Analytical-Chemistry-Informed Transformer for Infrared Spectroscopy

Anonymous Author(s)
Affiliation
Address
email

Abstract

Infrared (IR) spectroscopy is a crucial technique in the field of analytical chemistry. Recently, deep learning (DL) has drawn great interest as the modeling method of IR spectral data. Unlike vision or language tasks, IR spectral data modeling has distinctive characteristics and is faced with the problem of calibration transfer, which necessitates the assistance of prior knowledge. However, there is a lack of DL modules that incorporate the knowledge of analytical chemists. To this end, we propose Analytical-Chemistry-Informed Transformer (ACT) with two modules that incorporate the field knowledge in analytical chemistry. First, ACT features a module referred to as learnable spectral processing for generic spectral pre-processing, tokenization, and post-processing. Second, a specialized attention mechanism, namely spectral-attention, is incorporated into ACT. Spectral-attention utilizes the intra-spectral and inter-spectral correlations to extract intrinsic features. Empirical results show that ACT has achieved state-of-the-art (SOTA) results in 9 analytical tasks covering applications across pharmacy, chemistry, agriculture, and food science. Compared with SOTA networks, ACT reduces the root mean square error of prediction (RMSEP) by an average of 27% in calibration transfer tasks. These results indicate that DL modeling methods could benefit from the prior knowledge of IR spectroscopy. The code is publicly available at (masked for anonymity).

1 Introduction

Infrared (IR) spectroscopy plays a crucial role in both scientific research [11, 30, 2] and industrial application [29, 42, 18], which provides a rapid, convenient, non-destructive and economical solution to chemical analysis [38, 17]. As vibrational spectroscopy, IR spectroscopy studies the absorbance or reflection of light resulting from molecular vibrations, which manifests itself as spectra containing peaks or overtones at different wavelengths [22]. IR spectroscopy has been introduced to the analysis of complex samples in biological, medical, and chemical applications with the assistance of chemometrics [31, 15], where machine learning plays a key role. In a typical chemometric modeling pipeline, a batch of spectra with labels serve as the calibration (training) and the validation set for establishing a calibration model. The calibration model is then utilized for the analysis of spectra collected from testing samples.

Despite the success of chemometric methods, it is still challenging to establish a robust calibration model across capricious spectra-collecting environments [41]. It is recognized that the performance of calibration models will degrade when handling the data collected with different spectrometers, varied sampling protocols, or different environmental conditions [25]. Calibration transfer aims to improve the prediction ability under such a condition, where training (source) data and testing (target) data might not follow the i.i.d. assumption due to different spectra-collecting processes. Conventional

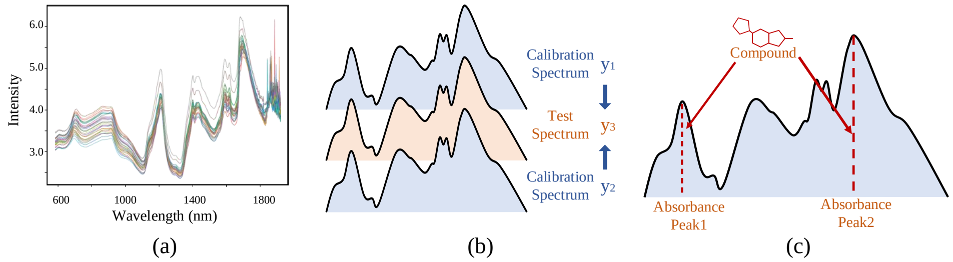


Figure 1: Characteristics of IR spectroscopy: (a) examples of NIR spectra from tablets; (b) interpreting the quantitative information of IR spectra relies on labeled calibration spectra; (c) non-adjacent absorbance peaks resulting from the same chemical compound are correlated.

calibration transfer methods either retrain [43, 20] or regularize [25, 44] the calibration model with the data from the target domain. However, collecting extra data from the target domain could be costly and such data collection is not always practical. A robust calibration model providing reliable predictions in varied spectra-collecting environments is therefore attractive.

In the last decade, deep learning (DL) has dramatically improved the state-of-the-art in nature language processing (NLP), computer vision (CV), and many other fields [12, 34, 39, 14, 24]. On the contrary, DL for IR spectroscopy is still in its infancy [21]. Besides the lack of large open datasets, the absence of specialized DL architectures is also responsible for this delay. The widespread deep neural networks, like convolutional networks (CNNs) and Transformers, have been introduced to IR spectra modeling [45, 10, 7]. Recently, some researchers try to boost the performance via data augmentation and learn the augmented data with CNNs [35, 40, 4]. Although the introduction of these networks does alleviate the reliance on spectral pre-processing [21], the learned models still suffer performance degradation when training data and testing data are generated from different spectral-collecting processes.

We attribute this performance degradation to the lack of inductive biases for infrared spectral analysis. The characteristics of IR spectral data differ from the characteristics of vision and language data. As shown in Fig. 1, both inter- and intra-spectral correlations are vital for IR spectra analysis. IR spectroscopy is an indirect method and thus inter-spectral correlations determine the quantitative meaning of a single spectrum [21]. A compound with multiple chemical groups could absorb infrared radiation at multiple wavelengths [22], and the intra-spectral correlation therefore contains the information of possible chemical compounds. Moreover, the absorbance peaks caused by a single compound are usually non-adjacent, resulting in long-range dependencies along the spectral axis. The characteristics of infrared spectra also present challenges to the modeling methods due to the presence of baseline (background) drift, peak shifting, and peak overlaps [6].

Incorporating the knowledge from analytical chemists is therefore crucial for fully realizing the potentiality of DL in IR spectroscopy. In this paper, a Transformer specialized in IR spectroscopy is proposed, namely Analytical-Chemistry-Informed Transformer (ACT). Compared with Vanilla Transformer [34], ACT implements several modifications in terms of spectral processing and attention mechanism. Firstly, a learnable spectral processing module is proposed and incorporated into ACT, which is an integration of pre-processing, tokenization, and post-processing. Inspired by chemometric baseline correction, our learnable spectral processing introduces reversible spectral pre-processing for the first time. Secondly, ACT introduces a spectral-attention mechanism in place of self-attention. Following the knowledge that both inter- and intra-spectral correlations are vital, spectral-attention is designed to discover the similarities among spectra and utilizes the correlations among spectral bands. Guided by the field knowledge, ACT could extract intrinsic representations that exhibits superior ability in calibration transfer, alleviating the need for extra pre-processing simultaneously.

We evaluate ACT in 9 different IR spectra modeling tasks (4 calibration transfer tasks and 5 regular tasks), covering the analytical tasks in pharmacy, chemistry, agriculture, and food science. The contributions are summarized as follows:

1. We propose ACT, a Transformer specialized in IR spectral data modeling, to provide a robust modeling method across capricious applications and spectra-collecting environments.

2. Inspired by conventional chemometric methods, we incorporate learnable spectral processing into ACT, which introduces reversible spectral pre-processing for the first time.
3. Following the prior knowledge of IR spectroscopy, we propose the spectral-attention mechanism that considers both the inter- and intra-spectral correlations.
4. We conduct comprehensive experiments in 9 IR spectral data modeling tasks, where ACT outperforms the baseline DL methods in terms of both accuracy and adaptability.

2 Preliminary and related work

Infrared spectroscopy and chemometrics IR spectroscopy can be generally divided into three categories: the near-infrared (NIR), the mid-infrared (MIR), and the far-infrared [6]. Among these categories, NIR and MIR spectroscopy have widespread applications in pharmacy [13], chemistry [18], agriculture [29], biology [30] and material science [9]. As an indirect method, IR spectroscopy requires a calibration model for qualitative and quantitative analysis. Chemometrics studies the modeling methods for chemical data including IR spectra, where machine learning techniques have a crucial role [16]. The conventional chemometric modeling pipeline is similar to the work pipeline of shallow machine learning, where knowledge-driven spectral pre-processing (feature engineering) is usually essential [23]. The effectiveness of these fixed spectral pre-processing methods varies on different datasets, resulting in an exhaustive search process. In this paper, the proposed ACT integrates with a unique learnable pre-processing module, providing better adaptability across different datasets.

Deep learning for infrared spectra modeling As DL blooms in the last decade, deep networks have been utilized for IR spectral data modeling [36, 5]. CNN along with its variants is one of the most popular deep networks in the field of infrared spectroscopy. DeepSpectra [45], one of the earliest deep spectral modeling methods, utilizes the structure of Inception [32]. Transformer also draws considerable attention [4, 37]. Spectraformer [4] combines the encoder of Vanilla Transformer with multi-layer perception. Despite the widespread DL applications, DL for IR spectroscopy is still in its infancy. Many studies adopt existing networks and merely adjust the hyper-parameters like layer numbers and kernel sizes [3]. Some recent studies like AggMapNet [35] and TeaNet [40] have introduced specialized data augmentation methods to CNNs. However, IR spectra have unique characteristics differing from those of data in CV or NLP. For instance, IR spectra usually contain long-range dependencies along the spectral axis due to correlations between non-adjacent spectral peaks. Meanwhile, the correlations between spectra are also vital since the quantitative meaning of a single spectrum is defined by the labeled calibration spectra. Therefore, ACT incorporates spectral-attention in place of self-attention to match the characteristics of IR spectra.

Calibration transfer A well-known problem with IR spectroscopy is that the predictions of a calibration model are reliable only if the calibrating (training) and testing spectra are collected with an identical process [25]. Differences in measurement environments, instruments, and sample-handling protocols could disturb the calibration model [33]. Both the conventional methods and the DL methods suffer a performance degradation during tackling data from a different spectra-collecting process [19]. Researchers try to alleviate this problem by retraining or regularizing the model with the data from target domain (data within the same domain as the testing data) [43, 20, 25, 44]. However, collecting extra data could be costly and the data from target domain is not always available. A generic model that could achieve satisfactory performance across different spectra-collecting processes (with good domain generalization) is therefore attractive. To this end, ACT incorporates the prior knowledge from analytical chemists to approach such an ideal generic model.

3 Proposed method

As mentioned above, IR spectral data is distinguished from the data in CV or NLP by its unique characteristics. Meanwhile, the changes in spectra-collecting process also matter, which will result in catastrophic model degradation. CNNs and Transformers utilized by previous studies neglect these characteristics, as these networks are originally designed for other purposes. To address this issue, we introduce prior knowledge to Transformer and propose an Analytical-Chemistry-Informed Transformer (ACT). ACT incorporates two specialized designs: learnable spectral processing and spectral-attention. The overall framework of ACT is illustrated in Fig. 2.

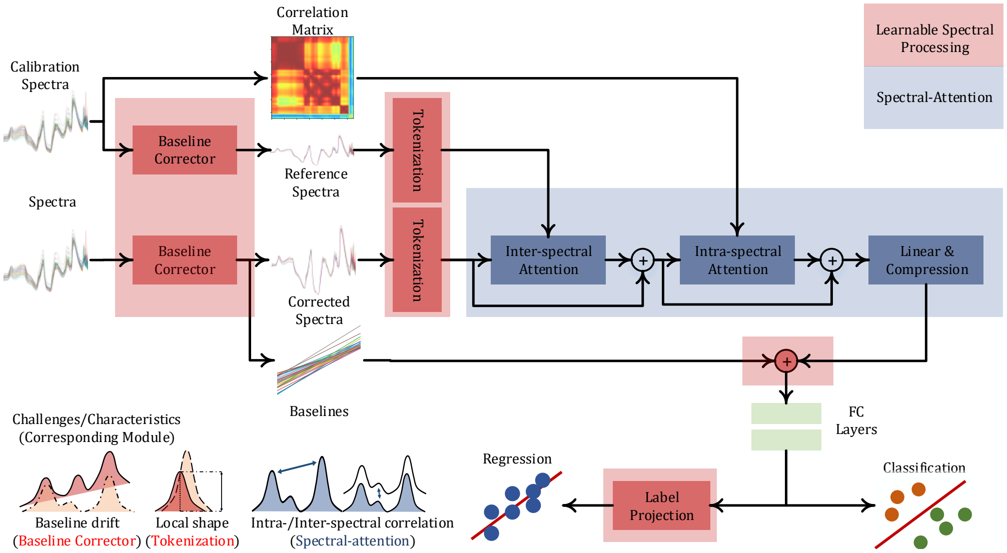


Figure 2: ACT architecture. The learnable spectral processing module (red blocks) first splits raw spectra into baselines and corrected spectra, then recovers the baseline information, and finally adjusts predictions for regression tasks. Spectral-attention (blue blocks) utilizes reference spectra and correlation matrix to refine the attention map. The challenges/characteristics along with corresponding modules are also presented at the bottom left of the figure.

In this paper, we denote a set of IR spectra as $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$, while the corresponding labels (chemical properties) are denoted as $\boldsymbol{Y}$. Given a set of calibration spectra and corresponding labels $\{\boldsymbol{X}_{cal}, \boldsymbol{Y}_{cal}\}$, ACT aims to analyze the testing spectra $\boldsymbol{X}_{test}$ and predict the labels $\boldsymbol{Y}_{test}$, where $\boldsymbol{X}_{test}$ might be collected with different instruments, sample preparing protocols, etc. In the following sections, n and b represent the number of samples and spectral bands respectively.

3.1 Learnable spectral processing

Learnable spectral processing incorporates a reversible baseline corrector, which separates baselines from spectra via iteratively fitting polynomial curves. Distinguishing from conventional spectral pre-processing [28], the reversible baseline correction keeps baselines and reverses the correction in the subsequent modules. For spectral data $\boldsymbol{X} \in \mathbb{R}^{n \times b}$, the baseline correction process is denoted as:

$$\begin{aligned} \boldsymbol{\mathcal{B}} &= \mathrm{BaseLineCorr}(\boldsymbol{X}) \\ \boldsymbol{\mathcal{X}} &= \boldsymbol{X} - \boldsymbol{\mathcal{B}}, \end{aligned} \tag{1}$$

where $\boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{X}} \in \mathbb{R}^{n \times b}$ denote the baselines and corrected spectra respectively. $\boldsymbol{\mathcal{X}}$ is subsequently normalized. For each spectrum $\S_i \in \boldsymbol{\mathcal{X}}$, the varied peak shape leads to different analytical meanings of absorbance/reflectance intensities at a single wavelength. As a result, the same absorbance intensities could represent different levels of chemical groups/compounds. It is hard to represent such neighboring information by utilizing the intensity values at each position exclusively.

To preserve the local spectral shape, the corrected spectral data $\boldsymbol{\mathcal{X}}$ are divided into overlapped patches. The patch at a single wavelength consists of the spectral intensities within a neighboring window. For a single spectral point at wavelength λ_j, l neighboring spectral points at higher wavelengths and l neighboring spectral points at lower wavelengths are included in the patch. We pad l zeros to the start and the end of corrected spectra. This patch with a size of $c = 2l + 1$ serves as the token of centering spectral point, namely the spectral point at λ_j. The tokenization process generates a sequence of patches $\boldsymbol{\mathcal{X}}_{(c)} \in \mathbb{R}^{n \times b \times c}$, which is utilized as the spectral embeddings after position encoding.

Baselines $\mathcal{B}$ are likely to contain discriminative information due to the estimation error. The sample property may also correlate with physical characteristics that contribute to baselines. We therefore restore baselines after the corrected spectra are encoded by spectral-attention layer:

$$\begin{aligned} \boldsymbol{\mathcal{X}}_{en} &= \mathrm{SpectrAttn}(\boldsymbol{\mathcal{X}}_{(c)}) \\ \boldsymbol{X}_{en} &= \alpha_1 \boldsymbol{\mathcal{X}}_{en} + \alpha_2 \boldsymbol{\mathcal{B}}, \end{aligned} \tag{2}$$

where $\boldsymbol{\mathcal{X}}_{en}$ is the output of spectral-attention layer. $\mathrm{SpectrAttn}$ is the notations of spectral-attention, the details of which will be introduced in the following subsection. Two trainable parameters, namely α_1 and α_2, are introduced to adjust the influence of baselines.

We also introduce a trainable label projecting module for regression tasks, which are ubiquitous in IR spectra analysis (e.g. quantitative estimation of chemical contents). The label projecting module records the statistical properties of training labels, namely mean μ and range r. Given a network output $\hat{\boldsymbol{\mathcal{Y}}}$, the label projecting module maps the output into the original label space:

$$\hat{\boldsymbol{Y}} = (\tanh(\hat{\boldsymbol{\mathcal{Y}}}) - \alpha_3) * r/\alpha_4 + \mu, \tag{3}$$

where $\hat{\boldsymbol{Y}} \in \mathbb{R}^{n \times 1}$ denotes the final prediction. α_3 and α_4 are two trainable parameters that provide flexibility to the projecting module, while $\tanh$ is introduced for scaling and nonlinearity.

3.2 Spectral-attention

In IR spectral data modeling, both the inter and intra-spectral correlations are crucial. As shown in Fig. 1 and Fig. 2, the quantitative meaning of a spectrum is determined by calibrating spectra with known labels. The inter-spectral correlations are therefore vital to IR spectroscopy. To enhance the ability to capture such correlations, we introduce inter-spectral attention. We select n_r spectra from the calibration (training) set and form a set of reference spectra $\boldsymbol{X}_r$. The reference spectra are corrected and embedded by the above learnable spectral processing module, deriving reference patches $\boldsymbol{\mathcal{X}}_{r(c)} \in \mathbb{R}^{n_r \times b \times c}$. We use $\boldsymbol{\mathcal{X}}^{(:,\lambda_j,:)}$ to denote the slice of $\boldsymbol{\mathcal{X}}$ at wavelength λ_j. Given a batch of processed data $\boldsymbol{\mathcal{X}}_{(c)}$, the inter-spectral attention (referred as $\mathrm{InterSpecAttn}$) at wavelength λ_j is computed as:

$$\begin{aligned} &\mathrm{InterSpecAttn}(\boldsymbol{\mathcal{X}}_{(c)}^{(:,\lambda_j,:)}, \boldsymbol{\mathcal{X}}_{r(c)}^{(:,\lambda_j,:)}) = \mathrm{Softmax}(\mathrm{Filter}(\frac{\boldsymbol{Q}\boldsymbol{K}_{\mathrm{cat}}^T}{\sqrt{d_k}}))\boldsymbol{V}_{\mathrm{cat}} \\ &\boldsymbol{K}_{\mathrm{cat}} = \mathrm{Concat}(\boldsymbol{K}, \boldsymbol{K}_r) \quad \boldsymbol{V}_{\mathrm{cat}} = \mathrm{Concat}(\boldsymbol{V}, \boldsymbol{V}_r), \end{aligned} \tag{4}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V} \in \mathbb{R}^{n \times d_k}$ are the queries, keys and values of $\boldsymbol{\mathcal{X}}_{(c)}^{(:,\lambda_j,:)} \in \mathbb{R}^{n \times c}$ respectively, while $\boldsymbol{K}_r$, $\boldsymbol{V}_r \in \mathbb{R}^{n_r \times d_k}$ are the keys and values of $\boldsymbol{\mathcal{X}}_{r(c)}^{(:,\lambda_j,:)}$. Operation Filter will suppress the attention between spectra within $\boldsymbol{\mathcal{X}}$ to 0, highlighting the attention from reference spectra. Inter-spectral attention calculates the attention between input samples and reference samples. Residual connection is also introduced and the output of inter-spectral attention is given as:

$$\boldsymbol{\mathcal{X}}_{\mathrm{Inter}} = \mathrm{Concat}(\mathrm{InterSpecAttn}(\boldsymbol{\mathcal{X}}^{(:,\lambda_j,:)}, \boldsymbol{\mathcal{X}}_r^{(:,\lambda_j,:)})_{j=1}^{b}) + \boldsymbol{\mathcal{X}}. \tag{5}$$

A single absorbance/reflectance peak covers several adjacent spectral bands, resulting in correlations between these bands. Non-adjacent spectral peaks might also correlate with each other, as a single chemical compound could result in multiple absorbance peaks. We introduce a correlation matrix $\boldsymbol{C} \in \mathbb{R}^{b \times b}$ to record the correlation between spectral bands and utilize it to guide intra-spectral attention. For data $\boldsymbol{\mathcal{X}}$, the intra-spectral attention of i th sample $\boldsymbol{\mathcal{X}}_{(c)}^{(i,:,:)} \in \mathbb{R}^{b \times c}$ is defined as:

$$\mathrm{IntraSpecAttn}(\boldsymbol{\mathcal{X}}_{(c)}^{(i,:,:)}, \boldsymbol{C}) = \mathrm{Softmax}((1 - \alpha_5) * \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}} + \alpha_5 * \frac{\boldsymbol{C}\boldsymbol{W}_c}{\sqrt{b}})\boldsymbol{V}, \tag{6}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V} \in \mathbb{R}^{b \times d_k}$ are the queries, keys and values of $\mathcal{X}_{(c)}^{(i,:,:)}$ respectively. $\boldsymbol{W}_c \in \mathbb{R}^{b \times b}$ is a learnable projection matrix, while α_5 is also a learnable weight that adjusts the influence of correlation matrix. In this paper, $\boldsymbol{\mathcal{X}}_{\mathrm{Inter}}$ serves as the input of intra-spectral attention and the corresponding representation is given as:

$$\boldsymbol{\mathcal{X}}_{\mathrm{Intra}} = \mathrm{Concat}(\mathrm{IntraSpecAttn}(\boldsymbol{\mathcal{X}}_{\mathrm{Inter}}^{(i,:,:)}, \boldsymbol{C})_{i=1}^{n}). \tag{7}$$

$\mathrm{IntraSpecAttn}$ calculates 'global' attention across the whole spectrum, which could capture the long-range dependencies among spectral peaks.

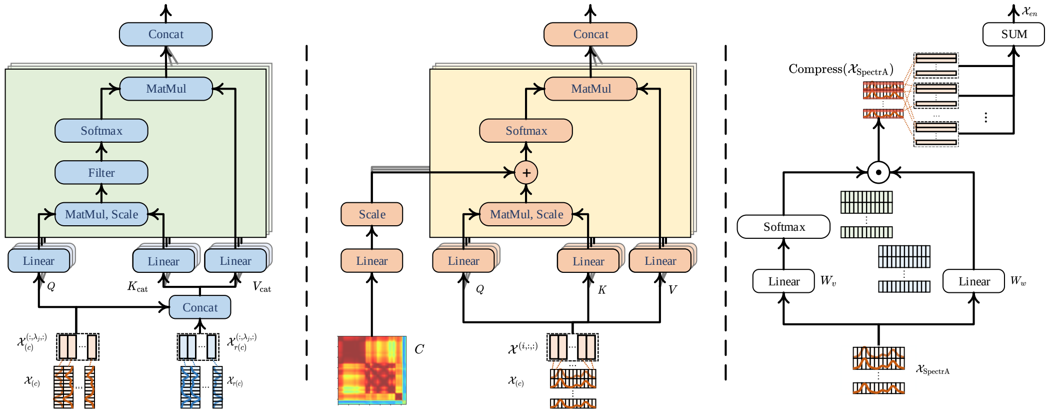


Figure 3: Spectral-attention: InterSpecAttn (left), IntraSpecAttn (middle) and compression of $\mathcal{X}_{\mathrm{SpectrA}}$ (right). We utilize reference spectra selected from the calibration set to capture the quantitative information within spectra, while a band-wise correlation matrix is introduced to guide the global intra-spectra attention. Two projection matrices are utilized to calculate the importance and embedded features of tokens, which enables the extraction of local intra-spectral correlations.

The final output of spectral-attention is the combination of inter- and intra-spectral attention, where another learnable parameter is introduced to control the influences of the two attentions. A fully-connected feed-forward network with batch normalization is also included:

$$\boldsymbol{\mathcal{X}}_{\mathrm{SpectrA}} = \mathrm{FeedForward}((1-\alpha_6)\boldsymbol{\mathcal{X}}_{\mathrm{Inter}} + \alpha_6 * \boldsymbol{\mathcal{X}}_{\mathrm{Intra}}), \tag{8}$$

where α_6 is a learnable weight that adjusts influences of inter- and intra- spectral attention. It should be noted that we use batch normalization instead of layer normalization between the feed-forward network to increase the interactions between spectra. The encoded feature $\boldsymbol{\mathcal{X}}_{\mathrm{SpectrA}} \in \mathbb{R}^{n\times b\times c}$ is then compressed into $\boldsymbol{\mathcal{X}}_{en} \in \mathbb{R}^{n\times b}$ with a weighted sum process along the token dimension. Two learnable projection matrix $\boldsymbol{W}_v, \boldsymbol{W}_w \in \mathbb{R}^{c\times c}$ are introduced:

$$\begin{aligned}
\boldsymbol{\mathcal{X}}_{en} &= \mathrm{SpectrAttn}(\boldsymbol{\mathcal{X}}_{(c)}) = \sum_{c}^{k=1} \mathrm{Compress}(\boldsymbol{\mathcal{X}}_{\mathrm{SpectrA}})^{(:,:,k)} \\
\mathrm{Compress}(\boldsymbol{\mathcal{X}}_{\mathrm{SpectrA}}) &= \mathrm{Concat}((\boldsymbol{\mathcal{X}}^{(i,:,:)}_{\mathrm{SpectrA}}\boldsymbol{W}_v \odot \mathrm{Softmax}(\boldsymbol{\mathcal{X}}^{(i,:,:)}_{\mathrm{SpectrA}}\boldsymbol{W}_w))_{i=1}^{n}),
\end{aligned} \tag{9}$$

where $\odot$ denotes Hadamard Product or element-wise product. $\mathrm{Compress}(\boldsymbol{\mathcal{X}}_{\mathrm{SpectrA}}) \in \mathbb{R}^{n\times b\times c}$ represents the weighted $\boldsymbol{\mathcal{X}}_{\mathrm{SpectrA}}$ where the significance of each token features has been added. The above compression operator calculates correlations within sliding windows that form tokens, which could capture 'local' information in complementary to the 'global' information captured by IntraSpecAttn. The encoded feature $\boldsymbol{\mathcal{X}}_{en}$ is then handled by the learnable spectral processing module and fully connected layers to get the final prediction of ACT.

4 Experiment

In this section, we evaluate the proposed ACT on 9 real-world tasks, including both calibration transfer tasks and ordinary calibration tasks. These evaluation tasks cover the analytical applications of IR spectroscopy in pharmacy, chemistry, agriculture, and food science.

4.1 Experimental setup

Datasets 9 real-world tasks originated from 5 datasets are used for evaluation. (1) *Tablet* dataset contains NIR spectra of tablets collected by two individual spectrometers (referred to as spectrometer No. 1 and No. 2). Tablet dataset gives four modeling tasks: *Tablet(1, 1)*, *Tablet(1, 2)*, *Tablet(2, 2)* and *Tablet(2, 1)*. (2) *Melamine* dataset consists of NIR spectra collected from melamine-formaldehyde with slightly different compositions. In this paper, we use two recipes R562 and R568 for evaluation, generating two tasks *MF(R562, R568)* and *MF(R568, R562)*. It should be noted that selected

Table 1: Calibration transferring tasks without any access to the target domain. IMP(%) stands for relative improvements and the best results are highlighted in **bold**.

	Tablet(1, 2)		Tablet(2, 1)		MF(R562,R568)		MF(R568,R562)	
	RMSEP	MAE	RMSEP	MAE	RMSEP	MAE	RMSEP	MAE
DeepSpectra	11.676 ±3.275	9.762 ±3.013	14.983 ±5.08	13.729 ±5.272	4.055 ±1.148	3.239 ±1.043	4.507 ±1.736	3.763 ±1.713
AggMapNet	10.142 ±0.901	7.031 ±0.746	13.212 ±1.154	10.322 ±1.111	2.631 ±0.227	**1.697** ±**0.131**	3.775 ±0.124	2.827 ±0.125
TeaNet	11.819 ±3.673	9.758 ±3.654	17.944 ±4.112	14.961 ±4.159	5.733 ±1.586	4.424 ±1.414	10.301 ±2.374	8.469 ±2.053
Spectraformer	8.810 ±2.048	6.497 ±1.837	9.279 ±2.239	7.445 ±2.272	3.168 ±0.633	2.563 ±0.561	4.247 ±0.451	3.384 ±0.254
ACT	**6.941** **±3.255**	**5.210** **±2.994**	**6.155** **±0.985**	**4.786** **±1.156**	**2.293** **±0.624**	1.798 ±0.51	**2.223** **±0.202**	**1.614** **±0.199**
Imp(%)	21.21%	19.81%	33.67%	35.72%	12.85%	-5.95%	41.11%	42.91%

absorbance peaks rather than whole spectra are utilized for analysis. (3) *Mango_DMC* [1] dataset contains NIR spectra of intact mango fruit, aiming at predicting the dry matter content across different seasons, location, and cultivar. (4) *Strawberry* [8] tries to classify MIR spectra of fruit purees collected by a Fourier transform infrared spectrometer. (5) *Apple_Leaf* aims to classify NIR spectra of apple leaves from 20 different varieties or cultivars. *Strawberry* and *Apple_Leaf* are classification tasks while the other tasks are regression tasks. Moreover, *Tablet(1, 2)*, *Tablet(2, 1)*, *MF(R562, R568)*, and *MF(R568, R562)* are calibration transfer tasks, while the other tasks are regular tasks. Further details are presented in Appendix A.

Baselines We include 4 deep learning baselines and 2 classical calibration transfer methods for comparison. (1) DeepSpectra [45] is a far-reaching end-to-end network for quantitative spectral analysis, which is based on Inception network. (2) AggMapNet [35] converts infrared spectra into 2D maps for feature augmentation and introduces 2D CNN for learning the maps. (3) TeaNet [40] masks and reconstructs the input spectra for data augmentation, where the augmented data are used to boost the modeling performance. (4) Spectraformer [4] is a hybrid network for IR spectroscopy that combines 1D convolutional layers with an attention layer. (5) di-PLS (domain invariant PLS) [25] introduces a domain regularizer for calibration transfer. (6) DIPALS [26] identifies a low-dimensional subspace and views the calibration transfer as a domain adaptation problem.

Experiment details ACT uses mean squared error (MSE) loss for regression tasks and cross-entropy (CE) loss for classification tasks, with Adam serving as the optimizer. We use root mean square error of prediction (RMSEP) and mean absolute error (MAE) as the evaluation metrics for regression, while accuracy (ACC), area under the curve (AUC), and weighted F1 (F1) score are used for classification. All the experiments are implemented based on PyTorch [27] and are repeated for 5 times with NVIDIA RTX 4090 24GB GPU. It should be noted that we adjust both the proposed ACT and the baselines to avoid obvious overfitting or underfitting. The parameter settings are presented in Appendix C.

4.2 Calibration transfer

We report the results of calibration transferring tasks in Table 1 and Table 2. The DL methods learn the calibration spectra from one instrument/recipe and predict the testing spectra from different instruments/recipes. Differing from conventional calibration transferring methods, the tested DL methods are trained without any access to the spectra generated from secondary spectra-collecting procedure (i.e. viewing calibration transfer as domain generation rather than domain adaptation).

As shown in Table 1, it is indicated that the proposed ACT outperforms DL methods across various tasks with a considerable margin in terms of both RMSEP and MAE. Specifically, ACT achieves a 27% reduction of RMSEP and a 23% reduction of MAE on tasks from the Tablet dataset. We attribute this improvement to the modules informed by analytical chemistry. Band-wise correlation map (illustrated in Fig. 2) shows that the absorbance peaks in the Tablet spectra are highly correlated and there are long-range dependencies. Meanwhile, the inter-spectral correlations are also vital,

Table 2: Comparison with classical calibration transfer methods. It should be noted that the two comparison methods achieve the below results **with** the assistance of 60% unlabeled target domain data, while the proposed ACT is trained **without** access to target domain data.

	Tablet(1, 2)		Tablet(2, 1)		MF(R562,R568)		MF(R568,R562)	
	RMSEP	MAE	RMSEP	MAE	RMSEP	MAE	RMSEP	MAE
di-PLS [25]	8.690	/	7.980	/	2.470	/	2.580	/
	±1.060	/	±0.830	/	±0.984	/	±0.997	/
DIPALS[26]	7.690	/	7.120	/	1.750	/	2.020	/
	±0.470	/	±0.680	/	±0.140	/	±0.140	/
ACT	6.941	5.210	6.155	4.786	2.293	1.798	2.223	1.614
	±3.255	±2.994	±0.985	±1.156	±0.624	±0.51	±0.202	±0.199

Table 3: Regular IR spectral data analysis (quantitative). Rank denotes the average rank across different datasets and the best results are highlighted in bold.

	Tablet(1, 1)		Tablet(2, 2)		Mango_DMC		
	RMSEP	MAE	RMSEP	MAE	RMSEP	MAE	Rank
DeepSpectra	6.218	4.269	5.738	4.054	1.039	0.791	4.333
	±1.034	±0.941	±0.676	±0.475	±0.09	±0.048	
AggMapNet	5.543	4.007	6.329	4.238	1.357	0.970	3.000
	±0.493	±0.609	±0.379	±0.323	±0.047	±0.018	
TeaNet	5.972	4.093	6.194	4.289	1.143	0.892	3.667
	±0.548	±0.392	±0.154	±0.166	±0.151	±0.067	
Spectraformer	6.679	4.404	5.956	3.958	1.038	0.781	3.000
	±0.839	±0.466	±0.242	±0.191	±0.067	±0.046	
ACT	**4.604**	**2.914**	**4.417**	**2.714**	**1.008**	**0.756**	1.000
	±0.661	**±0.688**	**±0.15**	**±0.186**	**±0.057**	**±0.039**	

especially the ones between calibration spectra and testing spectra, as the regression task needs quantitative information. ACT could therefore extract domain-invariant representations that conforms to IR spectroscopy.

Compared with the classical calibration transfer methods, ACT achieves comparable (even better in some cases) results without access to target domain data, as presented in Table 2. ACT has shown the potentiality of providing a generic model across different instruments and samples. Eliminating the need for extra samples from the target domain means lower cost and higher efficiency. We owe this improvement to the combination of DL and analytical chemistry. The deep structure of ACT provides abundant learning ability beyond classical methods, facilitating more robust model across different scenarios. On Tablet tasks, for example, several DL methods without specific modification achieves results comparable to those of traditional calibration transfer methods. Knowledge from analytical chemists guides ACT and further improves performance: the proposed modules introduce additional inductive bias which could guide ACT to learn domain invariant representations. It should be noted that the characteristics bands have been manually picked out on MF tasks, which lowers the need for learning ability and thus benefits the classical methods.

4.3 Regular IR spectral data analysis

The experimental results on regular IR spectral data analysis are reported in Table 3 and Table 4. The proposed ACT is further on regular IR spectral datasets where spectra are collected in relatively stationary environments. In both qualitative (classification) and quantitative (regression) tasks, ACT achieves state-of-the-art (SOTA) results across different datasets. Specifically, ACT achieves an average rank of 1.17 on both the regression and the classification tasks, while the second-best average rank is 2.83. We also attribute this improvement to the modules integrated with analytical chemistry. Spectra within Apple_leaf dataset, for example, have highly correlated absorbance peaks at the first overtone (1600-1800 nm) and combination band (around 2200 nm) region. Utilizing such prior information could guide the attention layers.

Table 4: Regular IR spectral data analysis (qualitative). Rank denotes the average rank across different datasets and the best results are highlighted in bold.

	Strawberry			Apple_leaf			
	ACC	AUC	F1	ACC	AUC	F1	Rank
DeepSpectra	0.960 ±0.008	0.981 ±0.003	0.960 ±0.007	0.661 ±0.051	0.804 ±0.04	0.608 ±0.061	2.833
AggMapNet	0.962 ±0.002	0.963 ±0.001	0.962 ±0.002	0.424 ±0.004	0.651 ±0.003	0.405 ±0.006	3.833
TeaNet	0.921 ±0.011	0.961 ±0.009	0.921 ±0.011	0.785 ±0.046	0.888 ±0.023	**0.791** **±0.045**	3.333
Spectraformer	0.960 ±0.009	0.975 ±0.004	0.959 ±0.009	0.502 ±0.038	0.695 ±0.025	0.450 ±0.046	3.833
ACT	**0.963** **±0.002**	**0.991** **±0.003**	**0.964** **±0.002**	**0.793** **±0.014**	**0.914** **±0.008**	0.779 ±0.013	**1.167**

Table 5: Ablation study. Notation '+' means incorporating specific modules and the best results are highlighted in **bold**.

	Tablet(1, 2)		Mango_DMC		Strawberry		
	RMSEP	MAE	RMSEP	MAE	ACC	AUC	F1
Base	9.977 ±2.288	7.521 ±1.681	1.706 ±0.365	1.277 ±0.021	0.907 ±0.039	0.977 ±0.005	0.904 ±0.043
Base + Token	8.491 ±2.837	6.477 ±2.376	1.130 ±0.087	0.854 ±0.07	0.953 ±0.003	0.974 ±0.001	0.953 ±0.003
Base + Token + LearnProc	7.833 ±2.571	5.690 ±1.995	1.446 ±0.327	1.129 ±0.265	0.951 ±0.018	0.979 ±0.002	0.951 ±0.017
Base + Token + SpectrAttn	8.213 ±2.751	6.252 ±2.404	1.091 ±0.054	0.747 ±0.034	0.962 ±0.004	0.990 ±0.004	0.962 ±0.004
ACT	**6.941** **±3.255**	**5.210** **±2.994**	**1.008** **±0.057**	**0.756** **±0.039**	**0.963** **±0.002**	**0.991** **±0.003**	**0.964** **±0.002**

4.4 Ablation results

The ablation study is conducted on three tasks: Tablet(1, 2) for calibration transfer, Mango_DMC for regular regression, and Strawberry for regular classification. ACT is decomposed into several hierarchical models for ablation study: (1) **Base** is the basic model containing an encoder layer based on self-attention and a fully connected network with 2 hidden layers. (2) **Base + Token** is integrated with the proposed tokenization method within learnable spectral processing. (3) **Base + Token + LearnProc** further incorporates the whole learnable spectral processing module. (4) **Base + Token + SpectrAttn** is the combination of base model and spectral-attention.

Results in Table 5 indicates the effectiveness of analytical-chemistry-informed modules. Firstly, the Base model achieves acceptable results in the Tablet(1, 2) task, which shows the potentiality of attention mechanism in calibration transfer tasks. Compared with the base model, the proposed tokenization improves the performance on Tablet(1, 2) and Mango_DMC, while the whole learnable spectral processing module achieves considerable improvement in all the three tasks. Besides, the introduction of spectral-attention also benefits IR spectra modeling in the three tasks.

4.5 Qualitative evaluation

Prediction results and corresponding ground truth of quantitative tasks are plotted for qualitative analysis. The qualitative results on task Tablet(1,2) are shown in Fig. 4 (a-e), where SOTA deep learning methods are included for comparison. The result of ACT is visually better, with lower prediction error and fewer outliers (predictions with large errors). We also plot the results of ablation methods which are presented in Fig. 4 (f-i). These results further indicate the effectiveness of analytical-chemistry-informed modules. As mentioned above, the inductive bias following analytical chemistry could enable ACT to learn domain-invariant representations. It should be noted that the

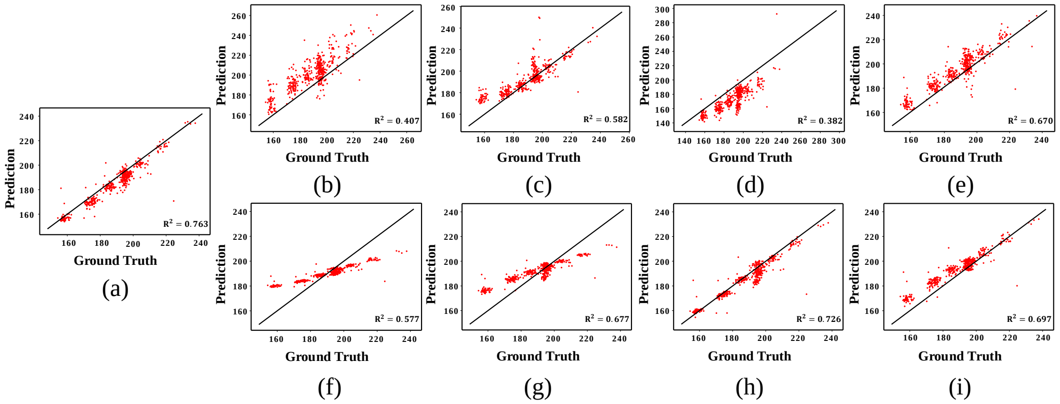


Figure 4: Qualitative evaluation on the Tablet(1,2) task. (a) ACT (b) DeepSpectra, (c) AggMapNet, (d) TeaNet, (e) Spectraformer, (f) Base, (g) Base + SPToken, (h) Base + LearnProc, (i) Base + SpectrAttn.

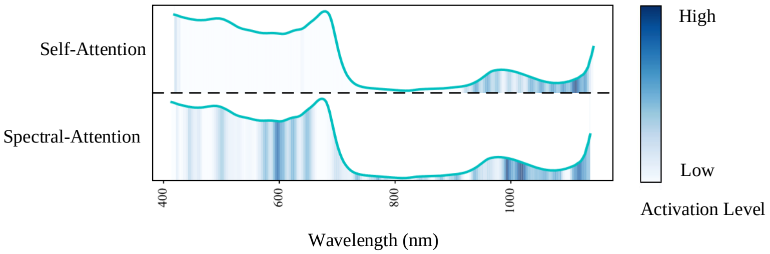


Figure 5: Attention maps of on Mango_DMC dataset: attention maps from self-attention (upper) and spectral-attention (lower). Activation levels are marked with blue, where darker color stands for higher activation and vice versa.

qualitative results are derived from a single experiment which might deviate from the results in the above tables (average results of 5 repeated experiments).

4.6 Interpreting spectral-attention

To interpret ACT, attention maps on the Mango_DMC dataset are presented in Fig. 5. Besides the attention map of spectral-attention within ACT, we also replace spectral-attention with self-attention and extract corresponding attention map for comparison. It is indicated that ACT within spectral-attention generates high activation around the absorbance peak of chlorophyll (680 nm) and O-H (800 – 1000 nm). Meanwhile, ACT without spectral-attention focuses exclusively on the absorbance peak located at 1000 nm and misses out the peak around 680 nm. In this sense, spectral-attention seems to be able to identify characteristic bands.

5 Conclusion

This paper studies the integration of deep learning and analytical chemistry knowledge, aiming to boost IR spectroscopy with novel spectral modeling methods. The calibration transfer problem and the distinctive properties of IR spectra hinder the further application of deep neural networks in IR spectroscopy. We propose the ACT, an IR-spectroscopy-oriented deep learning method that incorporates knowledge from analytical chemists. Within ACT, we design two modules integrated with chemical knowledge, namely learnable spectral processing and spectral-attention. ACT is evaluated in 9 IR spectral data modeling tasks covering calibration transfer, regression, and classification, where ACT outperforms state-of-the-art methods by a considerable margin. We believe this work promotes the development of deep learning in analytical chemistry and facilitates future work.

References

[1] N. Anderson, K. Walsh, P. Subedi, and C. Hayes. Achieving robustness across season, location and cultivar for a nirs model for intact mango fruit dry matter content. *Postharvest Biology and Technology*, 168:111202, 2020. ISSN 0925-5214.

[2] J. Bredenbeck, A. Ghosh, M. Smits, and M. Bonn. Ultrafast two dimensional-infrared spectroscopy of a molecular monolayer. *Journal of the American Chemical Society (JACS)*, 130(7): 2152–2153, 2008.

[3] Y.-Y. Chen and Z.-B. Wang. End-to-end quantitative analysis modeling of near-infrared spectroscopy based on convolutional neural network. *Journal of Chemometrics*, 33(5):e3122, 2019.

[4] Z. Chen, R. Zhou, and P. Ren. Spectraformer: deep learning model for grain spectral qualitative analysis based on transformer structure. *RSC Advances*, 14:8053–8066, 2024.

[5] C. Cui and T. Fearn. Modern practical convolutional neural networks for multivariate regression: Applications to nir calibration. *Chemometrics and Intelligent Laboratory Systems*, 182:9–20, 2018. ISSN 0169-7439.

[6] B. Debus, H. Parastar, P. Harrington, and D. Kirsanov. Deep learning in analytical chemistry. *TrAC Trends in Analytical Chemistry*, 145:116459, 2021. ISSN 0165-9936. doi: https://doi.org/10.1016/j.trac.2021.116459.

[7] A. A. Enders, N. M. North, C. M. Fensore, J. Velez-Alvarez, and H. C. Allen. Functional group identification for ftir spectra using image-based machine learning models. *Analytical Chemistry*, 93(28):9711–9718, 2021. doi: 10.1021/acs.analchem.1c00867. PMID: 34190551.

[8] J. K. Holland, E. K. Kemsley, and R. H. Wilson. Use of fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purées. *Journal of the Science of Food and Agriculture*, 76(2):263–269, 1998.

[9] A. John-Herpin, A. Tittl, L. Kühner, F. Richter, S. H. Huang, G. Shvets, S.-H. Oh, and H. Altug. Metasurface-enhanced infrared spectroscopy: An abundance of materials and functionalities. *Advanced Materials*, 35(34):2110163, 2023.

[10] G. Jung, S. G. Jung, and J. M. Cole. Automatic materials characterization from infrared spectra using convolutional neural networks. *Chemical Secience*, 14:3600–3609, 2023.

[11] D. A. Kalashnikov, A. V. Paterova, S. P. Kulik, and L. A. Krivitsky. Infrared spectroscopy with visible light. *Nature Photonics*, 10(2):98–101, 2016.

[12] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[13] L. Li, H. Zang, J. Li, D. Chen, T. Li, and F. Wang. Identification of anisodamine tablets by raman and near-infrared spectroscopy with chemometrics. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 127:91–97, 2014. ISSN 1386-1425.

[14] Z. Liu, M. Cheng, Z. Li, Z. Huang, Q. Liu, Y. Xie, and E. Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. *In NeurIPS 2023*, 36, 2024.

[15] M. Manley. Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chemical Society Reviews*, 43:8200–8214, 2014.

[16] F. Marini. Artificial neural networks in foodstuff analyses: Trends and perspectives a review. *Analytica Chimica Acta*, 635(2):121–131, 2009. ISSN 0003-2670.

[17] J. Martens, V. Koppen, G. Berden, F. Cuyckens, and J. Oomens. Combined liquid chromatography-infrared ion spectroscopy for identification of regioisomeric drug metabolites. *Analytical Chemistry*, 89(8):4359–4362, 2017.

[18] T. Miao, N. Sihota, F. Pfeifer, C. McDaniel, M. De Gea Neves, and H. W. Siesler. Rapid determination of the total petroleum hydrocarbon content of soils by handheld fourier transform near-infrared spectroscopy. *Analytical Chemistry*, 95(17):6888–6893, 2023. PMID: 37070825.

[19] P. Mishra and D. Passos. Deep calibration transfer: transferring deep learning models between infrared spectroscopy instruments. *Infrared Physics & Technology*, 117:103863, 2021.

[20] P. Mishra, R. Nikzad-Langerodi, F. Marini, J. M. Roger, A. Biancolillo, D. N. Rutledge, and S. Lohumi. Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? the answer is not always. *TrAC Trends in Analytical Chemistry*, 143:116331, 2021. ISSN 0165-9936.

[21] P. Mishra, D. Passos, F. Marini, J. Xu, J. M. Amigo, A. A. Gowen, J. J. Jansen, A. Biancolillo, J. M. Roger, D. N. Rutledge, and A. Nordon. Deep learning for near-infrared spectral data modelling: Hypes and benefits. *TrAC Trends in Analytical Chemistry*, 157:116804, 2022. ISSN 0165-9936.

[22] P. Mishra, D. Passos, F. Marini, J. Xu, J. M. Amigo, A. A. Gowen, J. J. Jansen, A. Biancolillo, J. M. Roger, D. N. Rutledge, and A. Nordon. Deep learning for near-infrared spectral data modelling: Hypes and benefits. *TrAC Trends in Analytical Chemistry*, 157:116804, 2022. ISSN 0165-9936.

[23] P. Mishra, J. M. Roger, F. Marini, A. Biancolillo, and D. N. Rutledge. Pre-processing ensembles with response oriented sequential alternation calibration (prosac): A step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling. *Chemometrics and Intelligent Laboratory Systems*, 222:104497, 2022. ISSN 0169-7439.

[24] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *In ICLR 2023*, 2022.

[25] R. Nikzad-Langerodi, W. Zellinger, E. Lughofer, and S. Saminger-Platz. Domain-invariant partial-least-squares regression. *Analytical Chemistry*, 90(11):6693–6701, 2018.

[26] R. Nikzad-Langerodi, W. Zellinger, S. Saminger-Platz, and B. A. Moser. Domain adaptation for regression under beer–lambert's law. *Knowledge-Based Systems*, 210:106447, 2020.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *In NeurIPS 2019*, 32, 2019.

[28] J. Peng, S. Peng, Q. Xie, and J. Wei. Baseline correction combined partial least squares algorithm and its application in on-line fourier transform infrared quantitative analysis. *Analytica chimica acta*, 690(2):162–168, 2011.

[29] J. U. Porep, D. R. Kammerer, and R. Carle. On-line application of near infrared (nir) spectroscopy in food production. *Trends in Food Science* & *Technology*, 46(2, Part A):211–230, 2015. ISSN 0924-2244.

[30] M. Ramzan, A. Raza, Z. un Nisa, R. M. Abdel-Massih, R. Al Bakain, F. M. Cabrerizo, T. E. Dela Cruz, R. K. Aziz, and S. G. Musharraf. Detection of antimicrobial resistance (amr) and antimicrobial susceptibility testing (ast) using advanced spectroscopic techniques: A review. *TrAC Trends in Analytical Chemistry*, 172:117562, 2024. ISSN 0165-9936.

[31] A. Rinnan, F. van den Berg, and S. B. Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201–1222, 2009. ISSN 0165-9936.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *In CVPR 2016*, pages 2818–2826, 2016.

[33] K. Varmuza and P. Filzmoser. *Introduction to multivariate statistical analysis in chemometrics*. CRC press, 2016.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *In NeurIPS 2017*, 30, 2017.

[35] T. Wang, Y. Tan, Y. Z. Chen, and C. Tan. Infrared spectral analysis for prediction of functional groups based on feature-aggregated deep learning. *Journal of Chemical Information and Modeling*, 63(15):4615–4622, 2023.

[36] X. Wang, S. Jiang, W. Hu, S. Ye, T. Wang, F. Wu, L. Yang, X. Li, G. Zhang, X. Chen, J. Jiang, and Y. Luo. Quantitatively determining surface–adsorbate properties from vibrational spectroscopy with interpretable machine learning. *Journal of the American Chemical Society (JACS)*, 144(35):16069–16076, 2022.

[37] Z. Wang, J. Zhang, X. Zhang, P. Chen, and B. Wang. Transformer model for functional near-infrared spectroscopy classification. *IEEE Journal of Biomedical and Health Informatics*, 26 (6):2559–2569, 2022.

[38] A. Weber, B. Hoplight, R. Ogilvie, C. Muro, S. R. Khandasammy, L. Pérez-Almodóvar, S. Sears, and I. K. Lednev. Innovative vibrational spectroscopy research for forensic application. *Analytical Chemistry*, 95(1):167–205, 2023. PMID: 36625116.

[39] H. Wu, J. Xu, J. Wang, and M. Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *In NeurIPS 2021*, 34:22419–22430, 2021.

[40] Y. Wu, J. Liu, Y. Wang, S. Gibson, M. Osadchy, and Y. Fang. Reconstructing randomly masked spectra helps dnns identify discriminant wavenumbers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(5):3845–3861, 2024.

[41] X. Yang, X. Zhuang, R. Shen, M. Sang, Z. Meng, G. Cao, H. Zang, and L. Nie. In situ rapid evaluation method of quality of peach kernels based on near infrared spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 313:124108, 2024. ISSN 1386-1425.

[42] H. Yu, W. Du, Z.-Q. Lang, K. Wang, and J. Long. A novel integrated approach to characterization of petroleum naphtha properties from near-infrared spectroscopy. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.

[43] Y. Yu, J. Huang, S. Liu, J. Zhu, and S. Liang. Cross target attributes and sample types quantitative analysis modeling of near-infrared spectroscopy based on instance transfer learning. *Measurement*, 177:109340, 2021. ISSN 0263-2241.

[44] J. Zhang, X. Zhou, and B. Li. Pfce2: A versatile parameter-free calibration enhancement framework for near-infrared spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 301:122978, 2023. ISSN 1386-1425.

[45] X. Zhang, T. Lin, J. Xu, X. Luo, and Y. Ying. Deepspectra: An end-to-end deep learning approach for quantitative spectral analysis. *Analytica Chimica Acta*, 1058:48–57, 2019. ISSN 0003-2670.

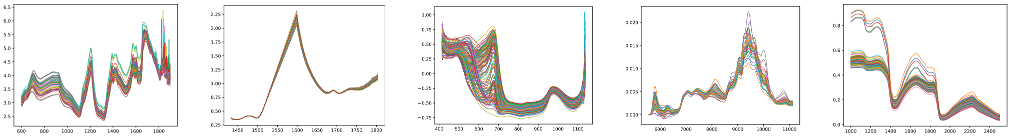
A Dataset details

Tablet [1] dataset is originally proposed in IDRC shoot-out. Tablet contains NIR measurements of pharmaceutical tablets from two spectrometers (referred as spectrometer No. 1 and No. 2 in this paper), ranging from 600 nm to 1898 nm. The spectra are collected from 655 pharmaceutical tablets from production runs and pilot runs. Each spectrum contains 650 sampling points with an interval of 2 nm. The content of active pharmaceutical ingredient (API) varies from 80% to 120% of target value (195 mg), following the requirements of International Conference on Harmonization (ICH) and the requirements of U.S. Food and Drug Administration (FDA). This dataset is divided into calibration set (155 tablets), validation set (40 tablets), and test set (460 tablets). The aim of this dataset is to predict the amount of API (in mg) within the tablets.

Melamine (MF) [2] dataset originates from a batch-condensation process at Metadynea GmbH (Krems, Austria) and consists of NIR spectra from different MF recipes with slightly different compositions. The spectra covers the first and second overtone regions, which are located at wavenumbers at 5546 cm^{-1}- 6254 cm^{-1} (1803 nm - 1598 nm) and 6596 cm^{-1} – 6975 cm^{-1} (1433 nm - 1516 nm). The spectra are recorded in 346 spectral bands. Two recipes, namely R562 and R568, are included in this

[1] https://eigenvector.com/resources/data-sets/

[2] https://github.com/RNL1/Melamine-Dataset



(a) Tablet (b) Melamine(MF) (c) Mango_DMC (d) Strawberry (e) Apple_leaf

Figure 6: Visualization of spectra within five datasets.

paper. There are 3032 samples in R562, while 733 samples in R568. When using a recipe for training, 70% samples are used as training set and the rest 30% are used as validation set. The analytical target of this dataset is turbidity point which indicates the degree of polymerization.

Mango_DMC [3] dataset consists of 11,691 spectra collected from 4675 mango samples across 112 populations and 4 seasons. The first three seasons are used for training (7413) and validating (2830), while the last season is used for testing (1448). Spectra are collected with a F750 Produce Quality Meter, ranging from 300 nm to 1100 nm. There is a 3 nm interval between neighboring spectral bands (242 sampling points), while the optical resolution is 10 nm. The analytical task of this dataset is to predict the dry matter content (DMC) of mango fruit. DMC is an index of total carbohydrates (starch and sugars) and correlates strongly to the Soluble Solids Content (SSC) of ripened fruit, which influences the eating flavor of mango fruits. DMC can also be used a harvest maturity guide, in conjunction with flesh color.

Strawberry [4] dataset contains 983 MIR fruit purees collected by Fourier transform infrared (FTIR) spectrometer with attenuated total reflectance (ATR) sampling. Spectra are recorded with 235 data points ranging from 899 cm^{-1} to 1802 cm^{-1}. Among the spectra, 337 spectra are used for training, 329 spectra are used for validation, and 317 spectra are used for testing. The single-beam spectra of the purees were ratioed to background spectra of water and then converted into absorbance units. Infrared spectroscopy is expected to replace the slow and expensive chemical analyses. The analytical task of this dataset is to detect adulteration in strawberry purees, where the samples are divided into two classes: strawberry purees and adulterated strawberry purees.

Apple_leaf [5] dataset contains 5,490 NIR spectra collected from the leaves of apple trees covering 20 different varieties. Training set consists of 2,500 spectra, validation set consists of 1,250 spectra, while testing set contains 1,740 spectra. Each apple leaf is measured by ten times, deriving ten spectra respectively. The spectrometer utilized in this dataset is ASD FieldSpec 3, which records spectra ranging from 300 nm to 2500 nm. Spectral resolution between 300 nm and 1000 nm is 3 nm, while spectral resolution between 1001 nm and 2500 nm is 6 nm. 300 spectral bands ranging from 1000 nm to 2500 nm are included in the experiments. The analytical task of this dataset is to classify the apple leaves from different varieties.

B Supplementary experiments

B.1 Additional experiments on Tablet dataset

We provide additional experimental results on Tablet tasks in Table 6 and 7. The results of two conventional chemometric methods (PLS and kNNR) are included. These methods are standard solution to the regression problems of infrared spectroscopy. Two evaluation metrices, namely RMSEP(%) and R^2, are also introduced. RMSEP(%) is the ratio of root mean square error to ground truth. The above results further illustrates the superiority of proposed ACT in both calibration transfer and ordinary IR spectral analysis.

Prediction results from additional comparison methods and corresponding ground truth of quantitative tasks are plotted for qualitative analysis, which are presented in Fig. 7. Generally, the prediction results of ACT deviates less from ground truth compared with the two conventional methods.

[3] https://data.mendeley.com/datasets/46htwnp833/1

[4] https://csr.quadram.ac.uk/example-datasets-for-download/

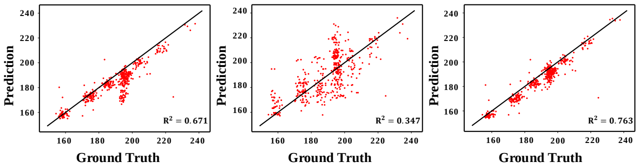
[5] https://www.scidb.cn/en/detail?dataSetId=633694460860956674&version=V1

Table 6: Additional results on Tablet(1,2) and Tablet(2,1) tasks.

	Tablet(1,2)				Tablet(2,1)			
	RMSEP	RMSEP(%)	MAE	R^2	RMSEP	RMSEP(%)	MAE	R^2
PLS	9.038	4.797%	6.349	0.671	16.702	8.865%	16.141	-0.125
	±0	±0	±0	±0	±0	±0	±0	±0
kNNR	12.724	6.754%	9.840	0.347	13.662	7.252%	10.888	0.247
	±0	±0	±0	±0	±0	±0	±0	±0
DeepSpectra	11.676	6.197%	9.762	0.407	14.983	7.953%	13.729	-0.010
	±3.275	±0.017	±3.013	±0.3	±5.08	±0.027	±5.272	±0.62
AggMapNet	10.142	5.383%	7.031	0.582	13.212	7.013%	10.322	0.291
	±0.901	±0.005	±0.746	±0.073	±1.154	±0.006	±1.111	±0.119
TeaNet	11.819	6.274%	9.758	0.382	17.944	9.524%	14.961	-0.367
	±3.673	±0.019	±3.654	±0.383	±4.112	±0.022	±4.159	±0.674
Spectrformer	8.810	4.676%	6.497	0.670	9.279	4.250%	7.445	0.633
	±2.048	±0.011	±1.837	±0.15	±2.239	±0.012	±2.272	±0.189
ACT	**6.941**	**3.684%**	**5.210**	**0.763**	**6.155**	**3.267%**	**4.786**	**0.843**
	±3.255	**±0.017**	**±2.994**	**±0.244**	**±0.985**	**±0.005**	**±1.156**	**±0.046**

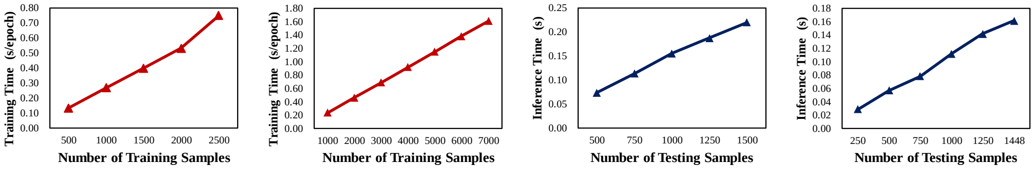
Table 7: Additional results on Tablet(1,1) and Tablet(2,2) tasks.

	Tablet(1,1)				Tablet(2,2)			
	RMSEP	RMSEP(%)	MAE	R^2	RMSEP	RMSEP(%)	MAE	R^2
PLS	5.027	2.668%	3.331	0.898	5.236	2.779%	3.261	0.889
	±0	±0	±0	±0	±0	±0	±0	±0
kNNR	10.987	5.832%	8.349	0.513	11.262	5.978%	8.831	0.488
	±0	±0	±0	±0	±0	±0	±0	±0
DeepSpectra	6.218	3.300%	4.269	0.840	5.738	3.046%	4.054	0.865
	±1.034	±0.005	±0.941	±0.051	±0.676	±0.004	±0.475	±0.034
AggMapNet	5.543	2.942%	4.007	0.875	6.329	3.360%	4.238	0.838
	±0.493	±0.003	±0.609	±0.023	±0.379	±0.002	±0.323	±0.02
TeaNet	5.972	3.170%	4.093	0.855	6.194	3.288%	4.289	0.845
	±0.548	±0.003	±0.392	±0.026	±0.154	±0.001	±0.166	±0.008
Spectrformer	6.679	3.545%	4.404	0.817	5.956	3.162%	3.958	0.857
	±0.839	±0.004	±0.466	±0.048	±0.242	±0.001	±0.191	±0.012
ACT	**4.604**	**2.444%**	**2.914**	**0.913**	**4.417**	**2.344%**	**2.714**	**0.921**
	±0.661	**±0.004**	**±0.688**	**±0.027**	**±0.15**	**±0.001**	**±0.186**	**±0.005**



(a) (b) (c)

Figure 7: Additional qualitative evaluation on the Tablet(1,2) task. (a) PLS, (b) kNNR, (c) ACT.



(a) (b) (c) (d)

Figure 8: Scalability evaluation: (a) training time on Apple_leaf, (b) training time on Mango_DMC, (c) inference time on Apple_leaf, (d) inference time on Mango_DMC.

Table 8: Training time and inference time (both in seconds) on Tablet(1, 2) and Apple_leaf.

	Tablet(1,2)			Apple_leaf(300)	
	Train(Total)	Inference		Train(Total)	Inference
PLS	0.00997	0.00996	SVM	0.49015	1.29785
	Train(Epoch)	Inference		Train(Epoch)	Inference
DeepSpectra	0.01080	0.01395	DeepSpectra	0.20105	0.06777
AggMapNet	0.07982	0.06302	AggMapNet	0.30030	0.23880
TeaNet	0.18502	0.21927	TeaNet	0.75287	0.20958
Spectraformer	0.01796	0.03588	Spectraformer	0.30030	0.23880
ACT	0.09595	0.11137	ACT	0.74937	0.24843

Table 9: Comparison between learnable spectral processing and traditional pre-processing methods in the framework of ACT.

	Tablet(1,2)		Tablet(2,1)		Tablet(1,1)		Tablet(2,2)	
	RMSEP	MAE	RMSEP	MAE	RMSEP	MAE	RMSEP	MAE
ACT_PBC	6.899	5.259	15.585	6.748	4.771	3.126	7.736	3.696
ACT_Deriv	8.032	6.069	9.421	7.091	6.128	4.513	8.006	5.875
ACT_SNV	9.077	6.962	17.975	6.809	8.971	5.432	6.678	4.493
ACT	6.941	5.210	6.155	3.389	4.604	2.914	4.254	3.209

B.2 Efficiency analysis

To evaluate the scalability of ACT, we gradually decrease the number of training and testing samples, aiming to study the relation between computation time and the number of samples. Experiments are conducted on two datasets with relatively abundant samples, namely Apple_leaf and Mango_DMC, with the results presented in Fig. 8. For training time, we record the time cost per epoch as the number of training epochs is not fixed due to early stopping. It is indicated that the training and inference time cost is roughly linear to the number of samples.

Moreover, we also compare ACT with other methods in terms of training and testing time. Deep networks (DeepSpectra, AggMapNet, TeaNet, and Spectrformer) along with traditional methods (PLS for regression, SVM for classification) are included for comparison. The results on Tablet(1, 2) and Apple_leaf are listed in Table 8. The efficiency of ACT is similar to that of other deep learning methods. It is indicated that the traditional methods are more efficient at training stage, as these methods usually have closed form solution. At inference stage, PLS remains efficient compared with the deep learning methods, while SVM fails to retain supremacy in terms of time cost. Additionally, ACC, AUC, and F1 of SVM on Apple_leaf are 0.526, 0.732, and 0.531 respectively.

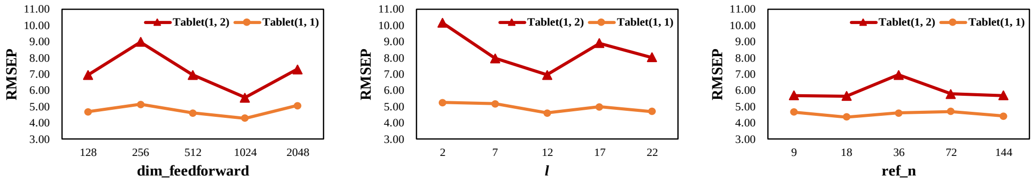
B.3 Additional experiments on pre-processing

To evaluate the effectiveness of proposed learnable spectral processing, an extra experiment has been conducted. Three traditional pre-processing methods are introduced for comparison, namely Standard Normal Variate (SNV), derivatives (Deriv), and polynomial-fitting-based baseline correction (denoted as PBC). Unlike traditional methods that are separate from subsequent classifiers, the proposed module is integrated with ACT and is trained along with the major network. The traditional pre-processing methods are therefore evaluated in the framework of ACT, forming ACT_PBC, ACT_Deriv, and ACT_SNV. The results are listed in Table 9. Although PBS performs slightly better on Tablet(1, 2), original ACT achieves more stable results across the Tablet tasks.

The traditional methods could suffer “over-processing”: removing the noise and the useful information simultaneously. Traditional pre-processing methods rely on certain assumptions to estimate the irrelevant signal within spectra. However, these assumptions are not always suitable for the real-world data. The learnable spectral processing module allows ACT to retrieve features from the removed baselines. Moreover, the learnable parameters also enhance the adaptability of the proposed method.

Table 10: Fine-grained ablation study. Notation '+' means incorporating specific modules.

	Tablet(1, 2)		Mango_DMC		Strawberry_puree		
	RMSEP	MAE	RMSEP	MAE	ACC	AUC	F1
Base	9.977	7.521	1.706	1.277	0.907	0.977	0.904
Base + SPToken	8.491	6.477	1.130	0.854	0.953	0.974	0.953
Base + SPToken + LearnProc'	11.969	8.357	1.459	1.099	0.953	0.973	0.953
Base + SPToken + LearnProc	7.833	5.690	1.083	0.813	0.951	0.979	0.951
Base + SPToken+ SpectrAttn	8.213	6.252	1.009	0.747	0.962	0.990	0.962
Base + SPToken+ IntraSpec	7.562	4.874	1.348	0.973	0.962	0.992	0.962
Base + SPToken+ InterSpec	10.182	6.135	1.151	0.858	0.969	0.993	0.969
ACT	6.941	5.210	1.008	0.756	0.963	0.991	0.964



(a) (b) (c)

Figure 9: The effect of hyper-parameters on Tablet(1, 1) and Tablet(1, 2): (a) dim_feedforward, (b) l, (c) ref_n. Each hyper-parameter varies in a range around the configured value (dim_feedforward: 512, l: 12, and ref_n: 36).

B.4 Additional ablation test

To further evaluate the effectiveness of proposed components, additional fine-grained ablation test is conducted. The baseline reconstruction, intra- and inter-spectral attention are included, forming three auxiliary methods respectively. **LearnProc'** stands for learning spectra processing without baseline reconstruction. **Base + SPToken + LearnProc'** methods tends to evaluate the effectiveness of baseline reconstruction. **IntraSpec** and **InterSpec** denote intra- and inter- spectral attention respectively. The additional results are illustrated in Table 10.

The baseline reconstruction is proved crucial to learnable spectral processing and ACT. This component enable ACT to reclaim useful information from the removed baselines. As mentioned above, the baseline correction method could have the risk of over-processing, leading to mis-removed information in the baselines. Meanwhile, above results indicate the effectiveness of both intra- and inter-spectral attention. Both the two components can improve the performance in certain scenarios, but lack adaptability across different tasks. The combination of these two component (namely spectral-attention) could ensure more stable performance in different scenarios.

B.5 Hyper-parameter analysis

Here we evaluate ACT's sensitivity to hyper-parameter setting. The effects of three hyper-parameters (dim_feedforward, l, ref_n) are evaluated on Tablet(1, 1) and Tablet(1, 2). dim_feedforward controls the projection dimension of spectral-attention which directly influences the learned representation. l determines the size of token patches, and ref_n is the number of reference spectra. For Tablet dataset, the value of dim_feedforward, l, ref_n are set to 512, 12, and 36 respectively. The results are presented in Figure 9. It can be found that ACT is more sensitive to the change of hyper-parameters when dealing with calibration transfer task. Meanwhile, the results also indicate that dim_feedforward, l have greater impact on the performance of ACT. It should also be noted that the reported results of ACT could be sub-optimal, as the performance is further improved when the hyper-parameters deviate from the configured values.

B.6 Additional case study

Attention maps derived from Tablet and Apple_leaf are also studied, which are presented in Fig. 10. It should be noted that spectra in Apple_leaf are recorded as reflectance. In both tasks, ACT tends to

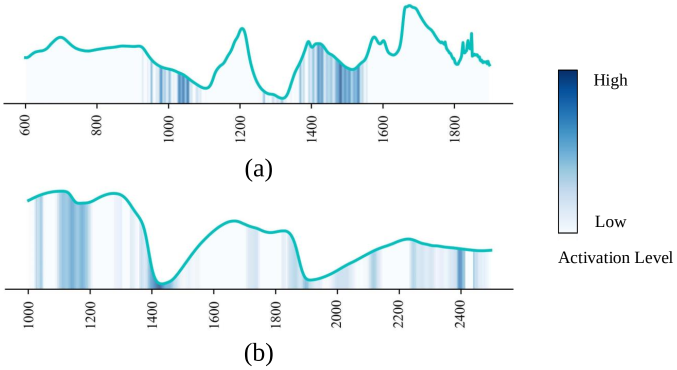


Figure 10: Attention maps of on: (a) Tablet(1, 2), (b) Apple_leaf. Activation levels are marked with blue, where darker color stands for higher activation and vice versa. It should be noted that Apple_leaf records reflectance.

Table 11: The hyper-parameter values of ACT.

	Tablet	Mango_DMC	Strawberry	Melamine	Apple_leaf
l	12	12	12	2	22
n_head	5	5	5	1	1
ref_n	36	36	36	36	36
dim_feedforward	512	512	512	128	256
Learning_rate	0.0005	0.001	0.001	0.002	0.001
FC_layers	[256,64]	[256,64]	[256,64]	[64, 16]	[128, 32]
Drop_out	0.1	0.1	0.1	0	0
Loss	MSE	MSE	CE	MSE	CE

generate high activation values around the absorbance peaks. These attention maps are generally in accord with the prior knowledge in IR spectroscopy. Experts usually select several peaks that are correlated with the prediction targets, and establish calibration model based on the selected peaks.

C Implementation details

C.1 Training settings

During all the above experiments, ACT selects reference spectra from training sets. In classification tasks, this process is fully randomized. In regression tasks, we first sort the training set according to target values and then split the training sets into several subsets (the number of reference spectra). From each subset, a reference spectrum is selected, ensuring the target values (labels) of reference spectra are evenly distributed.

For calibration transfer, the training set and validation set from the source domain are used for training, while the testing set from target domain is used for testing. In Tablet(1, 2) task for example, the training set and validation set are collected by spectrometer No. 1 (source domain), while testing set is collected by spectrometer No. 2.

C.2 Parameter setting

Due to varied samples and capricious spectra-collecting environments, DL methods will suffer serious overfitting or underfitting on certain tasks when tested with fixed hyper-parameters and structure. For thorough evaluation, we adjust the tested methods to simulate the real-world task of IR spectral data modeling. The parameter settings are listed in Table 11–15.

C.3 Details of ablation methods

In this subsection, we further illustrate the details of ablation methods.

Table 12: The hyper-parameter values of AggMapNet.

	Tablet	Mango_DMC	Strawberry	Melamine	Apple_leaf
n_inception	2	3	3	2	3
Conv_size	25	13	13	13	125
Dense_Layers	[256,64]	[128]	[128]	[128]	[512]
Learning_rate	0.005	0.0001	0.0001	0.001	0.0001
Cluster_channel	1	5	1	1	1
Drop_out	0.01	0.1	0.1	0.1	0.05
Loss	MSE	MSE	MSE	MSE	MSE

Table 13: The hyper-parameter values of TeaNet.

	Tablet	Mango_DMC	Strawberry	Melamine	Apple_leaf
ConvLayers	[8, 8, 'M']	[8, 8, 'M']	[64, 64, 'M']	[8, 8, 'M']	[64, 64, 'M']
mask_num	10	2	10	10	2
lr_base_D	0.001	0.001	0.001	0.001	0.001
lr_base_G	0.001	0.001	0.001	0.001	0.001
loss_ratio	0.7	0.7	0.7	0.7	0.7
kernel	21	21	21	21	21
kernel_unet	9	9	9	9	9

Table 14: The hyper-parameter values of Spectraformer.

	Tablet	Mango_DMC	Strawberry	Melamine	Apple_leaf
Conv1	3	3	5	3	5
Conv2	3	3	5	3	5
Conv3	1	3	3	3	3
Conv4	1	3	3	3	3
FC_layers	[128, 32]	[256,64]	[256,64]	[256,64]	[256,64]
Attn_ff	256	256	512	256	512
Loss	MSE	MSE	MSE	MSE	MSE
Learning_rate	0.01	0.01	0.01	0.005	0.001

Table 15: The hyper-parameter values of DeepSpectra.

	Tablet	Mango_DMC	Strawberry	Melamine	Apple_leaf
Kernel_size1	5	5	5	10	5
Kernel_size2	2	2	2	4	2
Kernel_size3	3	3	3	6	3
Stride1	3	3	3	6	3
Stride2	2	2	2	4	2
dim_fc_layer	128	128	64	64	64
Drop_out	0.5	0.1	0.1	0.1	0.001
Learning_rate	0.01	0.01	0.01	0.01	0.01

Base first encodes the spectra with an encoder layer following self-attention mechanism. The encoded features are then fed into two FC layers to get final output. Without the proposed tokenization method, infrared spectra are regarded as a sequence of univariates.

Base + Token has the same network structure as the **Base** method. The only difference is that spectra are tokenized based on the proposed method where the tokens are formed by neighboring patches. Infrared spectra are therefore regarded as sequences of representations.

Base + Token + LearnProc further introduces the whole learnable spectral processing namely the reversible pre-processing, tokenization, and post-processing. As a part of the learnable spectral processing, the proposed tokenization is tested separately to evaluate the effectiveness of the three parts within the learnable spectral processing.

Base + Token + SpectrAttn introduces the spectral-attention along with the proposed tokenization method to the **Base** network. Compared with **Base + Token**, spectral-attention is utilized in place of self-attention.

C.4 Baseline correction

Here, we give detailed descriptions of baseline correction method used in ACT, namely iterative polynomial fitting [6]. In general, a chemical signal (e.g. an IR spectrum) are viewed as the combination of true signal, baseline, and measurement error. Approximating and eliminating the baseline could guide the subsequent learning methods. The baseline correction method incorporated in ACT assumes that baseline within chemical signals can be estimated by a polynomial with lower power. Given an IR spectrum $\boldsymbol{x} \in \mathbb{R}^b$ recording absorbance intensity, we set an initial temporal signal $\boldsymbol{x}_t = \boldsymbol{x}$. Firstly, this method conducts polynomial fitting with the whole signal $\boldsymbol{x}_t$ and generates an estimated baseline $\boldsymbol{x}_b$. The signal will not be perfectly fitted as spectra are complicated signals with higher power. Secondly, the estimated baseline $\boldsymbol{x}_b$ serves as a threshold where the unfitted part above threshold is cut out and replaced by the baseline. The above two steps are carried out iteratively until every part of the original spectrum $\boldsymbol{x}$ is above the estimated baseline $\boldsymbol{x}_b$. The procedure of baseline correction is described in Algorithm 1.

Algorithm 1 The procedure of baseline correction.

```
Input: IR spectrum $\boldsymbol{x}$, power of polynomial $n$
Output: Baseline $\boldsymbol{x}_b$
Initialize $\boldsymbol{x}_t = \boldsymbol{x}$
repeat
    Generate $\boldsymbol{x}_b$ with polynomial fitting on $\boldsymbol{x}_t$: $\boldsymbol{x}_b = \mathrm{PolyFit}(\boldsymbol{x}_t, n)$
    Replace signal peaks above threshold with $\boldsymbol{x}_b$: $\boldsymbol{x}_t = \mathrm{Concat}(\min(\boldsymbol{x}_t^i, \boldsymbol{x}_b^i)_{i=1}^b)$
until $\max(\{\boldsymbol{x}_b^i - \boldsymbol{x}^i\}_{i=1}^b) < 0.001$
```

However, there could be a deviation between the estimated baselines and the true baselines. In other words, estimated baselines could contain the information of true signals. ACT therefore reconstruct the baselines in the learnable spectral processing module to recover the discriminative information.

D Limitations

Although current tokenization method utilized by ACT improves modeling performance (see ablation results), it attaches the size of local window to the dimension of tokens, which might limit the flexibility of proposed model. Meanwhile, the computation cost of spectral-attention is higher than self-attention. Compared with self-attention, intra-spectral attention introduces extra calculations based on correlation map, resulting in $O(b^3)$ complexity. Meanwhile, inter-spectral attention has a complexity of $O(bn_r d_k)$. Since $n_r \ll b$ and $d_k \ll b$, the complexity of spectral-attention is $O(b^3)$. Fortunately, b is usually limited due to the finite optical resolution of spectrometers. Finally, there is still room for improvement in terms of interpretability. Although the final attention maps of spectral-attention are in accord with the prior knowledge of IR spectroscopy, the intermediate attention maps between inter- and intra-spectral attention lack interpretability.

[6] Gan, F., Ruan, G., & Mo, J. (2006). Baseline correction by improved iterative polynomial fitting with automatic threshold. Chemometrics and Intelligent Laboratory Systems, 82(1-2), 59-65.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: [TODO]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Source code will be uploaded for reproduction in rebuttal stage.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Source code will be uploaded for reproduction in rebuttal stage.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: Detailed introduction will be updated latter.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.