scientific reports

Check for updates

OPEN

A deep learning framework for hepatocellular carcinoma diagnosis using MS1 data

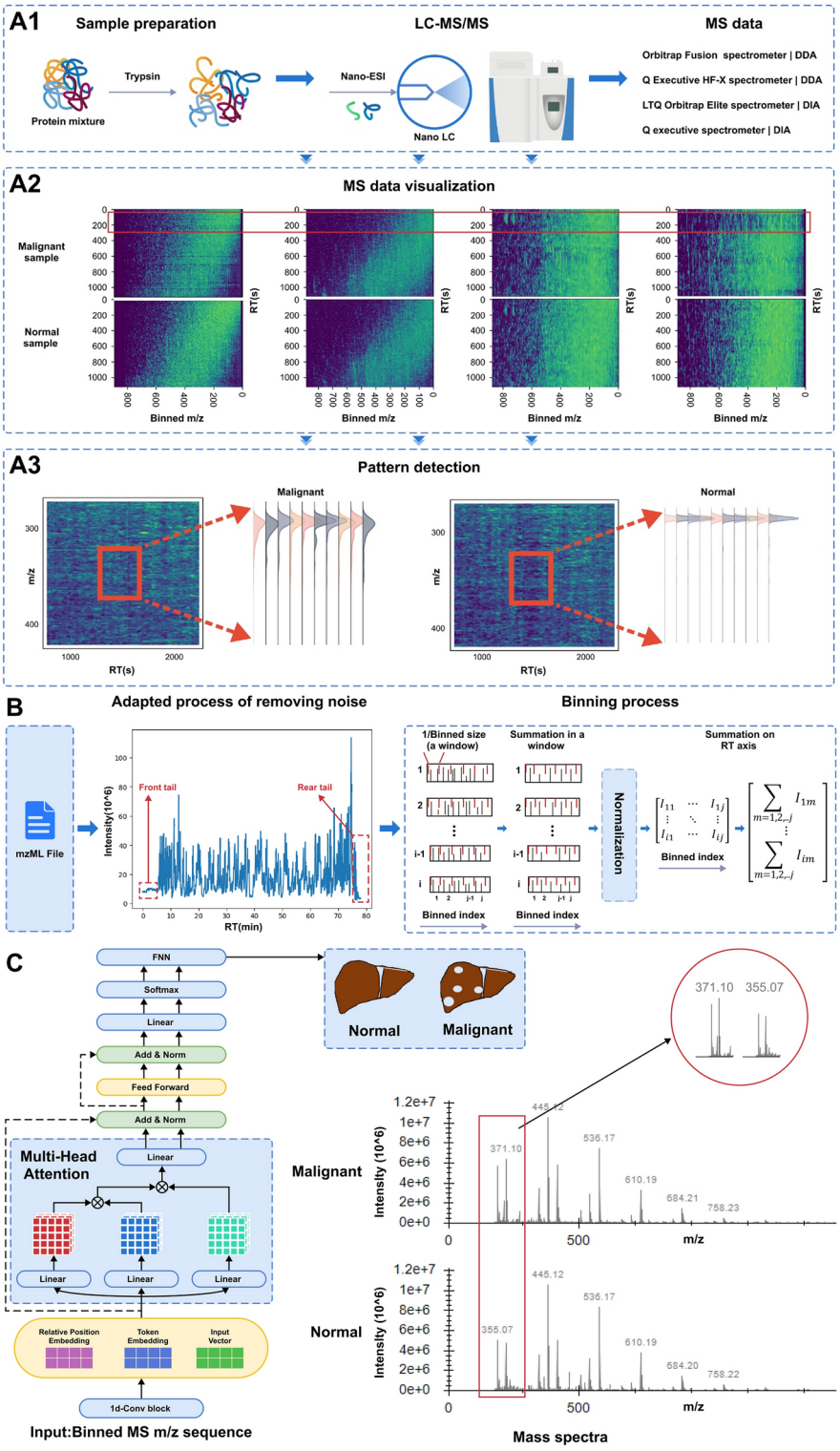
Wei Xu[1,2,12], Liying Zhang[3,12], Xiaoliang Qian[3,12], Nannan Sun[3,12], Xiao Tu[1,8], Dengfeng Zhou[3], Xiaoping Zheng[10], Jia Chen[5,6], Zewen Xie[3], Tao He[3], Shugang Qu[3], Yinjia Wang[9]✉, Keda Yang[11]✉, Kunkai Su[7]✉, Shan Feng[5,6]✉ & Bin Ju[3,4]✉

Clinical proteomics analysis is of great significance for analyzing pathological mechanisms and discovering disease-related biomarkers. Using computational methods to accurately predict disease types can effectively improve patient disease diagnosis and prognosis. However, how to eliminate the errors introduced by peptide precursor identification and protein identification for pathological diagnosis remains a major unresolved issue. Here, we develop a powerful end-to-end deep learning model, termed "MS1Former", that is able to classify hepatocellular carcinoma tumors and adjacent non-tumor (normal) tissues directly using raw MS1 spectra without peptide precursor identification. Our model provides accurate discrimination of subtle m/z differences in MS1 between tumor and adjacent non-tumor tissue, as well as more general performance predictions for data-dependent acquisition, data-independent acquisition, and full-scan data. Our model achieves the best performance on multiple external validation datasets. Additionally, we perform a detailed exploration of the model's interpretability. Prospectively, we expect that the advanced end-to-end framework will be more applicable to the classification of other tumors.

The number of cancer-related deaths has increased globally over the past few decades, with hepatocellular carcinoma (HCC) ranking among the top three in mortality rates. Although effective diagnosis and treatment have made significant progress in the early stages of HCC, the overall survival rate after 5 years is still low (only 50–70%). Clinically, the gold standard method for diagnosing HCC is histopathological observations performed with haematoxylin and eosin (H &E) or immunohistochemical staining, which is time-consuming and subjectivity from pathology specialists. Therefore, the rapid and accurate diagnosis methods for HCC are needed. Mass spectrometry (MS)-based proteomics as a high sensitivity analysis technique has a powerful capacity to quantify numerous proteins in complex biological samples in several hours, which is feasible for the examination of isolated tissue samples in a short time[1,2].

The mainstream method of traditional proteomics analysis is to first identify peptides based on secondary mass spectrometry (MS2), using MS analysis software such as MaxQuant[3] and MSGFPlus[4] to search databases, and then identify potential tumor biomarkers[5,6] and determine the probability of cancer through machine learning methods[7–9]. For example, Zhang et al. found five urine biomarkers associated with lung cancer using random forest algorithm based on urine from 231 patients[10]. Sun et al. employed deep neural networks (DNN) to identify 19 biomarkers of thyroid cancer from 288 thyroid tissue samples[11]. Zhu et al. performed a quantitative

[1]College of Basic Medical Science, Zhejiang Chinese Medical University, 548 Binwen Rd, Hangzhou 310053, China. [2]Key Laboratory of Chinese Medicine Rheumatology of Zhejiang Province, 548 Binwen Rd, Hangzhou 310053, China. [3]SanOmics AI Co., Ltd, Lingping District, Hangzhou 311103, China. [4]Innovative Institute of Basic Medical Sciences, Zhejiang University, Hangzhou 310022, Zhejiang, China. [5]School of Life Sciences, Key Laboratory of Structural Biology of Zhejiang Province, Westlake University, Hangzhou 310024, China. [6]The Biomedical Research Core Facility, Mass Spectrometry and Metabolomics Core Facility, Westlake University, Hangzhou 310024, China. [7]The First Affiliated Hospital, State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Zhejiang University School of Medicine, 79 Qingchun Road, Hangzhou 310013, China. [8]Key Laboratory of Zhejiang Province, Management of Kidney Disease, Hangzhou 310000, China. [9]The First People's Hospital of Kunming, Intensive Care Unit, Kunming 650032, China. [10]Pathology Department, Shulan (Hangzhou) Hospital, Hangzhou, China. [11]Key Laboratory of Artificial Organs and Computational Medicine in Zhejiang Province, Shulan International Medical College, Zhejiang Shuren University, Hangzhou 310015, China. [12]These authors contributed equally: Wei Xu, Liying Zhang, Xiaoliang Qian and Nannan Sun. ✉email: wangyinj@163.com; kdyang@zjsru.edu.cn; ksu@zju.edu.cn; fengshan@westlake.edu.cn; jubin@sanomics.ai



analysis in 19 patients with HCC to quantify 2579 swissProt proteins with high confidence, in which 16 promising candidate tumor markers were highlighted using a cluster heat map approach[12].

However, since certain errors occur in the process of peptide identification and protein, many researchers focus on the application of raw MS data, mainly using machine learning models to mine raw data features to promote clinical application of tumor prediction. Giordano et al. firstly adapted multivariate statistical analysis to collected features of ion peaks from 222 full-scan MS data, and the random forest model subsequently was trained based on features of ion peaks to classify tumor tissues and adjacent non-tumor tissues[13]. Deep learning, as a branch of machine learning[14–16], has the ability to handle large-scale, high-dimensional data[17–20], which makes it an ideal tool for processing raw MS data. In recent years, deep learning models have exhibited exceptional capabilities in various medical diagnostic applications, encompassing fields such as radiomics and proteomics[21–23]. There are also some studies that have leveraged deep learning approaches to analyze raw mass

◀**Fig. 1**. Overall pipeline of MS1Former. (A1) Tissue samples are hydrolyzed into peptide using trypsin, MS data was acquired using four different mass spectrometry instruments containing nano-electrospray ionization (Nano-ESI), (A2) The MS data was visualized, and (A3) Pattern detection for tumor and non-tumor(normal) samples across m/z dimension, the heatmaps show the distribution of the intensities of spectra data, where the color from deep viridis to deep yellow means that the intensity values change from low to large. The density graphs show the distribution of each channel's intensities, for example, the right density plots show the intensity distribution of 340-360 across the m/z dimension. (**B**) Data processing, the front tail and rear tail (highlighted in two red boxes) was removed using the adapted noise removal method; subsequently, binning and normalization on m/z were performed to obtain the processed MS data. (**C**) MS1Former model was trained to diagnose HCC based on input binned m/z sequences. By using the combination of 1d-convolutional neural network (1d-CNN) and transformer encoding module, it achieve an accurate prediction for HCC, which was further verified through the distinguishable features (m/z, intensity), as seen in the red box. The MS data visualization results are generated using ProteoWizard 3.0[38], and the overall graph is created using Figma online.

spectrometry data for the purpose of tumor classification. Wang et al. used a deep learning network to build the MSpectraAI platform based on public MS data of six types of tumor[24]. Zhang et al. converted the MS2 data of DIA into a (cycle, window, m/z) Tensor data structure, which was feed into the ResNet model to obtained a HCC classifier[25]. The above deep learning classification methods all convert raw MS data into 2D or 3D heatmaps. However, due to different settings of some parameters (such as chromatographic gradients) in different batches, the retention time (RT) and intensity information of the same peptide in different batches have great differences, which bring some difficulties for HCC diagnosis and biomarker discovery.

In the field of deep learning, Transformer[26] as a neural network model for sequence modeling[27–30] has been widespread used in natural language processing (NLP)[31–33] and achieves remarkable performance in various tasks such as machine translation[34,35] and information extraction[36,37] due to its ability to capture contextual dependencies between sequences. In this work, we developed a new model architecture named MS1Former (an improved transformer encoder model) for HCC disease prediction in humans from MS1 data. Here, we accumulated MS1 spectra along RT dimension to implement data dimensionality reduction from 2D heatmaps to 1D sequences. The input m/z feature sequence of our method is similar to a text sequence, so the HCC diagnosis task can be regarded as a text classification problem. In this paper, we further explored the capacity of MS1Former to learn highly heterogeneous peptide features and the robustness for the high-resolution mass spectrometry obtained from Orbitrap series instruments.

Results

Study design and analysis workflow

Figure 1 presented the overall pipeline of MS1Former. Wherein, Fig. 1(A1) showed the data were obtained based on the LC-MS/MS method. In detail, the samples were hydrolyzed into peptide fragments by trypsin, and then analyzed by Orbitrap series mass spectrometers equipped with Nano-ESI. Each mass spectrometer adopted different LC conditions, such as gradient time and flow rate, as well as different instrument parameters, including MS1 range. Detailed parameters can be seen in Table 1. Subsequently, we used the ThermoRawFileParser tool to convert raw data (.raw/.wif) into mzXML or mzML format, and the data set (.mzML) was parsed to obtain MS1 spectra(RT, m/z, intensity). Additionally, MS1 spectra were visualized as heatmap images (as seen in Fig. 1(A2)), and the MS1 data collected by the four Orbitrap mass spectrometers does not have a uniform pattern in the RT dimension (as shown in the red box). Figure 1(A3) showed the pattern detection for malignant and normal samples in m/z dimension. The heatmaps showed that the peak intensities distribution in malignant samples is much higher than those in normal samples(as shown in the red box), which means that the peaks of the proteome are more unstable in tumor. The density plots of the m/z are much more consistent and centralized in normal samples while the intensity distribution in malignant samples is dispersed and inconsistent (seen in the red arrow). Therefore, there is a significant difference in the raw data distribution between malignant and normal samples in the m/z dimension.

Figure 1B showed the detailed implementation of the data process in MS1Former. Before the liquid phase peak stabilizes, the interference peaks close to the starting time are removed from the data, and the short signal peaks within a short period of time before the end of the peak are removed from the data to reduce the interference of these impurity peaks on the input signal data. We applied an adapted process of removing noise to remove tailing data treated as noise at the beginning and end of the MS acquisition process (highlighted

	LC condition		Massspectrometer parameters	
	Gradient (min)	Flow rate (nl/min)	Resolution	Scan range (m/z)
Orbitrap fusion	78	350	120000	300–1400
Q Exactive HF-X	120/39	300	120000	350–1800
LTQ orbitrap elite	98	400	60000	350–2000
Orbitrap fusion lumos tribrid	120	N/A	N/A	N/A

Table 1. Summary of LC conditions and mass spectrometer parameters.

in two red boxes). Then, the binning and normalized process of m/z was performed. Herein a window across the dimension m/z (the range between two red dashed lines is referred to as one window or a bin, *i* means total window number) was split and all peak intensities in each window were summed. After summation, the intensities in each window are normalized by dividing the maximum intensity of all windows in a spectra data[24]. Intensity values of binned m/z were accumulated along the RT dimension to obtain the m/z sequence data (m/z, intensity), which served as input for subsequent modeling.

Finally, we developed a classification model framework with an improved transformer encoder model in MS1former to diagnose malignant and normal HCC tissue in humans (as seen in Fig. 1C). The MS1Former framework mainly consists of a CNN layer, a transformer encoder module, and a feed forward neural network block for classification. The encoder module of transformer is mainly composed of the multi-head attention mechanism that transform every position into the input sequence by computing a weighted sum across the representations of all other positions in the sequence, feed-forward neural network, residual connections and layer normalization. Here, the binned m/z sequences obtained from the data processing were subjected to weighted combinations through CNN kernels to obtain a local feature embedding for each m/z. The transformer encoder module further performed encoding on the hidden vector, which captures long-range dependencies between each m/z and the rest of the sequence. Due to the encoder model directly interacts with all other positions in the sequence, information can transform better between distant elements. So the model can effectively gather information from all relevant regions of the m/z sequence to improve the classification accuracy of malignant and normal tissues. Additionally, raw MS1 spectra of malignant and normal tissues were exhibited to further verify the high-performance of classification models through the distinguishable part of raw MS data (as seen in the red box).

Model performance evaluation and comparison

We performed five-fold cross-validation using the dataset PXD006512 (80% for training, 20% for validation) to evaluate the performance of the model. We obtained an mean accuracy of 0.934, mean precision of 0.926, mean recall of 0.930, and mean F1 score of 0.929 (Table 2). The ROC curve of four different model on PXD006512 dataset is shown in Fig. 2A. Figure 2B shown a UMAP plots indicating the separation between malignant and normal groups through latent space of m/z for PXD006512 dataset. To test whether MS1Former is generic, the model is applicable to other external test datasets including WL-2023, WL-Fast, PXD002171 and PXD021979. As seen in Table 3, for each dataset, the accuracy was over 0.84, and the highest accuracy was 0.95 on the PXD021979. The precision of the four test datasets was all over 0.89 and the Recall was over 0.82, meanwhile, the F1-score was over 0.85 and AUC was over 0.90 for four test datasets. The WL-Fast test datasets exhibited lowest assessment on accuracy, precision, recall, F1-score and AUC among those datasets because of the lack of information when adopting the full scan spectra with fastest gradient(only 39 minutes). The detailed classification results of the four external test datasets are shown in Fig. 2C.

We further compared the classification performance of MS1Former with other deep learning-based methods containing MSpectraAI[24], MaxQuant+RandomForest(MaxQuant+RF), and ResNet-18[25] based on the four external test datasets including WL-2023, WL-Fast and PXD002171 and PXD021979. The MSpectraAI, MaxQuant+RF, and ResNet-18 was likewise trained with PXD006512, notably, for the MaxQuant+RF, the protein identification was firstly performed to obtain potential relative proteins, and then the classified model was developed based on those potential proteins, the detail method was described in[25]. Our results showed that MS1Former outperformed other methods, as evaluated by the accuracy and AUC, nevertheless, the ResNet-18 indicated a worst generalization performance on the four test datasets (Fig. 2D, E). Particularly, the WL-Fast dataset was full scan data and there is no MS2 to identify protein via MaxQuant, so the MaxQuant+RF was not suitable to the WL-Fast. From the above result analysis, we can see that MS1Former has good performance in tumor and non-tumor samples classification.

In most cases, the range of ion m/z scans and the number of peaks in each spectrum were first discretized[19]. We employed a binning method that the whole m/z range was divided into equal bin (window), and the bin size (window size) here can be designed freely according to the complexity of data to improve the model's performance (Detailed in Methods). We used Ten bin sizes containing 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 for the samples to determine the peak-intensity distribution that affect model's performance as evaluated by accuracy, F1-score and AUC based on PXD006512. As seen in Fig. 2F, the model's performance gradually improved as the the bin size decreases.The accuracy, F1-score, and AUC of MS1Former was highest while bin size set as 0.1, the model learn high-resolution specific characteristic distribution of peptide while the peak-intensity distribution become more refined as the decrement of bin size. So, our results showed bin size of 0.1 or even higher should be determined to improve model's performance.

	Accuracy	Precision	Recall	F1-score
Fold0	0.926	0.918	0.929	0.923
Fold1	0.963	0.960	0.964	0.962
Fold2	0.916	0.900	0.914	0.910
Fold3	0.953	0.951	0.948	0.949
Fold4	0.906	0.897	0.907	0.902
Mean results	0.934	0.926	0.930	0.929

Table 2. 5-fold cross-validation results of MS1Former on the PXD006512 dataset.

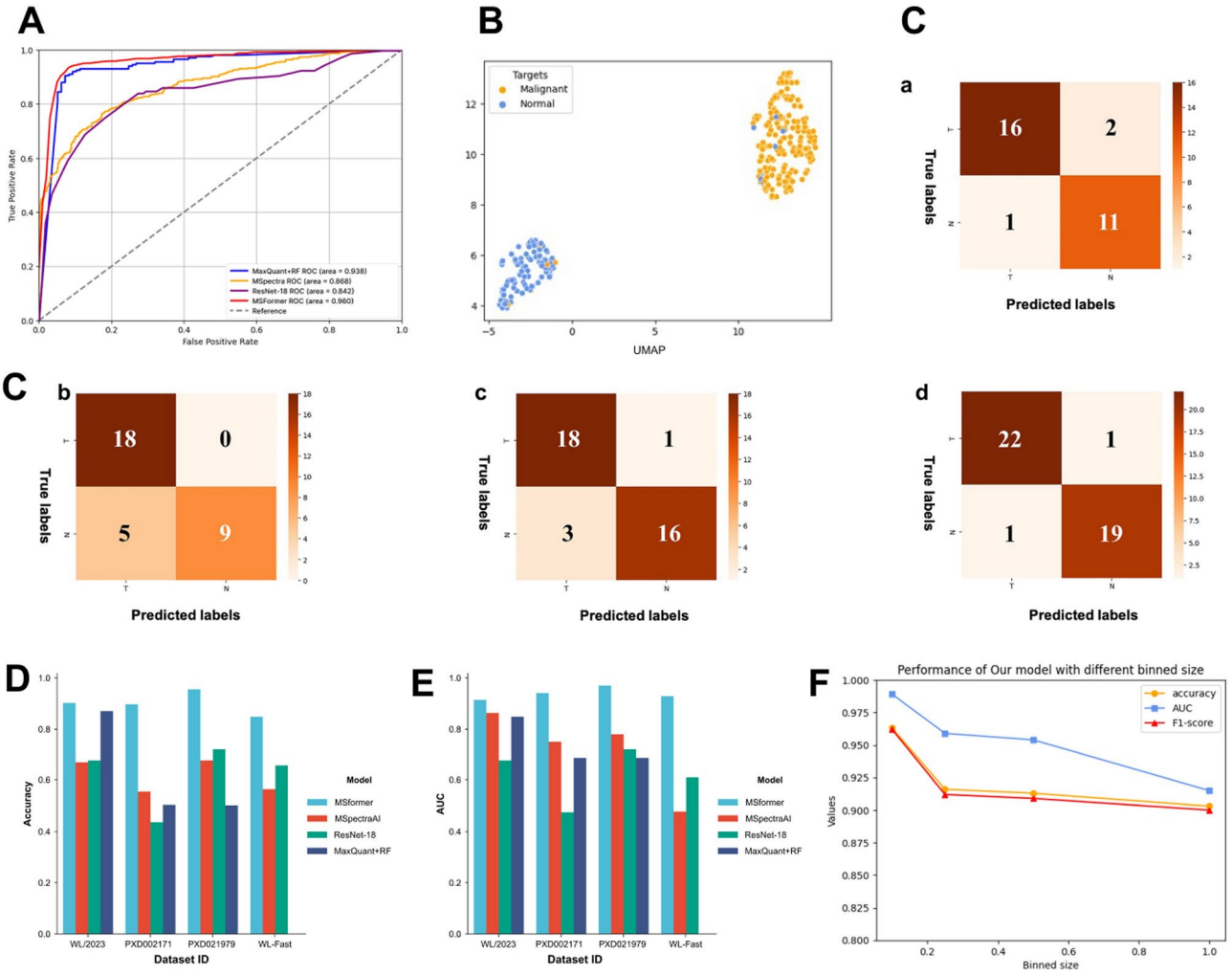


Fig. 2. MS1Former performance validation and testing in independent datasets. (**A**) ROC plot of four different machine learning models on PXD006512 dataset. (**B**) The UMAP plots showing the separation between malignant and normal groups in the PXD006512 dataset with latent space of m/z features. (**C**) Confusion matrix on four external test datasets including WL-2023, WL-Fast, PXD002171, and PXD021979, wherein, a showed the confusion matrix of WL-2023; b showed the confusion matrix of WL-Fast; c showed the confusion matrix of PXD002171; d showed the confusion matrix on PXD021979. (D, E) Performance comparison of Ours(MS1Former), MSpectraAI, MaxQuant+RF and ResNet-18 on four external test datasets including WL-2023, WL-Fast, PXD002171 and PXD021979. (F) The performance of the different binned size for the HCC prediction using MS1Former on dataset PXD006512.

	WL-2023	WL-Fast	PXD002171	PXD021979
Accuracy	0.90	0.84	0.90	0.95
Precision	0.90	0.89	0.90	0.95
Recall	0.90	0.82	0.90	0.95
F1-score	0.90	0.86	0.90	0.95
AUC	0.92	0.93	0.94	0.97

Table 3. Performance of MS1Former on four testing sets consisted of WL-2023, WL-Fast, PXD002171 and PXD021979.

Model interpretability

LIME (Local Interpretable Model-Agnostic Explanations)[39,40] analysis was used to identify the metabolites and the corresponding mass spectrometry-based method that contributed the most to the prediction of HCC using the WL-2023 data set. LIME mainly uses ablation methods to identify which binned m/z has more important influence on the classified results, meanwhile, give the importance value for the binned m/z. We have listed the

top ten m/z binning index using WL-2023 that had the highest contribution to a correct HCC prediction (Fig. 3A). When obtaining key classification features, we restored the m/z information in the RT time dimension to provide data information for peptide analysis. The raw m/z values corresponding to the top ten m/z binning indexes is also given in the Supplement A. Some of the raw m/z values corresponding to binned value 3333 are shown in Fig. 3B. Similar RT and m/z values are divided into one group, and different groups can be identified as different peptides, as described in DirectMS1[41,42]. Later, we can further study the segmented merging of adjacent RT times to better improve the classification prediction ability of the model.

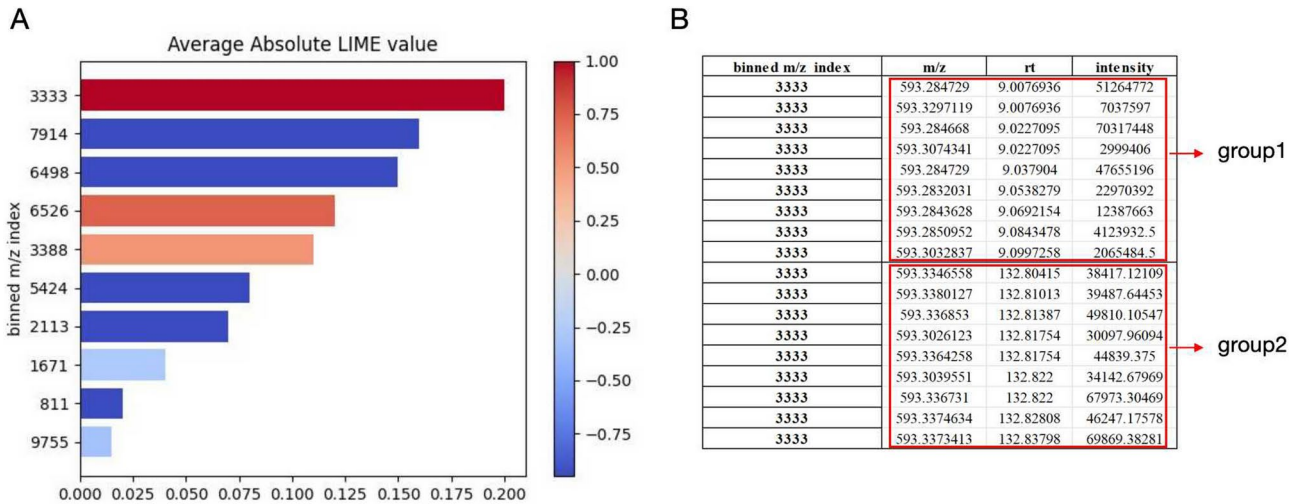
Additionally, Fig. 4 showed beeswarm plot[43] of important binned m/z features for WL-2023 and corresponding mass spectrometry data, the numbers of important features for malignant tissues was larger than that for normal tissues. The significant differences of binned m/z features shown in the beeswarm distribution between malignant and normal tissues was mapped with the raw MS. For example, as the range of binned m/z was 3400-4100, the significantly remarkable difference of the m/z distribution between HCC malignant and normal tissues, besides, the corresponding raw mass spectrometry also present a notable differences (seen in red box). Therefore, we observed that MS1Former is capable of effectively learning distinguishable m/z feature distributions between normal and malignant tumor tissues, which can better guide the inference of the model.

Discussion

In recent years, deep learning models have demonstrated remarkable performance in various medical diagnostic applications, including radiomics and proteomics. For instance, a systematic review by Martinino et al. highlighted the potential of artificial intelligence in enhancing the detection and characterization of HCC through the analysis of imaging data, such as CT, MRI, Ultrasound (US).[23] In this study, we focused on the use of mass spectrometry data for HCC diagnosis.

We established an end-to-end framework named MS1Former allowing us to directly analyze raw data with serialized LC-MS featurlization. This framework involves the featurlization and classification of raw MS data based on the deep learning method. The potential of MS1 data for HCC diagnosis is confirmed. The input m/z feature sequence of our method is similar to a text sequence, so the HCC diagnosis task can be regarded as a text classification problem. Additionally, our method bridges DDA and DIA proteomics directly to deep-learning technology such as a transformer encoder network, which can handle DDA, DIA, and full scan data. We foresee the potential of MS1Former framework in rapid clinical diagnosis, which could be applied to the diagnosis of many tumor types.

Several prior investigations have explored the integration of mass spectrometry data with machine learning methodologies for disease diagnosis. However, many of these approaches are limited by their reliance on additional data preprocessing or specific requirements for mass spectrometry data. In contrast, our proposed methodology, MS1Former, offers distinct advantages. For example, the MaxQuant+RF approach requires the identification and quantification of proteins before applying random forest models for prediction. Conversely, our method enables direct classification without the need for these additional steps. The ResNet-18 architecture was initially designed for image data analysis. Researchers have adapted it for mass spectrometry data by transforming MS2 data into three-dimensional tensors indexed by circle, m/z, and window, and then using the ResNet-18 framework for classification. In comparison, MS1Former is capable of directly modeling based on MS1 data, thereby simplifying the workflow and avoiding the need for complex data transformations. While MSpectraAI also employs MS1 data for disease diagnosis, it does not explicitly account for the relationships



binned m/z index	m/z	rt	intensity
3333	593.284729	9.0076936	51264772
3333	593.3297119	9.0076936	7037597
3333	593.284668	9.0227095	70317448
3333	593.3074341	9.0227095	2999406
3333	593.284729	9.037904	47655196
3333	593.2832031	9.0538279	22970392
3333	593.2843628	9.0692154	12387663
3333	593.2850952	9.0843478	4123932.5
3333	593.3032837	9.0997258	2065484.5
3333	593.3346558	132.80415	38417.12109
3333	593.3380127	132.81013	39487.64453
3333	593.336853	132.81387	49810.10547
3333	593.3026123	132.81754	30097.96094
3333	593.3364258	132.81754	44839.375
3333	593.3039551	132.822	34142.67969
3333	593.336731	132.822	67973.30469
3333	593.3374634	132.82808	46247.17578
3333	593.3373413	132.83798	69869.38281

Fig. 3. The result of Local Interpretable Model-Agnostic Explanations (LIME) for MS1Former. (**A**) LIME values shown for the top ten binned m/z index using WL-2023 that had the highest contribution to a correct HCC prediction. The average correlation corresponds to whether the feature is up- (red) or down- (blue) regulated. (**B**) Take binned value 3333 as an example, we listed some of the raw m/z values. The similar RT and m/z values are divided into one group, and different groups can be identified as different peptides.

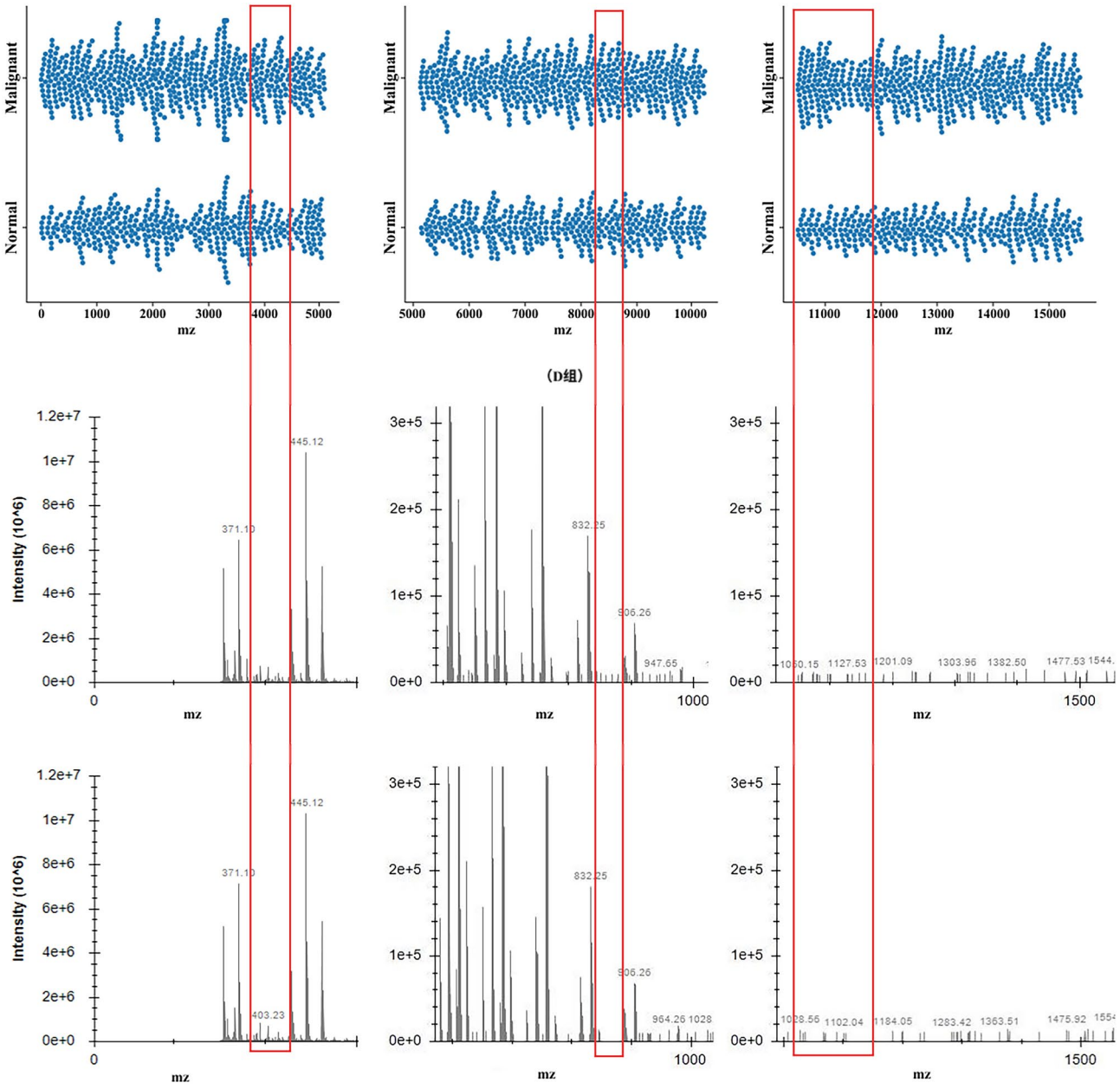


Fig. 4. Beeswarm distribution of important features for WL-2023 and corresponding raw mass spectra data. The certain binned m/z (binned size was set as 0.1) intervals is corresponding to the raw mass spectra data, as seen in red box.

among different scans within the same sample, classifying each scan independently. In contrast, MS1Former integrates all scan information from a sample, facilitating end-to-end diagnostic predictions that capture the comprehensive characteristics of the data. Through evaluations on various HCC datasets, we demonstrate that MS1Former consistently outperforms MaxQuant+RF, MSpectraAI, and ResNet-18 across multiple classification metrics, underscoring its superior performance and versatility in HCC diagnosis.

Overfitting and underfitting are inevitable risks in the field of deep learning. Especially, in the study of small sample classification, overfitting is a noteworthy problem. Here, we avoid overfitting in our work by adopting a relatively larger dataset to increase the sample size. Therefore, the MS data (PXD006512) of 220 HCC samples consisting of 1488 raw files was used as the training set. In this data set, tumor and paired non-tumor tissues in clinically early HCC patients were collected and analyzed by Orbitrap Fusion mass spectrometer. However, the MS data collection just through Obitrap series instruments, so in the future, more types of MS datasets were collected to validate our model. Additionally, during the training process, we implemented a combination of techniques, including dropout, L2 regularization, and early stopping, to further alleviate the overfitting problem.

Binning is a significant step to discretize data suitable for the deep learning method. As the bin size decreases, the precision of identifying peptides increases, which leads to a significant improvement in the model

performance (seen in Fig. 2F). Due to the memory and time-consuming limitations of the training machine, we only conducted experiments with bin sizes between 0.1 and 1. Additionally, although we can discover significant binned m/z features through the interpretability method, and then obtain a lot of important m/z features, it's not that precise. In the future, we can try to carry out more experiments on MS2 and identify biomarkers as a further research direction.

In order to further evaluate the ability of our model to judge the disease course and identify the key discriminant components in the model. It can be found that the differences in the pathological tissue images can be well captured by the model (seen in the corresponding feature heatmaps), and the characteristic differences are shown on the raw mass spectral peaks, the results are shown in Fig. 5. The figure shown that the model learned the m/z differences of different HCC tissue differentiations. In the future, we also can further train the model for the judgment of pathological staging and identify the key biological influencing factors that affect classification based on the model.

In conclusion, despite the promising results, our approach has several limitations that need to be addressed in future work:

1. Generalizability: The datasets used for training and testing were primarily generated using Orbitrap series mass spectrometers, and the generalizability of the model to other types of mass spectrometers and experimental conditions has yet to be fully established. To assess the performance of our method on non-Orbitrap series mass spectrometers, we conducted additional evaluations using a dataset (PXD004837) obtained from a TripleTOF 5600 mass spectrometer. Our model achieved an AUC of 0.8442 on this dataset. For a more detail analysis, please refer to Supplement B. Although this AUC is lower than that observed on the Orbitrap series datasets, it still indicates a certain degree of generalizability of our model. Even though, we acknowledge that further investigation is necessary to fully determine the model's performance across a broader range of mass spectrometry platforms. In future work, we plan to expand the diversity and size of the datasets to rigorously evaluate the robustness and generalizability of the model. Additionally, we will consider extending our approach to other tumor types, thereby enhancing the versatility and applicability of our framework.

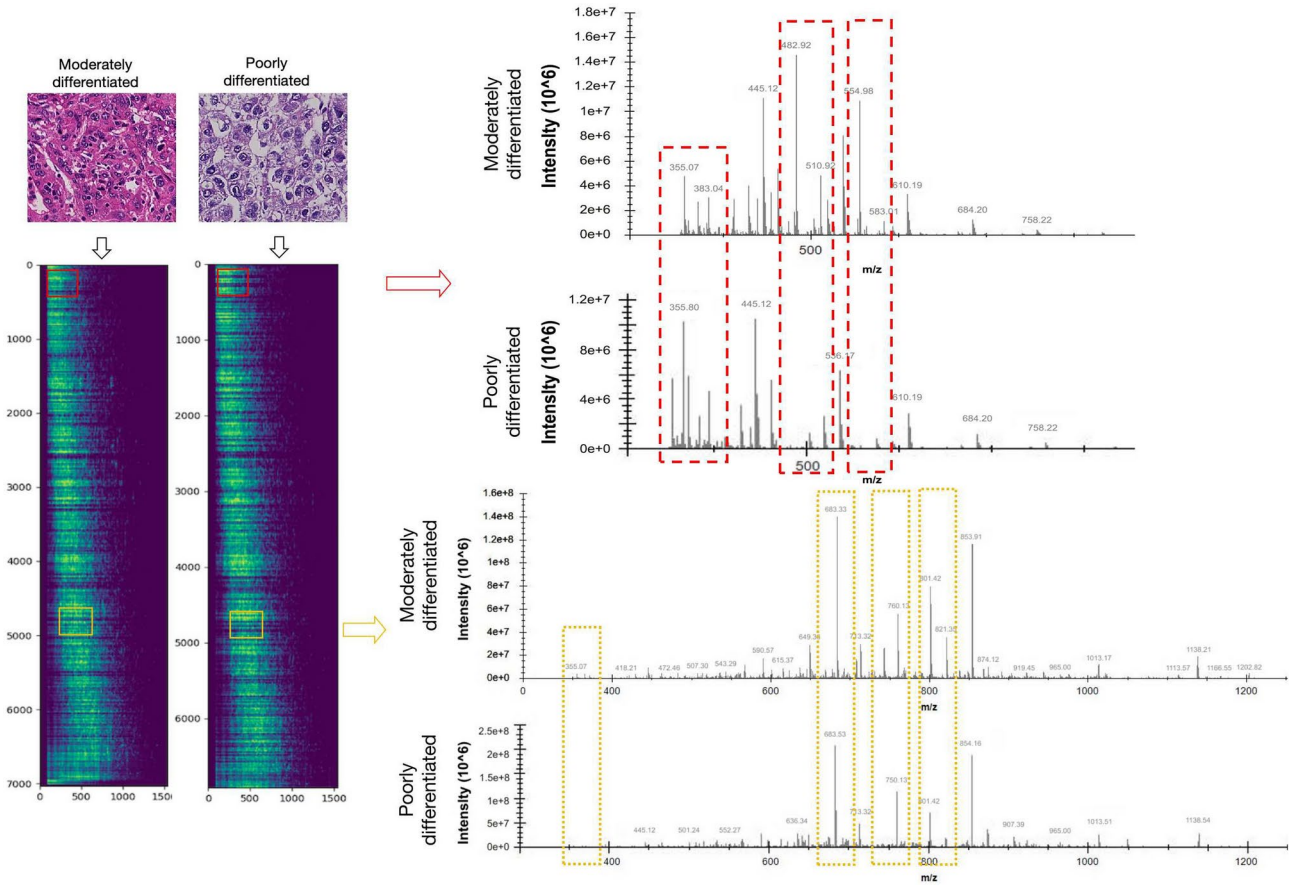


Fig. 5. Differences between different tumor differentiation types in pathological images, heatmap feature images and raw mass spectrometry data. The differences in the pathological tissue images can be well captured by the model (seen in the corresponding feature heatmaps), and these differences can be verified on the distinguishable features (m/z, intensity) on the raw mass spectrum peaks. The MS data visualization results are generated using ProteoWizard 3.0[38], and the overall graph is created using Figma online.

2. Biological Interpretation: While the interpretability framework LIME has been instrumental in identifying significant m/z bins, the biological relevance of these features remains inadequately explored due to the lack of MS2 data. Future work will focus on integrating MS2 data to perform peptide sequencing and to identify specific proteins and peptides corresponding to the significant m/z bins. This integration will enhance the biological interpretability of the model and facilitate the identification of potential biomarkers. To achieve this, we plan to leverage existing research on predicting proteins and peptides from MS2 spectra and incorporate these advancements into our framework. For example, the recently proposed MonoMS1 method[44] attempts to use peptides identified by MS2 spectra as labels and train a peptide recognition model based only on MS1 features, which provides a valuable reference for our future research. By doing so, we aim to bridge the gap between the identified m/z features and their corresponding biological entities, thereby enhancing the translational potential of our findings.By addressing these limiting, we believe the MS1Former framework can be further refined and applied to the clinical diagnosis of various tumor types, contributing to the advancement of precision medicine.

Methods

Hepatocellular Carcinoma dataset. The publicly available dataset PXD006512, PXD002171, and PXD021979 were obtained from Proteomics Identifications Database (PRIDE), more details are provided in previous publication[45–47]. Furthermore, all tissue samples for WL-2023 and WL-Fast were collected from the Shulan (Hangzhou) hospital, Hangzhou, China. This study was approved by Research Ethics Committee of Shulan(Hangzhou) Hospital(KY2023033). The study has obtained informed consent from the subjects or their guardians, and all subject information has been desensitized. All methods were performed in accordance with the relevant guidelines and regulations. For details on the processing of MS data for WL-2023 and WL-Fast, see Supplement B. The details of the five data sets were shown in Table 4.

Adapted process of removing noise. To reduce the interference of impurity peaks on the input signal data, an adapted function for elimination of tails was proposed, as seen in Fig. 1B. The steps as follow:

1. Firstly, summation of the intensities of all m/z at each RT point.
2. Subsequently, $D_{value}(r_i, r_{i+1})$ is the peak summed intensities difference between two adjacent retention time points r_i, r_{i+1}, where *i* represents the RT, *R* represents the maximum of RT, $0 \leqslant i \leqslant R$, and D_{value} as follows:

$$D_{value}(r_i, r_{i+1}) = \left| \sum_{j=0}^{M} intensity_{i,j} - \sum_{j=0}^{M} intensity_{i+1,j} \right| \tag{1}$$

where *j* represents as m/z and *M* is the maximum of m/z.

3. Finally, $D_{value}(r_i, r_{i+1})$ compare with the cutoff value, if it is greater than the cutoff value, then remove the summed intensity values before *i*. In this paper, the cutoff value is set as 3×10^7, which is counted from the raw MS data.

Binning of m/z. Binning index: $I = round(\frac{m/z - m_{min}}{\gamma})$; γ: bin size 0.1; $m_{min} = 260$; $m_{max} = 1800$; The intensity value corresponding to each binning index is the maximum intensity value within the interval; if there is no value in bin index, the intensity value is 0.

Model framework. The MS1Former framework consists of three parts: a 1-dimensional CNN with pooling, which is mainly used to perform word packet training operations in natural language process (NLP) to capture information between adjacent tokens within a local range; a transformer encoder block with relative position vector encoding was used to capture long-range interactions across m/z sequences[48]. Finally, the feedforward neural networks is used for HCC classification. The detailed model framework is shown in Fig. 6, and the settings for the key hyperparameters of the model are given in Supplement B.

Multihead Attention layers was used to share information across the sequence and model long-range interactions. Each head has an independent set of weights $w_q \in R^{C \times K}, w_k \in R^{C \times K}, w_v \in R^{C \times K}$, which transform

	Number of raw files	Number of malignant	Number of normal	Instrument	Data acquisition	PXD IDs
Training set	1488	939	549	Orbitrap Fusion	DDA	PXD006512
	30	18	12	Q Exactive	DDA	WL-2023
				HF-X		
Testing set	32	18	14	Q Exactive	DDA	WL-Fast
				HF-X LTQ		
	38	19	19	LTQ Orbitrap Elite	DDA	PXD002171
	43	23	20	Orbitrap Fusion Lumos Tribrid	DIA	PXD021979

Table 4. Summary of five datasets used in our experiments.

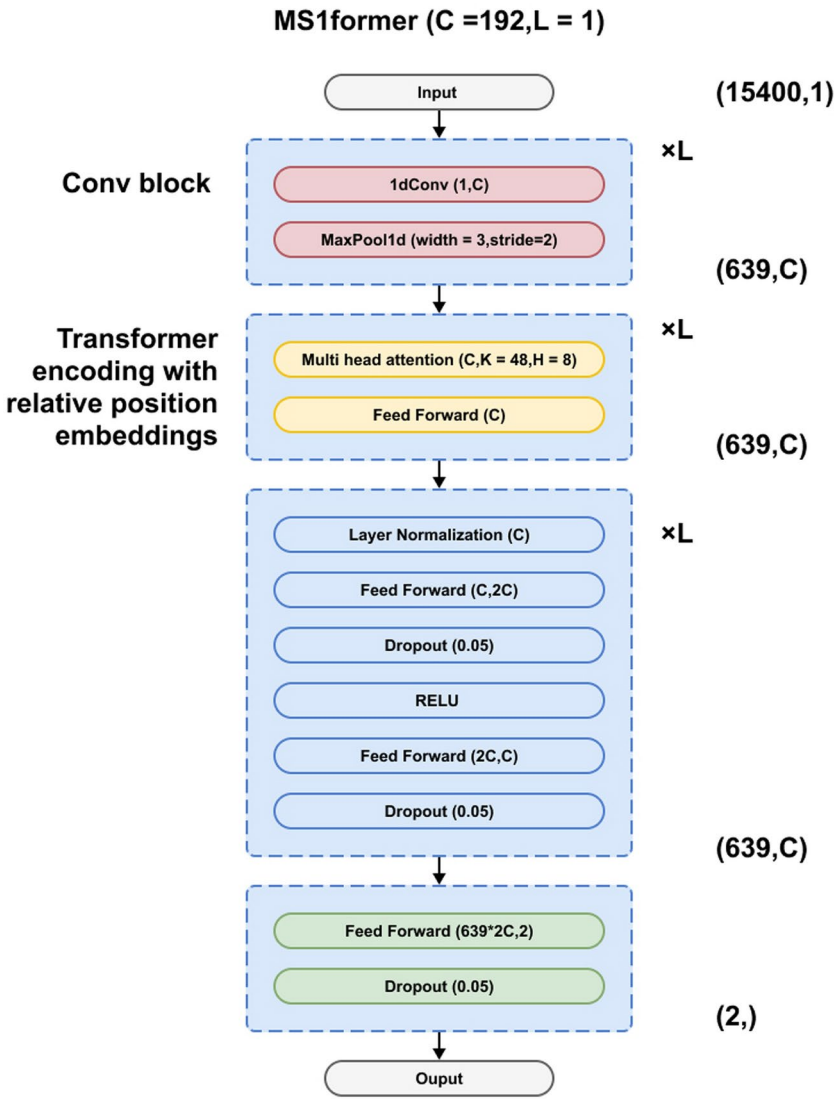


Fig. 6. MS1Former network structure consisted of 1d-CNN block, transformer encoder and feed forward neural network block. The number of trainable parameters for different parts of MS1Former are shown on the right side of the blocks.

the input sequence $X \in R^{L \times C}$ into query $q_i = X_i w_q$, keys $k_i = X_i w_i$,and values $v_i = X_i w_v$. The input consists of queries and keys of dimension d_k. We compute the sum of the matrix of outputs and relative positional encoding R_{ij} as attention matrix, which is determined as $a_{ij} = softmax\left(\frac{q_i k_j^T}{\sqrt{d_k}} v_i + R_{ij}\right)$, where the entry a_{ij} represents the amount of weight query placed at position *i* on the key at position *j*. Values represent the information that each position will propagate to the location that attends to it. Each single attention head computes its output as a weighted sum across all input positions. This allows each query position to use information across the whole sequence. The multiple heads compute with independent parameters, and we concatenate the outputs from each head to form the final layer output followed by a linear layer to combine them. Our layers used 8 heads, a value size of 24, and a key/query size of 48.

For injecting positional information, relative positional[49] encodings R_{ij} were added into $q_i k_j^T$. A parameterized baseline for how actively two positions in the sequence influence mutually during the layer's transformation as a function of their pairwise distance was provided by relative positional encodings. The funcions term as formulated in Enformer[48,50].

MS1Former training details. The 1488 HCC MS samples from public PXD006512 were divided into 5 groups. Four groups were treated as the training set, while the remaining one group was treated as the validation set. Additionally, three sets of DDA-MS (WL-2023, WL-Fast, PXD002171)and one set of DIA-MS (PXD021979) were used for testing. To differentiate spectra from tumor and non-tumor tissue areas, an active function *softmax*(·) is used to transform the output of the previous connection layer into a probability output, with a

leaning rate of 0.001. To prevent overfitting, a dropout layer was employed with a rate of 50%. Additionally, cross-entropy is then used to define the loss function:

$$loss = -\frac{1}{m}\sum^{m} y_{true} \log y_{pred} + (1 - y_{true}) \log(1 - y_{pred}) \quad (2)$$

m is the batch size, which is set to 8, y_{true} is the true label of the input x, and y_{pred} is the predicted score of the input x. During the training process, the model was iteratively trained for 200 times on the data set with Adam optimizer. The MS1Former was implemented in Pytorch and Python 3.8.

Conclusions

We present MS1Former, a powerful deep learning model that integrates CNN and Transformer blocks to facilitate HCC diagnosis using raw MS1 spectra. MS1Former heralds a novel era in the diagnosis of HCC through its ability to directly analyze raw MS1 spectra, offering a promising alternative to conventional, time-consuming diagnostic methods. The high accuracy and efficiency of our model, coupled with its potential for clinical applicability, underscore its potential for HCC diagnostics. In the future, we will focus on further exploring the efficacy of MS1Former in classifying multiple cancer types. Additionally, we plan to integrate MS2 spectra data into the model to enhance the predictive power and interpretability of the model, which is meaningful for advancing our understanding of HCC pathogenesis.

Data availability

The data involved in this study are available from the corresponding author on request. The PXD006512, PXD002171, and PXD021979 were obtained from Proteomics Identifications Database (PRIDE).

Code availability

The codes are freely available at GitHub (https://github.com/sanomics-lab/MS1Former).

Received: 6 February 2024; Accepted: 22 October 2024
Published online: 04 November 2024

References

1. Feng, J., Shang, S. & Beretta, L. Proteomics for the early detection and treatment of hepatocellular carcinoma. *Oncogene* **25**, 3810–3817 (2006).
2. Han, E. C. et al. Direct tissue analysis by maldi-tof mass spectrometry in human hepatocellular carcinoma. *Clin. Chim. Acta* **412**, 230–239 (2011).
3. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
4. Jo, J.-H., Kennedy, E. A. & Kong, H. H. Topographical and physiological differences of the skin mycobiome in health and disease. *Virulence* **8**, 324–333 (2017).
5. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* **49**, 107739 (2021).
6. Karayel, O. et al. Proteome profiling of cerebrospinal fluid reveals biomarker candidates for Parkinson's disease. *Cell Rep. Med.* **3**, 100661 (2022).
7. Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S. & Bacardit, J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS J. Integr. Biol.* **17**, 595–610 (2013).
8. Tyanova, S., Temu, T. & Cox, J. The maxquant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
9. Kim, S. & Pevzner, P. A. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
10. Zhang, C. et al. Urine proteome profiling predicts lung cancer from control cases and other tumors. *EBioMedicine* **30**, 120–128 (2018).
11. Sun, Y. et al. Artificial intelligence defines protein-based classification of thyroid nodules. *Cell Discov.* **8**, 85 (2022).
12. Zhu, Y. et al. Identification of protein abundance changes in hepatocellular carcinoma tissues using pct-swath. *Proteomics Clin. Appl.* **13**, 1700179 (2019).
13. Giordano, S. et al. Rapid automated diagnosis of primary hepatic tumour by mass spectrometry and artificial intelligence. *Liver Int.* **40**, 3117–3124 (2020).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
15. Gessulat, S. et al. Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
16. Tiwary, S. et al. High-quality ms/ms spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525 (2019).
17. Yang, Y. et al. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* **11**, 146 (2020).
18. Ma, C. et al. Improved peptide retention time prediction in liquid chromatography through deep learning. *Anal. Chem.* **90**, 10881–10888 (2018).
19. Xu, L. L., Young, A., Zhou, A. & Röst, H. L. Machine learning in mass spectrometric analysis of dia data. *Proteomics* **20**, 1900352 (2020).
20. Cox, J. Prediction of peptide mass spectral libraries with machine learning. *Nat. Biotechnol.* **41**, 33–43 (2023).
21. Le, N. Q. K. Hematoma expansion prediction: Still navigating the intersection of deep learning and radiomics. *Eur. Radiol.* **34**(5), 2905–2907 (2024).
22. Kha, Q.-H. et al. An interpretable deep learning model for classifying adaptor protein complexes from sequence information. *Methods* **207**, 90–96 (2022).
23. Martinino, A. et al. Artificial intelligence in the diagnosis of hepatocellular carcinoma: A systematic review. *J. Clin. Med.* **11**, 6368 (2022).

24. Wang, S., Zhu, H., Zhou, H., Cheng, J. & Yang, H. Mspectraai: A powerful platform for deciphering proteome profiling of multi-tumor mass spectrometry data by using deep neural networks. *BMC Bioinform.* **21**, 1–15 (2020).
25. Zhang, F. et al. Phenotype classification using proteome data in a data-independent acquisition tensor format. *J. Am. Soc. Mass Spectrom.* **31**, 2296–2304 (2020).
26. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. process. Syst.* **30** (2017).
27. Li, R., Li, L., Xu, Y. & Yang, J. Machine learning meets omics: Applications and perspectives. *Brief. Bioinform.* **23**, bbab460 (2022).
28. Pettini, F., Visibelli, A., Cicaloni, V., Iovinelli, D. & Spiga, O. Multi-omics model applied to cancer genetics. *Int. J. Mol. Sci.* **22**, 5751 (2021).
29. Li, Z., Jiang, X., Wang, Y. & Kim, Y. Applied machine learning in alzheimer's disease research: Omics, imaging, and clinical data. *Emerg. Top. Life Sci.* **5**, 765–777 (2021).
30. Sun, Y. Machine learning for the analysis of multi-omics data. *Methods (San Diego, Calif.)* **189**, 1–2 (2021).
31. Gillioz, A., Casas, J., Mugellini, E. & Abou Khaled, O. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 179–183 (IEEE, 2020).
32. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (2020).
33. Singh, S. & Mahmood, A. The nlp cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access* **9**, 68675–68702 (2021).
34. Huang, F., Zhou, H., Liu, Y., Li, H. & Huang, M. Directed acyclic transformer for non-autoregressive machine translation. In *International Conference on Machine Learning*, 9410–9428 (PMLR, 2022).
35. Raganato, A. & Tiedemann, J. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (The Association for Computational Linguistics, 2018).
36. Zhang, N. et al. Contrastive information extraction with generative transformer. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 3077–3088 (2021).
37. Nguyen, M.-T., Le, D. T. & Le, L. Transformers-based information extraction with limited data for domain-specific business documents. *Eng. Appl. Artif. Intell.* **97**, 104100 (2021).
38. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
39. Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M. & Costa da Silva, E. Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors* **19**, 2969 (2019).
40. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
41. Ivanov, M. V. et al. Directms1: Ms/ms-free identification of 1000 proteins of cellular proteomes in 5 minutes. *Anal. Chem.* **92**, 4326–4333 (2020).
42. Ivanov, M. V. et al. Boosting ms1-only proteomics with machine learning allows 2000 protein identifications in single-shot human proteome analysis using 5 min hplc gradient. *J. Proteome Res.* **20**, 1864–1873 (2021).
43. Peng, Z.-H. et al. Development of machine learning prognostic models for overall survival of prostate cancer patients with lymph node-positive. *Sci. Rep.* **13**, 18424 (2023).
44. Dai, Y., Yang, Y., Wu, E., Shen, C. & Qiao, L. Deep learning powers protein identification from precursor ms information. *J. Proteome Res.* **23**(9), 3837–46 (2024).
45. Jiang, Y. et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* **567**, 257–261 (2019).
46. Naboulsi, W. et al. Quantitative tissue proteomics analysis reveals versican as potential biomarker for early-stage hepatocellular carcinoma. *J. Proteome Res.* **15**, 38–47 (2016).
47. Zhang, Q. et al. Acox2 is a prognostic marker and impedes the progression of hepatocellular carcinoma via pparα pathway. *Cell Death Dis.* **12**, 15 (2021).
48. Avsec, Ž et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
49. Shaw, P., Uszkoreit, J. & Vaswani, A. Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018).
50. Dai, Z. *et al.* Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019).

Acknowledgements

We thank the patients and researchers involved in the public datasets for their data, the Shulan (Hangzhou) Hospital for providing tissue samples and Westlake University for providing MS data preprocessing of WL-2023 and WL-Fast datasets.

Author contributions

Bin Ju conceived the study. Xiaoping Zheng provided the tissue data. Shan Feng and Jia Chen completed mass spectrometry experiment and provided the raw files. Xu wei, Liying Zhang, Nannan Sun, and Xiao Tu implemented the pipeline, constructed the databases, developed the codes and performed. Yinjia Wang, Kunkai Su and Shan Feng peformed data analyses and interpreted all results. Dengfeng Zhou conducted further discussion and analysis. Xiaoliang Qian reviewed the code of the article. Zewen Xie, Tao He and Shugang Qu provided the figures. Wei Xu, Liying Zhang and Nannan Sun wrote manuscript. All authors reviewed the manuscript.

Funding

This work was supported by the Spring City Plan:the High-level Talent Promotion and Training Project of Kunming (Grant No.2022SCP002); the Fundamental Research Funds for the Central Universities (Grant No. 2022ZFJH003) and the National Key Research and Development Program of China (Grant No. 2019YFC0840600 and No. 2019YFC0840609).

Declarations

Competing interests

All authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1

0.1038/s41598-024-77494-4.

Correspondence and requests for materials should be addressed to Y.W., K.Y., K.S., S.F. or B.J.