1. The accuracy, F1–macro and F1–weighted scores of MLAN and baseline models on the test sets ($\bar{x} \pm s$)

模型	准确率	宏F1分数	加权F1分数
LR	0.825 ± 0.062	0.784 ± 0.099	0.805 ± 0.081
KNN	0.592 ± 0.100	0.470 ± 0.156	0.514 ± 0.144
SVM	0.742 ± 0.062	0.610 ± 0.092	0.665 ± 0.078
NB	0.658 ± 0.083	0.565 ± 0.111	0.606 ± 0.098
MLP	0.733 ± 0.116	0.679 ± 0.147	0.707 ± 0.129
MLAN	0.850 ± 0.066[a]	0.817 ± 0.103[a]	0.834 ± 0.087[a]

注：[a]所有模型中最佳结果。LR. 逻辑斯谛回归；KNN. K–近邻；SVM. 支持向量机；NB. 朴素贝叶斯；MLP. 多层感知器；MLAN. 基于元学习的注意力网络。

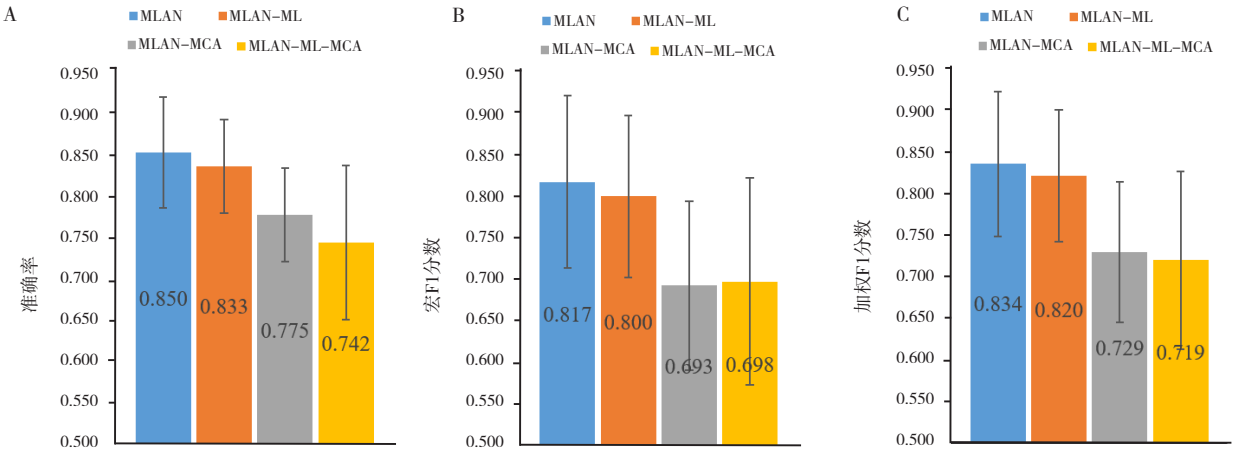

图3 MLAN和消融模型在测试集上的效果对比

Figure 3. Comparison of the effects of MLAN and the ablated models on the test sets

注：A. 模型的准确率；B. 模型的宏F1分数；C. 模型的加权F1分数；柱状图表示模型在多次实验上的均值，误差线表示模型在多次实验上的标准差。

此外，进一步测试了去除参数优化模块和多通道注意力机制的 MLAN–ML–MCA 模型在孤独症小样本数据上的效果。该模型的准确率、宏 F1 分数和加权 F1 分数分别为 0.742 ± 0.092、0.698 ± 0.124 和 0.719 ± 0.106，准确率和加权 F1 分数低于 MLAN 和其他消融模型；宏 F1 分数低于 MLAN 和 MLAN–ML，但略高于MLAN–MCA（0.698 ± 0.124 vs. 0.693 ± 0.101）。在基线模型中，LR 的表现明显优于 MLAN–ML–MCA，SVM 取得了与之相当的准确率，MLP 取得的综合表现只以较小的差距低于该模型（表 1），可见消融模型 MLAN–ML–MCA 并不具备显著优于基线模型的效果。

3 讨论

目前孤独症缺乏特效药物治疗，其诊断主要基于标准化工具。虽然学者们采用 LR、SVM、NB、卷积神经网等方法开发了基于儿童特定行为、脑核磁共振成像或脑电图、单核苷酸多态性等各类特征的孤独症诊断或分类模型 [31–33]，但能揭示多层次生物机制特征的多组学研究仍较为缺乏，基于多组学数据的诊断模型仍有待开发 [34–35]。

多组学数据常具有样本量少、特征维度高、组间数据异质性显著等问题，对于孤独症等患病率较低的疾病，小样本量问题进一步加剧，可能会使有意义的生物学信号或多组学互补信息被掩盖，影响数据建模的拟合效果和泛化性。本研究提出了一种基于注意力机制和元学习框架的小样本多组学数据整合和分析模型 MLAN，并在孤独症患儿和对照患儿的多组学数据中进行了实证研究。该模型首先利用差异表达分析法对高维多组学数据进行了特征预降维，之后通过多通道注意力机制进行了异质性多组学数据的特征融合，通过一个两层全连接网进一步学习融合特征的潜在表达并得到样本的分类预测标签，最后使用 Reptile 算法进行构建模型的参数优化，得到更利于小样本泛化的最终模型。MLAN 在测试数据上的 ROC 曲线以及与基线模型效果的比较，印证了该模型在小样本多组学数据处理和孤独症患儿诊断方面的有效性和优越性。

现有的基于多组学数据融合的疾病诊断模型大多直接串联原数据或数据特征，可能会忽略不同组学数据之间的异质性处理 [36–37]。注意力机制借鉴了人类注意力的工作方式，能从输入中动态选择信息，使模型聚焦于与任务最相关的部分，从而提高特征学习效果 [38]，然而现有的基于注意力的多组学数据融合分析模型往往只学习同组学内部的特征重要性，忽略了不同组学数据之间的差异 [13, 39–40]。本研究使用了基于 SENet 的多通道注意力机制实现不同组别的重要性学习和多组异质性数据的融合。结果显示，去除多通道注意力机制的消融模型 MLAN–MCA 在孤独症多组学小样本数据上的综合表现劣于 MLAN，体现了多通道注意力机制在处理并融合异质性多组学数据问题上的有效性。同理，消融模型 MLAN–ML 和 MLAN–ML–MCA 之间的表现差异也印证了注意力机制对提高模型诊断效果的积极作用。

对于小样本问题，直观的处理思路是进行数据增强或生成伪数据，然而这类方法需要合理设计数据增强或生成策略，否则可能引入与原数据无关的噪声并影响模型真实效果。元学习通过“学习如何学习”的策略，可为小样本问题的处理提供机会。本研究提出的参数优化模块旨在通过 Reptile 算法在多个小任务上优化模型的初始参数，从而学习到能够在新任务上快速泛化的最佳参数。消融实验的结果已验证 MLAN 中 Reptile 框架的有效性，也即元学习算法处理小样本多组学数据的有效性。目前基于元学习策略的多组学数据分析模型主要有用于癌症放疗预后及生存分析的 MMOSurv（Meta-learning for multi-omics data based survival analysis）[41] 和用于癌症诊断的 MUMA（multi-omics Meta-learning）[42]，本研究在 MMOSurv 模型的算法基础上针对样本多分类问题进行改造，在训练数据中多次采样得到具有较小样本量的不同任务，从而实现在内循环上的参数逐步优化并获得全局优化的元模型参数。此外，相比于 MUMA 模型中使用的 MAML（model-agnostic Meta-learning）[43] 算法，Reptile 算法通过在外循环中使用参数均值代替二阶梯度进行参数更新，可降低时空复杂度和敏感性，更适用于小样本场景 [15]。在本研究中，提出的 MLAN 模型采用了批次更新的 Reptile 算法，成功学习到基模型的最佳参数并最终提高了模型的孤独症综合诊断效果。

进一步观察 MLAN 诊断不同类别样本的 AUC 值，发现孤独症类样本的 AUC 显著低于单纯社交障碍和健康对照类（0.829 vs. 0.926 vs. 0.906）。这种现象可能与孤独症样本的内在特征相关，孤独症患儿的临床症状具有异质性，个体之间存在较大差异，因而数据本身可能存在较高的噪声，模型难以找到一致且稳定的多组学特征，诊断效果略差于单纯社交障碍类 [44-45]。对于单纯社交障碍类样本，所有儿童均仅存在社交障碍这一主要临床表征，其数据内部的同质性可能较强，因此模型具有较高的区分度 [46]，提示未来研究需要进一步区分孤独症的核心症状，以寻找更优的诊断分类特征和方法。此外，对于样本量小的类别（受限于临床样本的可获得性，单纯社交障碍仅采集到 12 例样本），其特征空间可能会更集中，需要额外警惕模型在该组数据上可能存在的过拟合倾向 [47]。

另一方面，通过观察基线模型的结果，发现 LR 的表现仅次于 MLAN。相比于深度学习方法，以 LR 为代表的广义线性模型简单易用，可解释性强，目前广泛用于疾病和药物的各类预测任务，但可能存在的问题是无法捕捉数据中的复杂非线性关系 [48]。在本研究中，可能由于使用的孤独症多组学数据集存在高维、相关且线性可分的特征，LR 可以直接建模线性决策边界、输出分类概率，受数据分布特征、维度以及样本量的影响较小，因此也取得了良好的疾病诊断效果 [49]。

本研究的优势在于创新性地使用了基于 Reptile 的元学习算法并设计了基于多通道注意力机制的多组学数据融合分析模型 MLAN，有效应对了多组学数据的异质性、高维稀疏性和小样本问题。通过在孤独症小样本多组学数据上进行模型验证和分析，结果表明，MLAN 具有良好的多组学数据建模效果和孤独症诊断能力。本研究的主要局限性在于模型中没有设计特征可视化算法以分析各类别特征分布情况，同时没有对模型进行解释以识别孤独症患儿、健康对照和单纯社交障碍类的关键特征。

综上，基于由唾液样本分析得到的小样本多组学数据，本研究提出的 MLAN 模型在孤独症诊断任务的应用中取得了良好的效果，提示该方法在小样本多组学数据学习上的良好性能。未来可对模型做进一步的优化，提高模型对各类疾病的诊断能力和对疾病关键特征的识别能力，并在其他少见病和罕见病的诊断和预测中做进一步验证和应用，从而为疾病的早期诊断、标志物和治疗靶点识别以及药物治疗提供更多的机会。

利益冲突声明：作者声明本研究不存在任何经济或非经济利益冲突。

参考文献

1 Volkmar FR. Encyclopedia of autism spectrum disorders[M]. Cham: Springer International Publishing, 2021: 1-691.

2 Zhou H, Xu X, Yan W, et al. Prevalence of autism spectrum disorder in China: a nationwide multi-center population-based study among children aged 6 to 12 years[J]. Neurosci Bull, 2020, 36(9): 961-971. DOI: 10.1007/s12264-020-00530-6.

3 Mottron L, Bzdok D. Autism spectrum heterogeneity: fact or artifact?[J]. Mol Psychiatry, 2020, 25(12): 3178-3185. DOI: 10.1038/s41380-020-0748-y.

4 李雨田，郭小芳，梁连娇. 孤独症谱系障碍早期诊断与早期

干预模式探讨 [J]. 中国现代药物应用 , 2015, 9(14): 260-262. DOI: 10.14164/j.cnki.cn11-5581/r.2015.14.180.

5 李丹丹 , 王鸿武 , 吴毅 , 等 . 孤独症谱系障碍筛查工具的演变和应用研究进展 [J]. 环境与健康杂志 , 2025, 42(2): 182-187. [Li DD, Wang HW, Wu Y, et al. Evolution and application of screening tools for autism spectrum disorder: a review of recent studies[J]. Environment & Health, 2025, 42(2): 182-187.] DOI: 10.16241/j.cnki.1001-5914.2025.02.017.

6 Tomiyama S, Yoshida K, Tani H, et al. Pharmacological treatment of autism spectrum disorder: a systematic review of treatment guidelines[J]. Pharmacopsychiatry, 2025, 58(3): 100-116. DOI: 10.1055/a-2514-4452.

7 Dawson G, Rieder AD, Johnson MH. Prediction of autism in infants: progress and challenges[J]. Lancet Neurol, 2023, 22(3): 244-254. DOI: 10.1016/S1474-4422(22)00407-0.

8 赵宇曦 . 基于蛋白质组学和代谢组学的自闭症血浆生物标志物研究 [D]. 广东深圳 : 深圳大学 , 2020.

9 徐宇航 . 基于图神经网络和 MRI 影像的自闭症辅助诊断方法研究 [D]. 成都 : 电子科技大学 , 2024. DOI: 10.27005/d.cnki.gdzku.2024.000861.

10 韩俊林 , 林玉梅 , 林堃 , 等 . 孤独症谱系障碍儿童的肠道菌群特点及其对患儿行为的影响 [J]. 中外医学研究 , 2023, 21(34): 63-67. [Han JL, Lin YM, Lin K, et al. Characteristics of intestinal flora in children with autism spectrum disorder and its influence on children's behavior[J]. Chinese and Foreign Medical Research, 2023, 21(34): 63-67.] DOI: 10.14033/j.cnki.cfmr.2023.34.016.

11 张彩云 , 候芳 , 陈艳琳 , 等 . 家系全外显子测序揭示自闭症谱系障碍中的关键遗传变异和生物通路 [C]// 深圳市康复医学会 , 香港职业治疗学院 , 大湾区康复医学会 , 等 . 2024 年深圳国际康复论坛 (第二十一届) 优秀摘要集 , 2024: 7-8. DOI: 10.26914/c.cnkihy.2024.026434.

12 龙智平 , 王帆 . 多组学整合分析的设计及统计方法在肿瘤流行病学研究中的应用 [J]. 中华流行病学杂志 , 2020, 41(5): 788-793. [Long ZP, Wang F. Study design and statistical methods used for integrative analysis on multi-omics in cancer epidemiology[J]. Chinese Journal of Epidemiology, 2020, 41(5): 788-793.] DOI: 10.3760/cma.j.cn112338-20190624-00461.

13 Choi JM, Chae H. moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks[J]. BMC Bioinformatics, 2023, 24(1): 169. DOI: 10.1186/s12859-023-05273-5.

14 Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 7132-7141.

15 Nichol A, Achiam J, Schulman J. On first-order Meta-learning algorithms[EB/OL]. (2018-04-04) [2025-04-25]. https://doi.org/10.48550/arXiv.1803.02999.

16 American Psychiatric Association. Diagnostic and statistical manual of mental disorders, DSM-5-TR[M]. Arlington, VA: American Psychiatric Publishing, 2022: 56-68.

17 Krug DA, Arick J, Almond P. Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior[J]. J Child Psychol Psychiatry, 1980, 21(3): 221-229. DOI: 10.1111/j.1469-7610.1980.tb01797.x.

18 Schopler E, Reichler RJ, DeVellis RF, et al. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS)[J]. J Autism Dev Disord, 1980, 10(1): 91-103. DOI: 10.1007/BF02408436.

19 谢堃 , 陈雯 , 陈锋 , 等 . 一种唾液采集装置 : 中国专利 , 221770063U[P]. 2024-09-27.

20 Hu S, Loo JA, Wong DT. Human saliva proteome analysis[J]. Ann Ny Acad Sci, 2007, 1098(1): 323-329. DOI: 10.1196/annals.1384.015.

21 Mota FSB, Nascimento KS, Oliveira MV, et al. Potential protein markers in children with Autistic Spectrum Disorder (ASD) revealed by salivary proteomics[J]. Int J Biol Macromol, 2022, 199: 243-251. DOI: 10.1016/j.ijbiomac.2022.01.011.

22 Dame ZT, Aziat F, Mandal R, et al. The human saliva metabolome[J]. Metabolomics, 2015, 11: 1864-1883. DOI: 10.1007/s11306-015-0840-5.

23 李圆圆 . 基于质谱的微量样品代谢组与蛋白质组新方法研究 [D]. 沈阳 : 东北大学 , 2022. DOI: 10.27007/d.cnki.gdbeu.2022.003435.

24 Wei R, Wang J, Su M, et al. Missing value imputation approach for mass spectrometry-based metabolomics data[J]. Sci Rep, 2018, 8(1): 663. DOI: 10.1038/s41598-017-19120-0.

25 Yang H, Chen R, Li D, et al. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data[J]. Bioinformatics, 2021, 37(16): 2231-2237. DOI: 10.1093/bioinformatics/btab109.

26 Shehab M, Abualigah L, Shambour Q, et al. Machine learning in medical applications: A review of state-of-the-art methods[J]. Comput Biol Med, 2022, 145: 105458. DOI: 10.1016/j.compbiomed.2022.105458.

27 Levy S, Duda M, Haber N, et al. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism[J]. Mol Autism, 2017, 8: 65. DOI: 10.1186/s13229-017-0180-6.

28 Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark[J]. Nucleic Acids Res, 2018, 46(20): 10546-10562. DOI: 10.1093/nar/gky889.

29 Wang Y, Wang Z, Yu X, et al. MORE: a multi-omics data-driven hypergraph integration network for biomedical data classification and biomarker identification[J]. Brief Bioinform, 2025, 26(1): bbae658. DOI: 10.1093/bib/bbae658.

30 Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification[J]. Pattern Recognit. Lett, 2009, 30(1): 27-38. DOI: 10.1016/j.patrec.2008.08.010.

31 Ben-Sasson A, Guedalia J, Ilan K, et al. Predicting autism traits from baby wellness records: a machine learning approach[J]. Autism, 2024, 28(12): 3063-3077. DOI: 10.1177/13623613241253311.

32 Jones W, Klaiman C, Richardson S, et al. Eye-tracking-based

measurement of social visual engagement compared with expert clinical diagnosis of autism[J]. JAMA, 2023, 330(9): 854-865. DOI: 10.1001/jama.2023.13295.

33 Jiao Y, Chen R, Ke X, et al. Single nucleotide polymorphisms predict symptom severity of autism spectrum disorder[J]. J Autism Dev Disord, 2012, 42(6): 971-983. DOI: 10.1007/s10803-011-1327-5.

34 Ottosson F, Russo F, Abrahamsson A, et al. Unraveling the metabolomic architecture of autism in a large Danish population-based cohort[J]. BMC Med, 2024, 22(1): 302. DOI: 10.1186/s12916-024-03516-7.

35 Szoko N, McShane AJ, Natowicz MR. Proteomic explorations of autism spectrum disorder[J]. Autism Res, 2017, 10(9): 1460-1469. DOI: 10.1002/aur.1803.

36 Guo LY, Wu AH, Wang YX, et al. Deep learning-based ovarian cancer subtypes identification using multi-omics data[J]. BioData Min, 2020, 13(1): 10. DOI: 10.1186/s13040-020-00222-x.

37 Hira MT, Razzaque MA, Angione C, et al. Integrated multi-omics analysis of ovarian cancer using variational autoencoders[J]. Sci Rep, 2021, 11(1): 6265. DOI: 10.1038/s41598-021-85285-4.

38 Shao W, Zuo Y, Shi YY, et al. Characterizing the survival-associated interactions between tumor-infiltrating lymphocytes and tumors from pathological images and multi-omics data[J]. IEEE Trans Med Imaging, 2023, 42(10): 3025-3035. DOI: 10.1109/TMI.2023.3274652.

39 Pan L, Wang X, Liang Q, et al. DEDUCE: Multi-head attention decoupled contrastive learning to discover cancer subtypes based on multi-omics data[J]. Comput Methods Programs Biomed, 2024, 257: 108478. DOI: 10.1016/j.cmpb.2024.108478.

40 Sun Q, Cheng L, Meng A, et al. SADLN: self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition[J]. Front Genet, 2023, 13: 1032768. DOI: 10.3389/fgene.2022.1032768.

41 Wen G, Li L. MMOSurv: Meta-learning for few-shot survival analysis with multi-omics data[J]. Bioinformatics, 2024, 41(1): btae684. DOI: 10.1093/bioinformatics/btae684.

42 Huang HH, Shu J, Liang Y. MUMA: a multi-omics Meta-learning algorithm for data interpretation and classification[J]. IEEE J Biomed Health Inform, 2024, 28(4): 2428-2436. DOI: 10.1109/JBHI.2024.3363081.

43 Finn C, Abbeel P, Levine S. Model-agnostic Meta-learning for fast adaptation of deep networks[C]. International conference on machine learning. PMLR, 2017: 1126-1135.

44 Geschwind DH, Levitt P. Autism spectrum disorders: developmental disconnection syndromes[J]. Curr Opin Neurobiol, 2007, 17(1): 103-111. DOI: 10.1016/j.conb.2007.01.009.

45 Lord C, Elsabbagh M, Baird G, et al. Autism spectrum disorder[J]. Lancet, 2018, 392(10146): 508-520. DOI: 10.1016/S0140-6736(18)31129-2.

46 Heinsfeld AS, Franco AR, Craddock RC, et al. Identification of autism spectrum disorder using deep learning and the ABIDE dataset[J]. Neuroimage Clin, 2018, 17: 16-23. DOI: 10.1016/j.nicl.2017.08.017.

47 He H, Garcia EA. Learning from imbalanced data[J]. IEEE Trans Knowl Data Eng, 2009, 21(9): 1263-1284. DOI: 10.1109/TKDE.2008.239.

48 Wang S, Wang S, Wang Z. A survey on multi-omics-based cancer diagnosis using machine learning with the potential application in gastrointestinal cancer[J]. Front Med (Lausanne), 2023, 9: 1109365. DOI: 10.3389/fmed.2022.1109365.

49 Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis, 2nd edition[M]. Cham: Springer, 2015: 63-102.

收稿日期：2024 年 12 月 24 日 修回日期：2025 年 04 月 30 日
本文编辑：杨 燕 周璐敏