



Noninvasive detection and interpretation of gastrointestinal diseases by collaborative serum metabolite and magnetically controlled capsule endoscopy



Xiang-Tian Yu ^{a,*}, Ming Chen ^{b,1}, Jingyi Guo ^a, Jing Zhang ^b, Tao Zeng ^{c,*}

^a Clinical Research Center, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China

^b Department of Gastroenterology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China

^c Guangzhou Laboratory, Guangzhou, China

ARTICLE INFO

Article history:

Received 15 April 2022

Received in revised form 15 September 2022

Accepted 2 October 2022

Available online 6 October 2022

Keywords:

Gastrointestinal diseases

Noninvasive detection

Serum metabolite

Magnetically controlled capsule endoscopy

Machine learning

Convolutional neural networks

ABSTRACT

Gastrointestinal diseases are complex diseases that occur in the gastrointestinal tract. Common gastrointestinal diseases include chronic gastritis, peptic ulcers, inflammatory bowel disease, and gastrointestinal tumors. These diseases may manifest a long course, difficult treatment, and repeated attacks.

Gastroscopy and mucosal biopsy are the gold standard methods for diagnosing gastric and duodenal diseases, but they are invasive procedures and carry risks due to the necessity of sedation and anesthesia. Recently, several new approaches have been developed, including serological examination and magnetically controlled capsule endoscopy (MGCE). However, serological markers lack lesion information, while MGCE images lack molecular information.

This study proposes combining these two technologies in a collaborative noninvasive diagnostic scheme as an alternative to the standard procedures. We introduce an interpretable framework for the clinical diagnosis of gastrointestinal diseases. Based on collected blood samples and MGCE records of patients with gastrointestinal diseases and comparisons with normal individuals, we selected serum metabolite signatures by bioinformatic analysis, captured image embedding signatures by convolutional neural networks, and inferred the location-specific associations between these signatures.

Our study successfully identified five key metabolite signatures with functional relevance to gastrointestinal disease. The combined signatures achieved discrimination AUC of 0.88. Meanwhile, the image embedding signatures showed different levels of validation and testing accuracy ranging from 0.7 to 0.9 according to different locations in the gastrointestinal tract as explained by their specific associations with metabolite signatures. Overall, our work provides a new collaborative noninvasive identification pipeline and candidate metabolite biomarkers for image auxiliary diagnosis. This method should be valuable for the noninvasive detection and interpretation of gastrointestinal and other complex diseases.

© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Gastrointestinal diseases occur in various parts of the gastrointestinal tract. Common gastrointestinal diseases include chronic gastritis, peptic ulcers, inflammatory bowel disease, and gastrointestinal tumors. Gastrointestinal diseases often are associated with abdominal pain, diarrhea, hematochezia, changes in stool charac-

teristics, and other symptoms, and there are often varying degrees of congestion, erosion, ulcers, bleeding, and other manifestations under endoscopy. Genetic susceptibility, epithelial barrier defects, immune response disorders, and environmental factors play important roles in the pathogenesis of the gastrointestinal tract [1,2]. Gastrointestinal diseases may manifest a long course, difficult treatment, and repeated attacks. Chronic atrophic gastritis (CAG) is an established precursor of gastric cancer (GC), which has high morbidity [3]. The stages in the progression of gastric cancer include CAG, intestinal metaplasia (IM), and dysplasia [4]. The small intestine includes the duodenum, jejunum, and ileum. Diseases of the small intestine are important components of digestive tract diseases [5]. Inflammatory bowel disease (IBD) is a set of

* Corresponding authors at: Clinical Research Center, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Yishan Road 600, Shanghai, China (X.-T. Yu); Guangzhou Laboratory, Guangzhou, China (T. Zeng).

E-mail addresses: graceyu1985@163.com (X.-T. Yu), zengtao@sibs.ac.cn, zeng_tao@gzlab.ac.cn (T. Zeng).

¹ Authors contributed equally to this work.

chronic relapsing diseases that includes Crohn's disease (CD, affecting the terminal ileum) and ulcerative colitis (UC, affecting the colon and rectum). The prevalence of IBD has steadily increased in newly industrialized countries and Western countries [6,7].

Gastroscopy and mucosal biopsy are the gold standard methods for diagnosing these gastric and duodenal diseases, but they are invasive procedures and carry risks of sedation and anesthesia [8]. The accuracy of such tests largely depends on the physician's skill in obtaining biopsy samples [3,9]. Meanwhile, double balloon enteroscopy (DBE) allows complete visualization, biopsy, and treatment of the small bowel [10]. However, this technology is time-consuming and also carries the risk of small intestinal perforation [11]. Therefore, noninvasive diagnoses are desirable for clinical practice, and they have played an increasingly important role in the diagnosis and treatment of gastrointestinal diseases in recent years. In particular, the development and application of new noninvasive technologies for early diagnosis of gastrointestinal diseases presents a serious challenge, one that needs to be solved with the help of biological and computational methods.

Serological examinations can provide effective indicators of gastrointestinal diseases. They are noninvasive and relatively convenient compared to markers in mucosal biopsies and in body fluids. Markers of gastric function such as PG-I, PG-II, PG-I/PG-II ratio (PGR), gastrin 17 (G-17), anti-*Helicobacter pylori* antibodies (HP-IgG) are used to jointly diagnose chronic atrophic gastritis [9,12]. $PGR \leq 3$ and $PG-I \leq 70$ ng/ml have been widely applied for CAG and GC prediction [13,14]. Reese et al. found that the combination of ASCA+/pANCA– had a specificity of 92.8 % and a sensitivity of 54.6 % for CD [15]. Pavlidis et al. illustrated that anti-MZGP2 antibodies were detected in 31 % of CD patients and that they had a high specificity for CD (96 %) [16,17]. In addition, studies have found that anti-granulocyte macrophage colony-stimulating factor (antiGM-CSF) antibodies also have higher concentrations in CD patients compared with healthy people [18]. The metabolites in serum have important effects on host physiology and can also be detected in many biological samples, including feces, urine, and cerebrospinal fluid [19,20]. The 1H NMR-based metabolomics with correlative analysis was recently developed to analyze the metabolic features of CAG. Cui et al. found that three plasma biomarkers (arginine, succinate, and 3-hydroxybutyrate) had the potential to indicate risks of CAG [21]. Several metabolites are associated with IBD and intestinal inflammation. For example, bile acid derivatives, short-chain fatty acids (SCFAs), and tryptophan metabolites are the focus of intense research [22]. Thus, serum metabolites should have generality for detecting early-warning signals of gastrointestinal disease.

Magnetically controlled capsule endoscopy (MGCE) is a noninvasive and safe technology with high clinical application value [23]. MGCE can be used for a complete examination of the entire stomach by actively controlling the movement of the capsule in the stomach through an external magnetic field. This also achieves highly accurate detection rate in the diagnosis of small intestine disease in patients who cannot undergo small intestine endoscopic examination. For example, MGCE has improved the clinical diagnosis of gastrointestinal diseases such as CG and small-bowel erosion [24], and it has diagnosed >60 % of cases of CG and small-bowel erosion in China [25]. Therefore, MGCE is capable of capturing the disease signals across the entire gastrointestinal tract.

Compared to the standards of gastroscopy and mucosal biopsy, serological markers lack lesion information, and MGCE images lack molecular information. Clearly, the combination of these two technologies would provide a more suitable collaborative diagnostic scheme as an alternative noninvasive method. Thus, this work introduces an interpretable framework for the clinical diagnosis of gastrointestinal diseases (Fig. 1). Based on our study enrollment, we collected blood samples (Fig. 1A) and MGCE records (Fig. 1B)

from patients with gastrointestinal diseases and compared these with normal individuals. The study workflow includes several steps (Fig. 1C): (i) detecting the serum metabolites from blood samples and selecting metabolite signatures related to gastrointestinal diseases by bioinformatic analysis; (ii) extracting the image features from MGCE records and capturing embedding signatures associated with gastrointestinal diseases by machine learning models based on convolutional neural networks; and (iii) inferring the location-specific association between serum metabolite signatures and MGCE image embedding signatures. Our study successfully identified five key metabolite signatures, i.e., Biliverdin, Estrone glucuronide, Tetrahydrocortisone, Jaceidin 4'-glucuronide, and PC (20:4(8Z,11Z,14Z,17Z)/14:0), whose combination achieved a discrimination AUC of 0.88. Meanwhile, the image embedding signatures showed different levels of validation and test accuracy ranging from 0.7 to 0.9 at different locations in the gastrointestinal tract. The results may be explained by their specific associations with key metabolite signatures. Collectively, our work provides a new collaborative noninvasive identification pipeline and identifies candidate metabolite biomarkers with image auxiliary diagnosis; the method should be valuable in the fight against gastrointestinal and other complex diseases.

2. Methods

2.1. Ethical approval of the study protocol

This study protocol was approved by the Ethics Committee of Sixth People's Hospital affiliated with Shanghai Jiao Tong University (Shanghai, China). Written informed consent was obtained from all individuals. The personal data were anonymized and omitted.

2.2. Study enrollment

This study ran at the Sixth People's Hospital from October 2020 to October 2021. Individuals who agreed to complete metabolite examination of serum samples and MGCE examination were recruited. The procedures of enrollment, serum metabolism, and MGCE classification were completed independently by different investigators who were blind to the results of each other's examinations.

2.3. Sample collection

Age and sex were recorded for every participant (Table S1). Serum samples were collected from each eligible individual and stored at -80 °C until analysis. The matched clinical diagnosis information was collected from the corresponding medical records in the hospital and included gastroscopic diagnosis, pathological diagnosis of gastroscopic biopsy, colonoscopic diagnosis, pathological diagnosis of colonoscopic biopsy, and capsule endoscopic diagnosis. The gastroscopic diagnosis that is the current standard was used to classify individuals into a normal group (NL group with seven samples) and a gastric and duodenal disease group (GD group with seven samples).

2.4. Metabolism

First, the metabolite extraction from our collected serum samples was carried out according to the following protocol: (i) 50 μ L of each sample was transferred to an EP tube; (ii) 200 μ L of extract solution was added (acetonitrile:methanol = 1:1, containing isotopically labeled internal standard mixture); (iii) samples were vortexed for 30 s, sonicated for 10 min in an ice-water

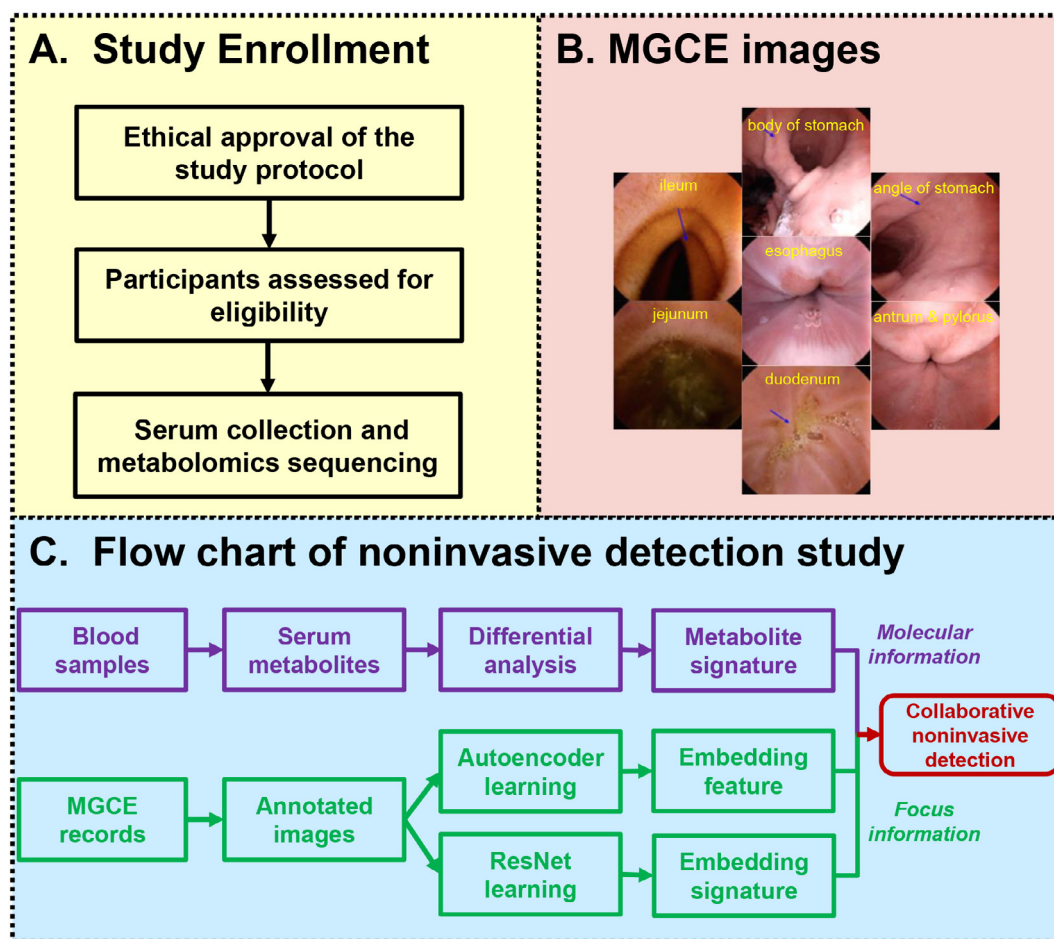


Fig. 1. Workflow of collaborative noninvasive detection for gastrointestinal diseases. A. The study enrollment designed in this study. B. Images and annotations from magnetically controlled capsule endoscopy. C. Study and analysis workflow used in this study, including (i) Blood sample collection and detection of serum metabolites, which supply metabolite signatures representing the molecular information related to GD; (ii) MGCE record collection and learning of image features, which produce embedding signatures representing focus information associated with GD; (iii) the inference of correlations between serum metabolite signatures and image embedding signatures, which support the collaborative noninvasive detection of GD.

bath, and incubated for one hour at -40°C to precipitate proteins; (iv) each sample was centrifuged at 12000 rpm ($\text{RCF} = 13800 \times g$, $R = 8.6\text{ cm}$) for 15 min at 4°C ; (v) the resulting supernatant was transferred to a fresh glass vial for analysis; and (vi) additional quality control (QC) samples were prepared by mixing equal aliquots of the supernatants from all samples.

Next, LC-MS/MS detection was executed on a UHPLC system (Vanquish, Thermo Fisher Scientific) with a UPLC BEH Amide column (Waters, Massachusetts, USA; $2.1\text{ mm} \times 100\text{ mm}$, $1.7\text{ }\mu\text{m}$) that was coupled to a Q Exactive HFX mass spectrometer (Orbitrap MS, Thermo). In this study, the mobile phase consisted of 25 ammonia hydroxide (mmol/L) in water ($\text{pH} = 9.75$) (A) and acetonitrile (B), and 25 mmol/L ammonium acetate. The auto-sampler temperature was 4°C , and the injection volume was $3\text{ }\mu\text{L}$. Here, the QE HFX mass spectrometer was adopted because it can acquire MS/MS spectra in information-dependent acquisition (IDA) mode in the control of the acquisition software (Xcalibur, Thermo), where the acquisition software continuously evaluated the full scan MS spectrum. The ESI source conditions were set at a capillary temperature of 350°C , a sheath gas flow rate of 30 Arb, collision energy as 10/30/60 in NCE (Normalized Collision Energy) mode, spray voltage as 3.6 kV (positive) or -3.2 kV (negative), auxiliary gas flow rate as 25 Arb, full MS resolution as 60000, and MS/MS resolution as 7500.

The raw data from the above detection were converted to the mzXML format using ProteoWizard. These data were further processed with an in-house program developed using R and based on XCMS that featured peak extraction, alignment, and integration. The metabolite annotation was applied to an in-house MS2 database (BiotreeDB) where the cutoff for annotation was set at 0.3.

2.5. Magnetically guided capsule endoscopy

All individuals (14 individuals involved in the above serum metabolite analysis and an additional 10 individuals involved in the MGCE analysis) underwent intestinal preparation with an electrolyte solution of polyethylene glycol, fasted all night, and completed MGCE examination (Ankon Medical Technologies, Shanghai, China) in the morning.

Gastric inflammation based on MGCE of each individual was identified based on the Updated Sydney System (Dixon et al., 1996). Each MGCE report recorded: (i) mucosal lesions such as erosions; (ii) changes of villi (flat mucosa, coarsened villi); (iii) lymphangiectasias/lymphocellular infiltrates; (iv) capillary lesions (angiodysplasias, petechiae); and (v) mucosal changes (erythema, edema, prominent mucosal folds). The collected medical images from each MGCE report were annotated with location information, including the esophagus, body of the stomach, angle of the stomach, antrum and pylorus, duodenum, jejunum, and ileum. Thus,

the entire MGCE image dataset could also be reorganized as seven location-specific image datasets for detailed analysis and discussion of potential tissue/lesion specificity of gastrointestinal diseases.

2.6. Bioinformatic analysis for serum metabolite signatures

For the two groups of participants, the clinical characteristics were analyzed between NL and GD groups (Table S1). The comparison of these clinical characteristics included signed-rank tests for continuous variables and chi-square tests for categorical variables; these were carried out in SAS 9.3 (SAS Institute, Cary, NC, USA). The clinical data were presented as mean \pm standard deviation (SD) or median (interquartile range) after normality testing for continuous variables. The level of significance was set at $P < 0.05$.

To analyze changes in metabolism between NL and GD groups, univariate and multivariate analyses were conducted, including differential expression analysis using a *t*-test, fold-change with a volcano plot, principal component analysis (PCA), partial least square discriminant analysis (PLS-DA), and orthogonal partial least square discriminant analysis (OPLS-DA). For screening differentially expressed metabolites (DEMs), the univariate analysis used the standards of $P < 0.05$ and $|\log_2FC| > 0$, and the multivariate analysis used a standard of variable importance in the projection (VIP) > 1 . Combining the assessments from the *t*-test, fold-change, and VIP score, a set of key DEMs were selected as metabolite signatures for discriminating GD samples from NL samples. Then, according to the co-expression network, the local network (i.e., neighboring pattern) of each key DEM was identified as a key DEM module. For evaluating the biological significance of DEMs and DEM modules, an enrichment analysis was performed in the small molecule pathway database (SMPDB) and the Kyoto Encyclopedia of Genes and Genomes database (KEGG), with $P < 0.05$ regarded as the significance level. Finally, the predictive efficiency of the selected key DEMs was analyzed using random forest (RF), and the area under the curve (AUC) for receiver operating characteristic (ROC) curves was used for assessing the RF performance. All metabolite data processing and analysis were performed using MetaboAnalyst [26].

2.7. Convolutional neural network analysis for image embedding signatures

As an unsupervised artificial neural network, autoencoder (AE) applies back-propagation by setting the output values equal to the input values. AE can nonlinearly transform the data into a low-dimensional latent space, as AE can force the neural networks to compress the high-dimensional data into a low-dimensional representation that captures the nonlinear relationships in the original data [27,28]. The hidden layers of AE can be thought of as some abstract representation of the input (e.g., image) [29]. For image data such as MGCE reported in this study, a convolutional autoencoder (CAE) is widely used in the feature extraction of images, being a special variant of basic AE [30]. Here, a CAE was implemented and applied in PyTorch. The CAE consisted of an input layer, a convolutional layer, a flattening layer, a deconvolutional layer, and an output layer (the structures are shown below). The internal output of flattening layers was used as the embedding features of input images from such unsupervised learning.

As a supervised artificial neural network, a typical deep residual network (ResNet) can directly infer high-level representations from low-level data by residual learning [31–33], and it can effectively solve the vanishing gradient problem. This work adopted the pre-trained ResNet-18 model from PyTorch for transfer learning and trained a new simplified model after network structure mod-

ification on our MGCE datasets (the structure is shown below). The last fully connected layer for softmax and final binary prediction (i.e., normal vs gastrointestinal disease) was used to produce the embedding signatures of each input image from such supervised learning. The performance evaluation of the learning model was the average value of 100 times the accuracy (ACC) on training data (randomly selected 60 % of the samples), validation data (randomly selected 20 % of the samples), and test data (the remaining 20 % of the samples).

3. Results and discussion

3.1. Metabolite signatures associated with gastrointestinal diseases

By comparing the metabolite profiles between NL and GD groups, the metabolite spectrum demonstrated the potential non-linear discrimination between the two group samples in the 2D space of the PCA (Fig. 2A). The hierarchical clustering also suggested the sample clusters corresponding to two groups (Fig. 2B). After the supervision, PLS-DA improved the metabolite discrimination between NL and GD groups (Fig. 2C) based on several important metabolites with high VIP scores (Fig. 2D). Meanwhile, the statistical analysis of metabolite differential expression supplied some DEMs with significant *P* values and large fold-changes, as shown in the volcano plot (Fig. 2E). These DEMs showed that in gastrointestinal diseases, many metabolites have been significantly down-regulated (Fig. 2F), suggesting the functional loss during disease occurrence and development. Taken together, five metabolite signatures were selected according to their contributions observed in all analyses; these were Biliverdin, Estrone glucuronide, Tetrahydrocortisone, Jaceidin 4'-glucuronide, and PC (20:4(8Z,11Z,14Z,17Z)/14:0). Biliverdin and bilirubin have antioxidant properties and thus can effectively scavenge free radicals and inhibit lipid peroxidation [34]; bilirubin, as a powerful antioxidant, is oxidized to biliverdin during the cycle [35,36].

Several metabolite signatures have been reported concerning their specific roles in gastrointestinal diseases. For example, biliverdin reduced cyclooxygenase 2 and the expression of the inflammatory factors IL-6 and IL-1 β mRNA in a rat small intestine transplantation model and reduced the infiltration of neutrophils into jejunum muscle layer, thereby having a protective effect on the intestine [37]. Heme oxygenase (HO) catalyzes the degradation of toxic free heme to biliverdin and Fe²⁺ and also releases carbon monoxide (CO) [38]. The Heme-Oxygenase 1 (Ho-1)/Biliverdin/CO Pathway protected against ethanol-induced gastric injury in mice through a co-dependent (CO) or biliverdin-independent mechanism [39]. Estrone-3-glucuronide can be activated by intestinal microbial β -glucuronidase (GUS) during the promotion and development of cancer [40].

3.2. Functional relevance of metabolite signatures

For the above DEMs, their functional enrichments indicated certain pathogen relevance (Fig. 3A). In particular, the five metabolite signatures were involved in co-expression modules (Fig. 3B), indicating their different functional roles. For example, Biliverdin, Jaceidin 4'-glucuronide, and PC (20:4(8Z,11Z,14Z,17Z)/14:0) were co-expressed and down-regulated in disease states (Fig. 3C). In contrast, Estrone glucuronide and Tetrahydrocortisone were in another co-expression module (Fig. 3B) and were up-regulated in disease conditions (Fig. 3C). Thus, their co-expressed neighboring patterns, i.e., DEM modules, were further captured (Fig. 3D), and the corresponding partner metabolites were combined for functional enrichment analysis (Fig. 3E).

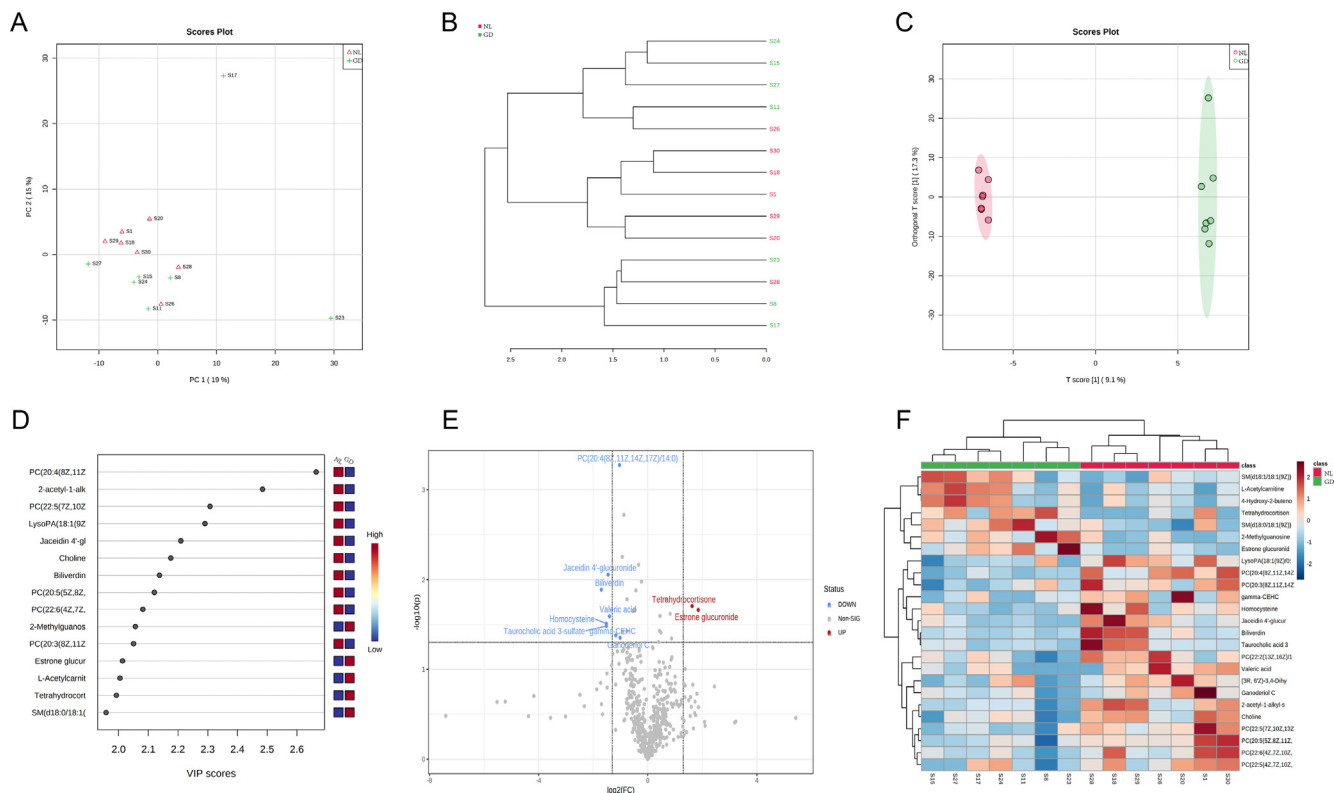


Fig. 2. Differential analysis of serum metabolites involved in gastrointestinal diseases. A. The PCA plot of metabolite profiles from disease and normal groups. B. Hierarchical clustering of samples based on their metabolite abundances. C. PLSDA plot of metabolite profiles discriminating GD and NL groups. D. Important metabolites ranked by VIP scores from PLSDA. E. Differentially expressed metabolites selected by volcano plot based on their differential significance and fold change. F. The expression heatmap of DEMs.

Considering the function in SMPDB as an example (Fig. 3D), B vitamins and gene polymorphisms that encode single-carbon metabolic enzymes may affect DNA synthesis and methylation and thus be associated with cancer. In a case-control study of the European Prospective Investigation into Cancer and Nutrition cohort, vitamin B6 species were measured in plasma. The adjusted relative risk per quartile (95 % confidence interval, P(trend)) was 0.78 (0.65–0.93, < 0.01) for vitamin B6. The relation was strong in individuals with severe chronic atrophic gastritis. The results showed a significant negative association between vitamin B6 and gastric cancer risk. This conclusion was clearer for atrophic gastritis [41]. Supplementation with sodium hydroxide (NaHS) and vitamin B6 (VB6) can partially reverse microbial dysregulation and has therapeutic potential for stress gastritis [42]. Vitamin B6 inhibited TNF- α -induced NF- κ B activation by inhibiting I κ B α degradation in human colon cancer HT-29 cells [43].

Indeed, many functions in the KEGG database were enriched in these DEM modules (Table S2). For example, synthesis and degradation of ketone bodies were enriched in the DEM module induced by PC (20:4(8Z,11Z,14Z,17Z)/14:0). Butyric acid provides more than 90 % of the total energy requirements of colon cells. The energy of intestinal mucosal epithelial cells comes from bacterial fermentation products, the most important of which is butyric acid. Butyric acid is absorbed directly by colon epithelial cells and rapidly forms ketone bodies for ATP synthesis. Impaired butyric acid oxidation was observed in the colonic mucosa of patients with ulcerative colitis. In addition, impaired butyric acid transport and oxidation were evident from the gene expression levels [44,45]. Chia et al. found that in the mammalian small intestine, the expression of Hmgcs2 (3-hydroxy-3-methylglutaryl-CoA synthetase 2), the gene encoding the rate-limiting enzyme in the production of ketone bodies, distinguishes self-renewing Lgr5 + stem

cells (ISCs) from other cell types. Ketone body signal transduction can mediate intestinal stem cell homeostasis [46]. Haruka et al. showed that the microtubule hyperacetylation induced by ketone bodies may be a causal factor linking diabetes to colorectal cancer [47].

Glycerophospholipid metabolism is enriched in the DEM module induced by Estrone glucuronide. Palmatine can restore the body function of Chronic Atrophic Gastritis in rats and reduce gastric mucosa damage. Metabolomics analysis showed that the therapeutic effect of this drug on CAS was primarily realized through the glycerophospholipid metabolic pathway [48]. Huang-Lian-Jie-du formula (HLJDD) and its effective fraction may inhibit the expression of COX-2 protein and the activities of PLA2 and 5-LOx. Inhibition of the arachidonic acid metabolic pathway and the glycerophospholipid metabolic pathway could alleviate acute ulcerative colitis [49]. A cohort study in Italy conducted lipid and polar profiling of plasma samples from 200 individuals with IBD and healthy individuals. The changes in phosphatidylcholine, fatty acids, and glycerophospholipids in pathological specimens were significant. In addition, decreased amino acid levels suggest mucosal damage in IBD [50].

Linoleic acid metabolism has significant enrichment in many DEM modules. It was found that activation of colon PPAR G by conjugated linoleic acid (CLA) could mediate the protective effect against experimental IBD in mice [51]. Under CLA treatment or a cLA-rich diet, the expression and activity of PPAR G in colon mucosa [52] and macrophages were increased [53]. Danoyo et al. found that linoleic acid epigenetic modification of the Farnesoid-X-receptor (FXR) led to the activation of downstream factors involved in bile acid homeostasis and induced epigenetic changes related to colon inflammation and cancer [54]. Conjugated linoleic acid (CLA) can prevent intestinal mucositis induced by 5-

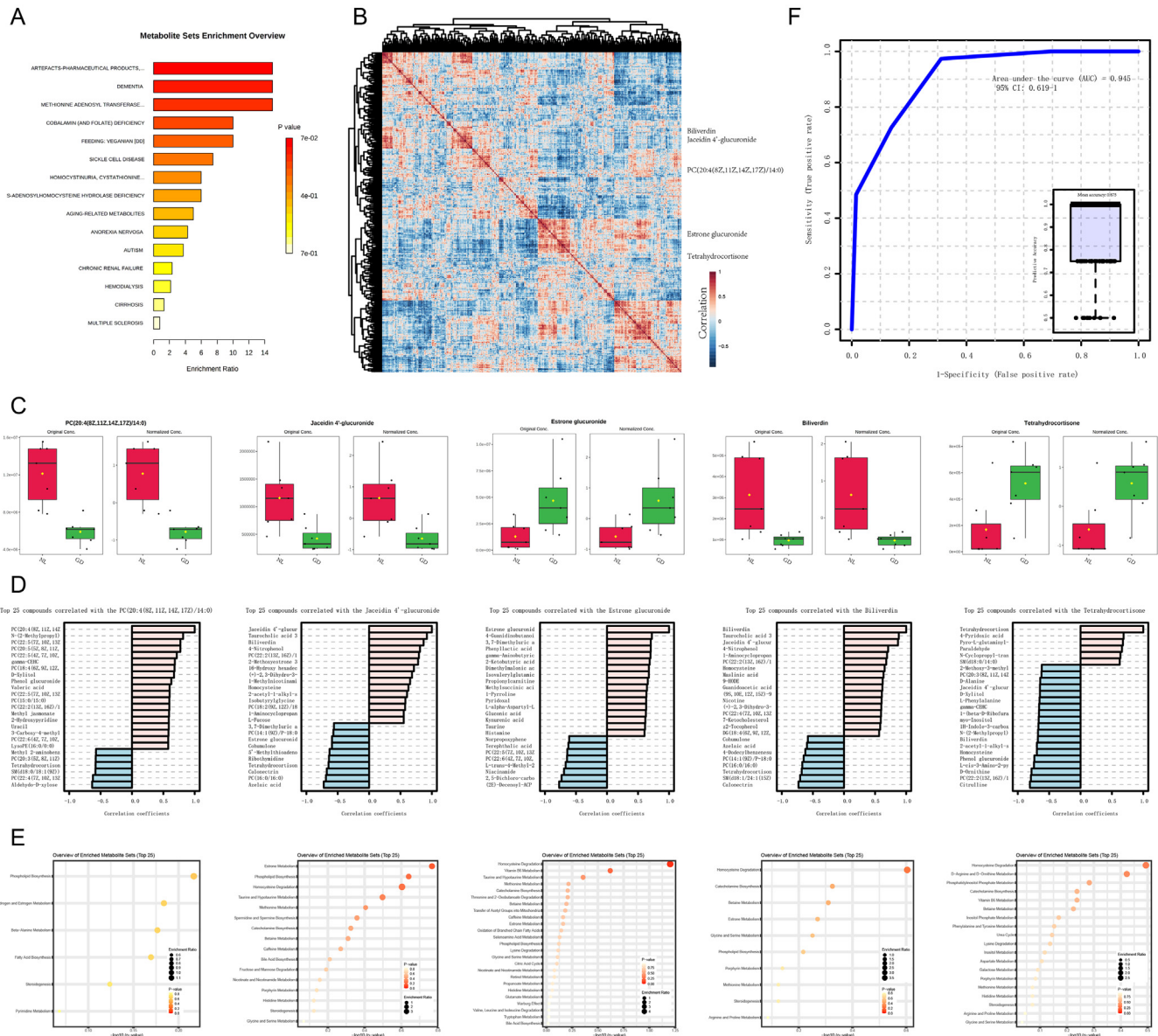


Fig. 3. Functional analysis of key metabolite signatures relevant to gastrointestinal diseases. A. The functional enrichment of all DEMs. B. The co-expression pattern of serum metabolites and five metabolite signatures. C. The detailed abundance change of metabolite signatures between disease and normal groups. D. The co-expression module of each metabolite signature, which contains a list of neighbouring partner metabolites on co-expression network. E. The enriched functions of each DEM module. F. The predictive performance of the combination of five metabolite signatures based on Randomforest model.

fluorouracil, and CLA treatment maintained intestinal epithelial integrity and a good balance between inflammatory and regulatory cytokines [55].

In addition to the biological significance, the combination of the above five metabolite signatures indicated their high efficiency and potential in distinguishing a disease state from a normal state, achieving an average AUC of approximately 0.94 in a cross-validation manner and an average accuracy of approximately 0.885 in multiple replications (Fig. 3F) for the RF model calculated in MetaboAnalyst [26].

3.3. Embedding features of MGCE and their association with metabolite signatures

The metabolite signatures were detected in serum so that they could be associated with the disease phenotype observed in specific tissues, e.g., images characterized by MGCE at different loca-

tions of the gastrointestinal tract. Thus, embedding features, the quantitative characteristics of MGCE images, were first extracted by auto-encoder technology. The encoder network transforms or reduces the original image data with pixel features at the input layer into embedding data with vector features at the hidden layer (implemented as shown in Fig. 4A). The decoder network transforms or reconstructs such embedding data into recovered data at the output layer (implemented as shown in Fig. 4B). Due to the constraint of the auto-encoder, the recovered data will be the same as the original data (examples are shown in Fig. 4C and D), and thus the embedding data can represent the essential information contained in the original data. Then, each embedding feature will have a score vector across all samples (the scores of images from the same individual can be averaged), so that the association between metabolite signatures and embedding features can be estimated. This provides a new way to explain the observed clinical image characteristics by molecular (metabolic) features. In the

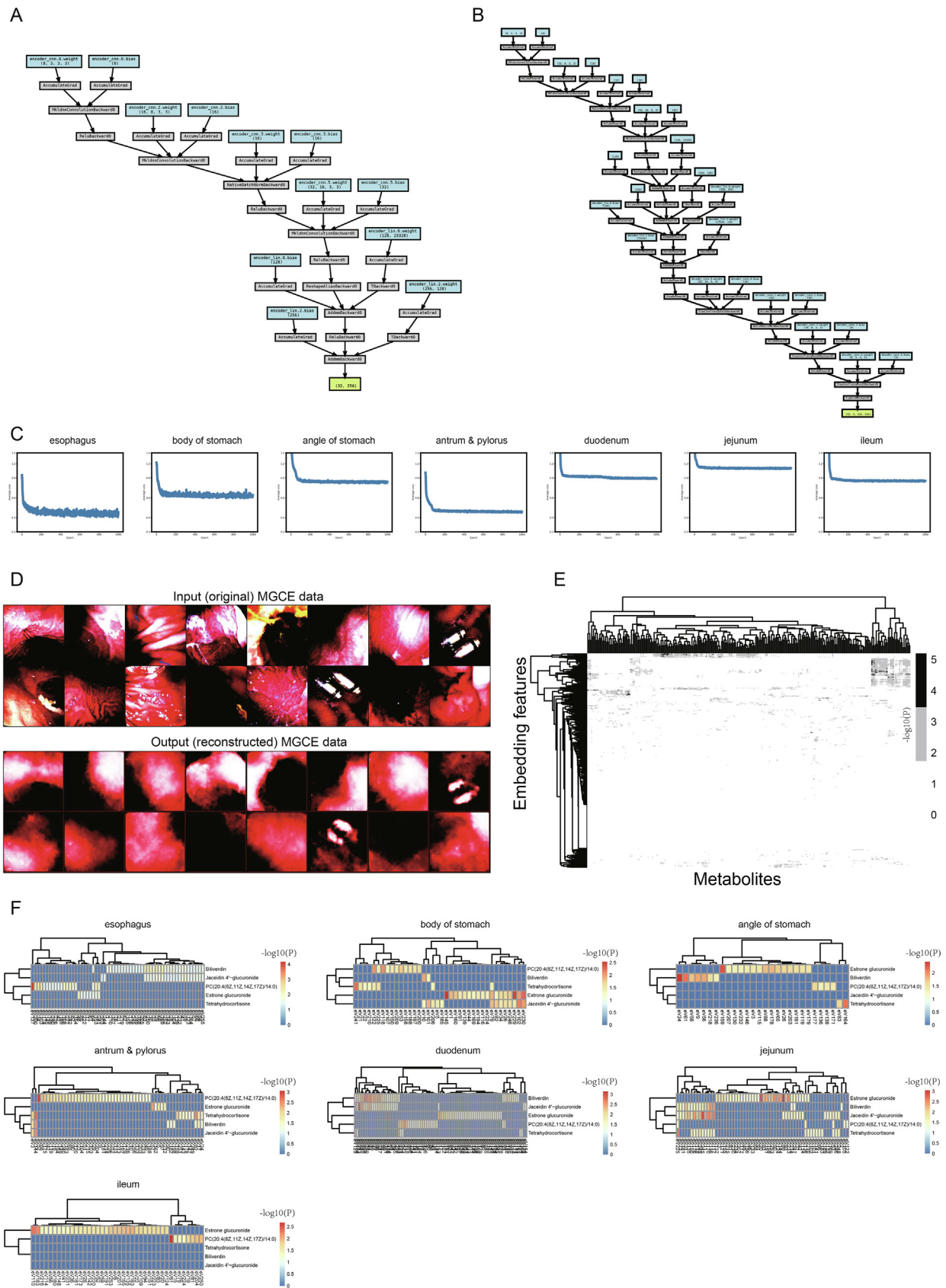


Fig. 4. Image embedding features from autoencoding model and their associations with metabolite signatures. A. The network structure of encoder produced by torchviz package. B. The network structure of decoder produced by torchviz package. C. The learning procedure of autoencoding models corresponding to different locations in gastrointestinal tract, measured by loss index. D. The example of input original data and output recovered data by such unsupervised learning. E. The association matrix (with $-\log_{10}(P)$ values) between abundance vector of serum metabolites and score vector of embedding features. F. The location specific association sub-matrix (with $-\log_{10}(P)$ values) for five metabolite signatures, where each row represents a key metabolite and each column represents one embedding feature.

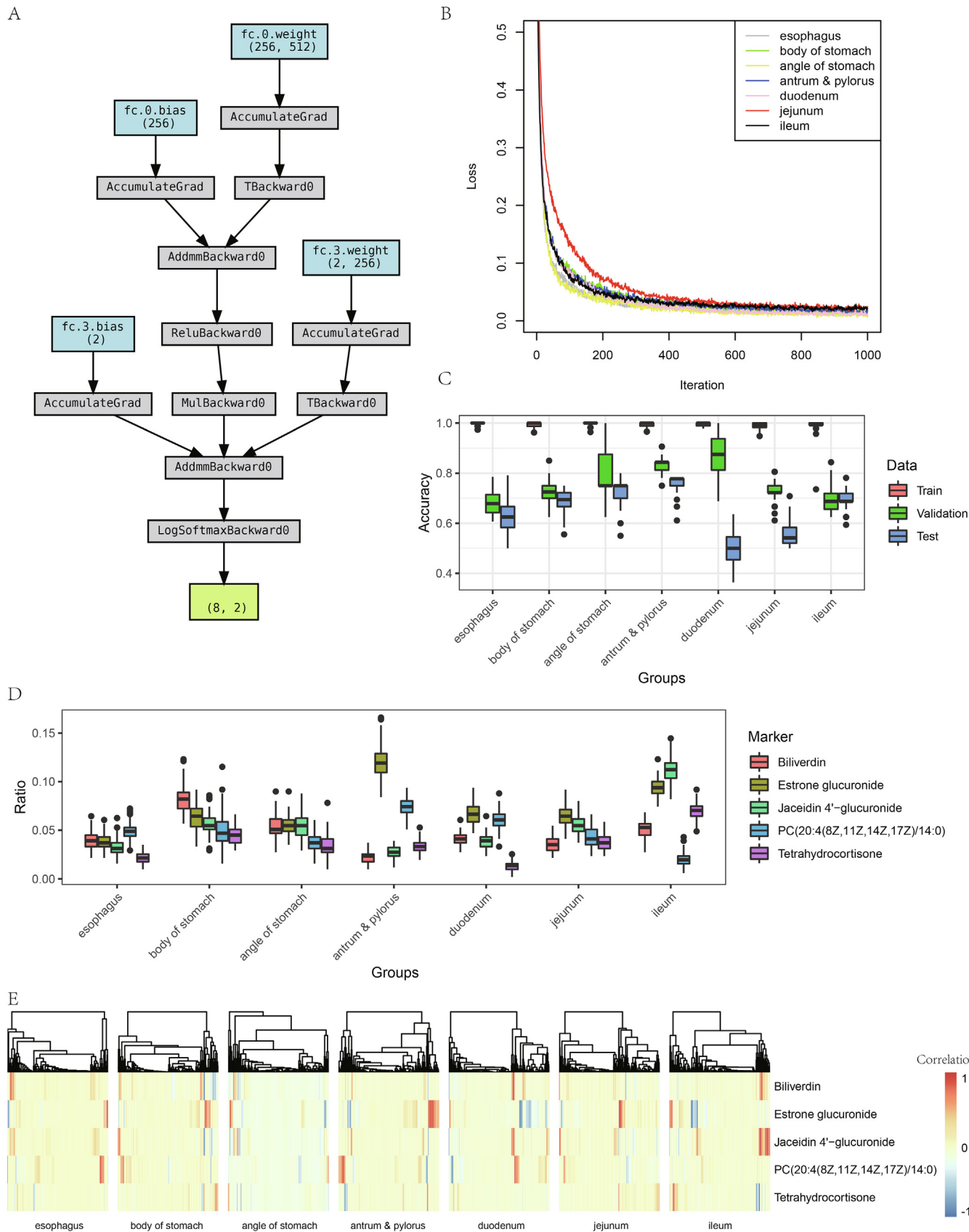


Fig. 5. Image embedding signatures from residue network model and their associations with metabolite signatures. A. The network structure of simplified ResNet produced by torchviz package. B. The learning procedure of ResNet models corresponding to different locations of gastrointestinal tract. C. The performances of location specific models evaluated by Accuracy on train, validation and test datasets respectively. D. The location specific distribution of the number of embedding signatures associated with each metabolite signature. E. The location specific association sub-matrix (with correlation values) between embedding signatures and metabolite signatures.

association matrix illustrated in Fig. 4E, a group of embedding features has significant correlations with a subset of metabolites, indicating the ability of molecular indicators to explain the image characteristics.

For the above five recognized metabolite signatures, each could be correlated with or explain several embedding features at different locations of the gastrointestinal tract (Fig. 4F). For example, Estrone glucuronide displayed an association with embedding features at many locations across the gastrointestinal tract, suggesting its potential universal participation in gastrointestinal diseases. Of note, according to the annotation from HMDB (Table S3), Estrone glucuronide exists in many tissues, including the intestine, kidney, liver, and pancreas (Table S4). Thus, it is reasonable to assume that this key metabolite has a location-specific association within the gastrointestinal tract. In contrast, PC (20:4(8Z,11Z,14Z,17Z)/14:0) demonstrated a certain association preference within the gastrointestinal tract, although it was reported to exist in all tissues by HMDB (Table S4), indicating the potential functional tissue-specificity of this metabolite feature. In addition, biliverdin is said to exist in neurons and prostate by HMDB (Table S4), but a previous literature report stated that biliverdin can selectively modulate the inflammatory cascade involved in intestinal muscularis function so as to attenuate morbidity to the intestine [56]. This suggested that our study has found new evidence that biliverdin has functional roles in the gastrointestinal tract and corresponding dysfunction.

3.4. Embedding signatures of MGCE relevant to gastrointestinal diseases

The above embedding features obtained by the unsupervised approach can explain the essential image information. Meanwhile, the discriminative image information on classes/phenotypes (e.g., disease vs normal) is also required to further explain the clinical diagnosis and corresponding metabolite signatures. Thus, the residue network (ResNet) (implemented as displayed in Fig. 5A) was used to learn such discriminative features, i.e., embedding signatures, from the same MGCE image data with additional group labels (i.e., GD vs NL in a binary class manner).

Similarly, at different locations across the gastrointestinal tract, the ResNet was taught using the training data (Fig. 5B) and assessed with validation data and test data (Fig. 5C). We could observe tissue specificity from these supervised models. The models achieved better performance for the esophagus, the body of the stomach, the angle of the stomach, and the antrum and pylorus. Meanwhile, they tended to have reduced test performance for the duodenum, jejunum, and ileum. These results supported the finding that the pathological information captured by MGCE indeed can match with the diagnosis from the standard gastroscopy and mucosal biopsy methods.

With effective learning, the embedding signatures were extracted to quantitatively measure the images from each individual. Indeed, many embedding signatures had a remarkable similarity to embedding features (Fig. S1), suggesting the consistency of feature extraction in unsupervised and supervised manners and the preferred features associated with targeted phenotypes (e.g., diseases). These embedding signatures were correlated with metabolite signatures, thus supplying new molecular explanations of these discriminative image characteristics.

The images and molecular associations had remarkable location specificity. Globally, Estrone glucuronide was associated with more embedding signatures than other signatures (Fig. 5D), indicating again its universal participation at different locations of the gastrointestinal tract. Locally, Estrone glucuronide further displayed a significant positive association with stomach-specific embedding signatures while having negative correlations with intestine-

specific embedding signatures, suggesting the possibility of functional specificity of molecules reflected by biomedical images (Fig. 5E).

4. Conclusions

This study aimed to propose a method for collaborative noninvasive identification of gastrointestinal diseases via serum metabolite signatures together with MGCE embedding signatures. The standard methods for diagnosing gastric and duodenal diseases in the current clinical application are gastroscopy and mucosal biopsy, procedures that are invasive and that carry high risks of sedation and anesthesia [8]. Thus, a new noninvasive approach or standard is urgently required. Routine blood testing is practical in clinics, so that blood-originated markers should be very suitable [57]. The serum metabolites have been widely studied and are thought to be indicative of many gastrointestinal diseases. However, with this type of method, it is difficult to identify the pathogenic condition of on-set tissues, a general issue in current noninvasive detection. Typical biomedical imaging technology can help reveal the disease signals from targeted tissue locations. In particular, the newly developed MGCE method can supply complete image information for nearly the entire gastrointestinal tract [58]. Thus, the combination of serum metabolites and MGCE imaging can provide effective (complementary) noninvasive identification of gastrointestinal diseases.

In conclusion, this work first recognized five serum metabolite signatures for distinguishing gastrointestinal diseases from normal controls. These metabolite signatures demonstrated their potential correlation with MGCE images by auto-encoder learning. Furthermore, they are also associated with embedding signatures representing the active region on images responsive to gastrointestinal disease identification, as shown by ResNet learning. The detection framework and analysis results provide a new noninvasive identification pipeline and metabolite biomarkers with image auxiliary diagnosis, making it possible to further collaborate with other candidate noninvasive approaches such as circulating cell-free DNA [59], gut metagenomics [60], or single circulating tumor cells [61,62] by multi-omics integration methods [63–68]. The continuous clinical study and application of such approaches are worthwhile for examining a large population for gastrointestinal diseases.

Funding

This study was supported by the National Natural Science Foundation of China (No. 11871456 and No. 61803360), the Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01), the Shanghai Jiao Tong University Affiliated Sixth People's Hospital Basic Scientific Research (No. ynms202118), and the Clinical Research Plan of SHDC (SHDC2022CRS044).

CRedit Author Contribution statement

X.-T Yu: Conceptualization, Investigation, Formal analysis, Writing – Original Draft, Supervision, Project administration. M. Chen: Investigation, Formal analysis, Writing – Original Draft. J Guo: Formal analysis. J. Zhang: Formal analysis. T. Zeng: Conceptualization, Methodology, Writing – Original Draft, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2022.10.001>.

References

- Hall SN, Appelman HD. Autoimmune Gastritis. *Arch Pathol Lab Med* 2019;143:1327–31.
- Ananthakrishnan AN, Bernstein CN, Iliopoulos D, Macpherson A, Neurath MF, Ali RAR, et al. Environmental triggers in IBD: a review of progress and evidence. *Nat Rev Gastroenterol Hepatol* 2018;15:39–49.
- Shah SC, Piazuelo MB, Kuipers EJ, Li D. AGA Clinical Practice Update on the Diagnosis and Management of Atrophic Gastritis: Expert Review. *Gastroenterology* 2021;161:1325–1332.e1327.
- Banks M, Graham D, Jansen M, Gotoda T, Coda S, di Pietro M, et al. British Society of Gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. *Gut* 2019;68:1545–75.
- Rondonotti E, Spada C, Adler S, May A, Despott EJ, Koulaouzidis A, et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Technical Review. *Endoscopy* 2018;50:423–46.
- Kaplan GG, Ng SC. Understanding and Preventing the Global Increase of Inflammatory Bowel Disease. *Gastroenterology* 2017;152:313–321.e312.
- Kobayashi T, Siegmund B, Le Berre C, Wei SC, Ferrante M, Shen B, et al. Ulcerative colitis. *Nat Rev Dis Primers* 2020;6:74.
- Multisociety Sedation Curriculum for Gastrointestinal Endoscopy. *Hepatology* 2012, 56:E1–25.
- Botezatu A, Bodrug N. Chronic atrophic gastritis: an update on diagnosis. *Med Pharm Rep* 2021;94:7–14.
- Kopacova M, Tachei I, Rejchrt S, Bartova J, Bures J. Double balloon enteroscopy and acute pancreatitis. *World J Gastroenterol* 2010;16:2331–40.
- Delvaux M, Gérard G. Capsule endoscopy in 2005: facts and perspectives. *Best Pract Res Clin Gastroenterol* 2006;20:23–39.
- Zagari RM, Rabitti S, Greenwood DC, Eusebi LH, Vestito A, Bazzoli F. Systematic review with meta-analysis: diagnostic performance of the combination of pepsinogen, gastrin-17 and anti-*Helicobacter pylori* antibodies serum assays for the diagnosis of atrophic gastritis. *Aliment Pharmacol Ther* 2017;46:657–67.
- Bang CS, Lee JJ, Baik GH. Prediction of Chronic Atrophic Gastritis and Gastric Neoplasms by Serum Pepsinogen Assay: A Systematic Review and Meta-Analysis of Diagnostic Test Accuracy. *J Clin Med* 2019;8.
- Kim EH, Kang H, Park CH, Choi HS, Jung DH, Chung H, et al. The optimal serum pepsinogen cut-off value for predicting histologically confirmed atrophic gastritis. *Dig Liver Dis* 2015;47:663–8.
- Reese GE, Constantiniades VA, Simillis C, Darzi AW, Orchard TR, Fazio VW, et al. Diagnostic precision of anti-*Saccharomyces cerevisiae* antibodies and perinuclear antineutrophil cytoplasmic antibodies in inflammatory bowel disease. *Am J Gastroenterol* 2006;101:2410–22.
- Pavlidis P, Shums Z, Koutsoumpas AL, Milo J, Papp M, Umemura T, et al. Diagnostic and clinical significance of Crohn's disease-specific anti-MZGP2 pancreatic antibodies by a novel ELISA. *Clin Chim Acta* 2015;441:176–81.
- Roggenbuck D, Hausdorf G, Martinez-Gamboa L, Reinhold D, Büttner T, Jungblut PR, et al. Identification of GP2, the major zymogen granule membrane glycoprotein, as the autoantigen of pancreatic antibodies in Crohn's disease. *Gut* 2009;58:1620–8.
- Gathungu G, Kim MO, Ferguson JP, Sharma Y, Zhang W, Ng SM, et al. Granulocyte-macrophage colony-stimulating factor autoantibodies: a marker of aggressive Crohn's disease. *Inflamm Bowel Dis* 2013;19:1671–80.
- Holmes E, Li JV, Athanasiou T, Ashrafian H, Nicholson JK. Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends Microbiol* 2011;19:349–59.
- Krishnan S, Ding Y, Saedi N, Choi M, Sridharan GV, Sherr DH, et al. Gut Microbiota-Derived Tryptophan Metabolites Modulate Inflammatory Response in Hepatocytes and Macrophages. *Cell Rep* 2018;23:1099–111.
- Cui J, Liu Y, Hu Y, Tong J, Li A, Qu T, et al. NMR-based metabolomics and correlation analysis reveal potential biomarkers associated with chronic atrophic gastritis. *J Pharm Biomed Anal* 2017;132:77–86.
- Lavelle A, Sokol H. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol* 2020;17:223–37.
- Han Y, Liao Z, Li Y, Zhao X, Ma S, Bao D, et al. Magnetically Controlled Capsule Endoscopy for Assessment of Antiplatelet Therapy-Induced Gastrointestinal Injury. *J Am Coll Cardiol* 2022;79:116–28.
- Aktas H, Mensink PB. Small bowel diagnostics: current place of small bowel endoscopy. *Best Pract Res Clin Gastroenterol* 2012;26:209–20.
- Ding Z, Shi H, Zhang H, Meng L, Fan M, Han C, et al. Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning Model. *Gastroenterology* 2019;157(1044–1054):e1045.
- Pang Z, Chong J, Zhou G, de Lima Morais DA, Chang L, Barrette M, et al. *MetaboAnalyst 5.0*: narrowing the gap between raw spectra and functional insights. *Nucl Acids Res* 2021;49:W388–96.
- Zhang S, Li X, Lin Q, Lin J, Wong KC. Uncovering the key dimensions of high-throughput biomolecular data using deep learning. *Nucl Acids Res* 2020;48:e56.
- Zhang C, Chen Y, Zeng T, Zhang C, Chen L. Deep latent space fusion for adaptive representation of heterogeneous multi-omics data. *Brief Bioinform* 2022;23.
- Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;313:504–7.
- Ye Q, Liu C. An Unsupervised Deep Feature Learning Model Based on Parallel Convolutional Autoencoder for Intelligent Fault Diagnosis of Main Reducer. *Comput Intell Neurosci* 2021;2021:8922656.
- Roy RS, Quadir F, Soltanikazemi E, Cheng J. A deep dilated convolutional residual network for predicting interchain contacts of protein homodimers. *Bioinformatics* 2022.
- Zhang Z, Zhou T, Zhang Y, Pang Y. Attention-based deep residual learning network for entity relation extraction in Chinese EMRs. *BMC Med Inform Decis Mak* 2019;19:55.
- Tang H, Yu X, Liu R, Zeng T. Vec2image: an explainable artificial intelligence model for the feature representation and classification of high-dimensional biological data by vector-to-image conversion. *Brief Bioinform* 2022;23.
- Stocker R, Yamamoto Y, McDonagh AF, Glazer AN, Ames BN. Bilirubin is an antioxidant of possible physiological importance. *Science* 1987;235:1043–6.
- Sedlak TW, Snyder SH. Bilirubin benefits: cellular protection by a biliverdin reductase antioxidant cycle. *Pediatrics* 2004;113:1776–82.
- Chang M, Xue J, Sharma V, Habtezion A. Protective role of hemoxygenase-1 in gastrointestinal diseases. *Cell Mol Life Sci* 2015;72:1161–73.
- Nakao A, Otterbein LE, Overhaus M, Sarady JK, Tsung A, Kimizuka K, et al. Biliverdin protects the functional integrity of a transplanted syngeneic small bowel. *Gastroenterology* 2004;127:595–606.
- Magierowska K, Brzozowski T, Magierowski M. Emerging role of carbon monoxide in regulation of cellular pathways and in the maintenance of gastric mucosal integrity. *Pharmacol Res* 2018;129:56–64.
- Gomes AS, Gadelha GG, Lima SJ, Garcia JA, Medeiros JV, Havt A, et al. Gastroprotective effect of heme-oxygenase 1/biliverdin/CO pathway in ethanol-induced gastric damage in mice. *Eur J Pharmacol* 2010;642:140–5.
- Ervin SM, Li H, Lim L, Roberts LR, Liang X, Mani S, et al. Gut microbial β -glucuronidases reactivate estrogens as components of the estrobolome that reactivate estrogens. *J Biol Chem* 2019;294:18586–99.
- Eussen SJ, Vollset SE, Hustad S, Middtun Ø, Meyer K, Fredriksen A, et al. Vitamins B2 and B6 and genetic polymorphisms related to one-carbon metabolism as risk factors for gastric adenocarcinoma in the European prospective investigation into cancer and nutrition. *Cancer Epidemiol Biomarkers Prev* 2010;19:28–38.
- Han Y, Li Y, Hu Z, Wang X, Liu J, Ren X, et al. Hydrogen sulfide-mediated resistance against water avoidance stress-induced gastritis by maintenance of gastric microbial homeostasis. *Microbiologyopen* 2020;9:e00951.
- Yanaka N, Ohata T, Toya K, Kanda M, Hirata A, Kato N. Vitamin B6 suppresses serine protease inhibitor 3 expression in the colon of rats and in TNF- α -stimulated HT-29 cells. *Mol Nutr Food Res* 2011;55:635–43.
- Roediger WE. The colonic epithelium in ulcerative colitis: an energy-deficiency disease? *Lancet* 1980;2:712–5.
- De Preter V, Arijis I, Windey K, Vanhove W, Vermeire S, Schuit F, et al. Impaired butyrate oxidation in ulcerative colitis is due to decreased butyrate uptake and a defect in the oxidation pathway. *Inflamm Bowel Dis* 2012;18:1127–36.
- Cheng CW, Biton M, Haber AL, Gunduz N, Eng G, Gaynor LT, et al. Ketone Body Signaling Mediates Intestinal Stem Cell Homeostasis and Adaptation to Diet. *Cell* 2019;178:1115–1131.e1115.
- Sudo H, Kubo A. The Aneugenicity of Ketone Bodies in Colon Epithelial Cells Is Mediated by Microtubule Hyperacetylation and Is Blocked by Resveratrol. *Int J Mol Sci* 2021;22.
- Chen X, Zhang J, Wang R, Liu H, Bao C, Wu S, et al. UPLC-Q-TOF/MS-Based Serum and Urine Metabonomics Study on the Ameliorative Effects of Palmatine on *Helicobacter pylori*-Induced Chronic Atrophic Gastritis. *Front Pharmacol* 2020;11:586954.
- Yuan Z, Yang L, Zhang X, Ji P, Hua Y, Wei Y. Mechanism of Huang-lian-jie-du decoction and its effective fraction in alleviating acute ulcerative colitis in mice: Regulating arachidonic acid metabolism and glycerophospholipid metabolism. *J Ethnopharmacol* 2020;259:112872.
- Murgia A, Hinz C, Liggi S, Denes J, Hall Z, West J, et al. Italian cohort of patients affected by inflammatory bowel disease is characterised by variation in glycerophospholipid, free fatty acids and amino acid levels. *Metabolomics* 2018;14:140.
- Bassaganya-Riera J, Reynolds K, Martino-Catt S, Cui Y, Hennighausen L, Gonzalez F, et al. Activation of PPAR gamma and delta by conjugated linoleic acid mediates protection from experimental inflammatory bowel disease. *Gastroenterology* 2004;127:777–91.
- Bassaganya-Riera J, Hontecillas R. Dietary conjugated linoleic acid and n-3 polyunsaturated fatty acids in inflammatory bowel disease. *Curr Opin Clin Nutr Metab Care* 2010;13:569–73.
- Yang M, Cook ME. Dietary conjugated linoleic acid decreased cachexia, macrophage tumor necrosis factor- α production, and modifies splenocyte cytokines production. *Exp Biol Med (Maywood)* 2003;228:51–8.
- Romagnolo DF, Donovan MG, Doetschman TC, Selmin OI. n-6 Linoleic Acid Induces Epigenetics Alterations Associated with Colonic Inflammation and Cancer. *Nutrients* 2019;11.
- de Barros PAV, Rabelo Andrade ME, de Vasconcelos GS, Mendes Miranda SE, Dos Reis DC, Lacerda Leocádio PC, et al. Conjugated linoleic acid prevents

- damage caused by intestinal mucositis induced by 5-fluorouracil in an experimental model. *Biomed Pharmacother* 2018;103:1567–76.
- [56] Panes J: Treatment of angiodysplasia and arterial malformations of the gastrointestinal tract. *Acta Endoscopica* 1998;28:688–688.
- [57] Lu W, Luo M, Fang X, Zhang R, Li S, Tang M, et al. Discovery of metabolic biomarkers for gestational diabetes mellitus in a Chinese population. *Nutr Metab (Lond)* 2021;18:79.
- [58] Chen XY, Fan HN, Zhang HK, Qin HW, Shen L, Yu XT, et al. Rewiring of Microbiota Networks in Erosive Inflammation of the Stomach and Small Bowel. *Front Bioeng Biotechnol* 2020;8:299.
- [59] Yuan K, Zeng T, Chen L. Interpreting Functional Impact of Genetic Variations by Network QTL for Genotype-Phenotype Association Study. *Front Cell Dev Biol* 2021;9:720321.
- [60] Zeng T, Yu X, Chen Z. Applying artificial intelligence in the microbiome for gastrointestinal diseases: A review. *J Gastroenterol Hepatol* 2021;36:832–40.
- [61] Zeng T, Dai H. Single-Cell RNA Sequencing-Based Computational Analysis to Describe Disease Heterogeneity. *Front Genet* 2019;10:629.
- [62] Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc Natl Acad Sci USA* 2013;110:21083–8.
- [63] Guo WF, Zhang SW, Feng YH, Liang J, Zeng T, Chen L. Network controllability-based algorithm to target personalized driver genes for discovering combinatorial drugs of individual patients. *Nucl Acids Res* 2021;49:e37.
- [64] Guo WF, Yu X, Shi QQ, Liang J, Zhang SW, Zeng T. Performance assessment of sample-specific network control methods for bulk and single-cell biological data analysis. *PLoS Comput Biol* 2021;17:e1008962.
- [65] Guo WF, Zhang SW, Zeng T, Akutsu T, Chen L. Network control principles for identifying personalized driver genes in cancer. *Brief Bioinform* 2020;21:1641–62.
- [66] Yu X, Zhang J, Sun S, Zhou X, Zeng T, Chen L. Individual-specific edge-network analysis for disease prediction. *Nucl Acids Res* 2017;45:e170.
- [67] Zeng T, Zhang W, Yu X, Liu X, Li M, Chen L. Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals. *Brief Bioinform* 2016;17:576–92.
- [68] Tang H, Zeng T, Chen L. High-Order Correlation Integration for Single-Cell or Bulk RNA-seq Data Analysis. *Front Genet* 2019;10:371.