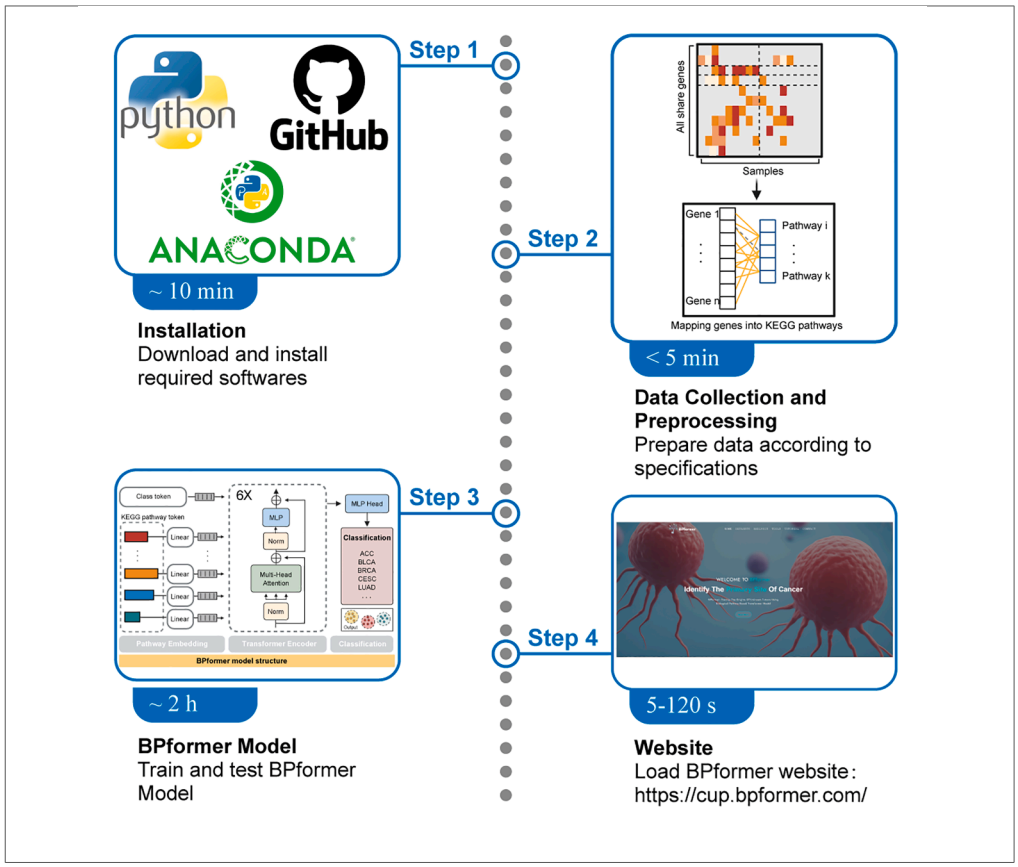
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol to predict the origin of cancer of unknown primary from gene expression data using BPformer



Patients diagnosed with cancer of unknown primary (CUP) have poor prognoses, highlighting the need to identify primary sites for targeted therapeutic intervention. Here, we present a protocol to predict the primary sites of CUP based on gene expression data using BPformer. We describe steps for installing software, collecting and preprocessing data, and training the BPformer model. We then detail procedures on using BPformer via a visualized web server.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Jiajing Xie, Tongxu Qi, Ying Chen, Chaoyong Yang, Mengsha Tong, Rongshan Yu

mstong@xmu.edu.cn (M.T.)
rsyu@xmu.edu.cn (R.Y.)

Highlights

Steps for installing BPformer and preparing software environment

Steps for collecting and preprocessing gene expression data

Steps for training BPformer and performing evaluation

Instructions for deploying BPformer via web server

Xie et al., STAR Protocols 6, 103881
December 19, 2025
https://doi.org/10.1016/j.xpro.2025.103881

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol to predict the origin of cancer of unknown primary from gene expression data using BPformer

Jiajing Xie,[1,7,9] Tongxu Qi,[2,3,9] Ying Chen,[4] Chaoyong Yang,[2,3] Mengsha Tong,[1,5,*] and Rongshan Yu[1,4,6,7,8,*]

[1]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian 361102, China

[2]College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China

[3]State Key Laboratory of Physical Chemistry of Solid Surfaces, Xiamen University, Xiamen 361005, China

[4]School of Informatics, Xiamen University, Xiamen, Fujian 361005, China

[5]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Faculty of Medicine and Life Sciences, Xiamen University, Xiamen, Fujian 361102, China

[6]Aginome Scientific, Xiamen, Fujian 361005, China

[7]Technical contact

[8]Lead contact

[9]These authors contributed equally

*Correspondence: mstong@xmu.edu.cn (M.T.), rsyu@xmu.edu.cn (R.Y.)
https://doi.org/10.1016/j.xpro.2025.103881

SUMMARY

Patients diagnosed with cancer of unknown primary (CUP) have poor prognoses, highlighting the need to identify primary sites for targeted therapeutic intervention. Here, we present a protocol to predict the primary sites of CUP based on gene expression data using BPformer. We describe steps for installing software, collecting and preprocessing data, and training the BPformer model. We then detail procedures on using BPformer via a visualized web server.

For complete details on the use and execution of this protocol, please refer to Xie et al.[1]

BEFORE YOU BEGIN

Metastasis accounts for approximately 90% of cancer-related deaths and is the primary driver of cancer mortality.[2,3] Among them, a significant subset of cancer patients cannot have their primary tumor sites accurately determined by conventional diagnostic methods and are diagnosed as cancer of unknown primary (CUP).[4,5] Patients with CUP generally have poor prognoses, with median survival typically less than one year.[6,7] Traditional diagnostic approaches identify the primary site in only about 25% of CUP cases,[8–11] and treatment outcomes remain limited. Given the therapeutic challenges and clinical uncertainties, developing auxiliary diagnostic tools for CUP is crucial.

This protocol introduces BPformer, a Transformer based model that integrates biological pathway information for accurate identification of primary tumor sites for CUPs.[1] BPformer employs three core modules—pathway embedding, Transformer encoder, and classification head to processes transcriptomes, and leverage their complex relationships within biological pathways for the prediction of primary tumor sites in CUPs. Furthermore, BPformer offers additional insights into the biological mechanisms of tumor metastasis and supports the development of personalized treatment strategies. Additionally, an online platform is developed using the Django web framework. This publicly accessible platform allows users to upload gene expression data to perform BPformer analysis, and download the final prediction reports and visualizations plots for further analysis.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw data of CUP samples	Xie et al.[1]	https://ngdc.cncb.ac.cn/gsa-human
Preprocessed data of CUP samples	Xie et al.[1]	https://zenodo.org/records/11179119
Transcriptomics 1 (TCGA)	Cancer Genome Atlas Research Network[12]	https://portal.gdc.cancer.gov/; https://zenodo.org/records/11174764
Transcriptomics 2 (ICGC)	International Cancer Genome Consortium et al.[13]	https://dcc.icgc.org/; https://zenodo.org/records/11174764
Transcriptomics 3 (GEO)	Barrett T et al.[14]	https://www.ncbi.nlm.nih.gov/geo/; https://zenodo.org/records/11174764
Software and algorithms		
CPU - Intel(R) Xeon(R) CPU E5-2696 v.3 @ 2.30 GHz	-	-
GPU - 02:00.0 VGA compatible controller: NVIDIA Corporation GP104 [GeForce GTX 1080] (rev a1)	-	-
BPformer	Xie et al.[1]	https://github.com/xmuyulab/BPformer; Zenodo code: https://zenodo.org/records/11174895
Other		
Website for BPformer	Xie et al.[1]	https://cup.bpformer.com/index/

MATERIALS AND EQUIPMENT

The BPformer platform is a freely accessible web server available at https://cup.bpformer.com/index/, which is developed using the Django web framework. BPformer provides access to datasets, models, and codebases from previous CUP studies. The image illustrating cancer cell metastasis on the homepage is credited to https://immunooncologynews.com/2019/11/27/surface-oncology-asking-open-srf61-antibody-trial-advanced-solid-cancers/. Additionally, the BPformer network architecture is developed using PyTorch 2.0.1 and Python 3.8.13. The source code is available on https://github.com/xmuyulab/BPformer. The CUDA version is 11.8. The GLIBC version is 2.35 (Ubuntu GLIBC 2.35-0ubuntu3.9). The Ubuntu version is 22.04.4.

STEP-BY-STEP METHOD DETAILS

Collect and prepare data

Timing: 5 min

Raw RNA-seq FASTQ files are quality-trimmed, aligned to the reference genome, and quantified at the gene level. The resulting gene-by-sample expression matrix is $\log_2$-transformed and saved as. csv and/or .pkl for direct downstream analyses.

1. Process the raw sequencing data (in FASTQ format) using the preprocessing pipeline provided by the sequencing platform.
   a. Transform the processed data into a matrix format, with rows representing gene names and columns representing sample names.
   b. Save this data in .csv or .pkl formats.

   Note: For users interested in using publicly available datasets, this protocol outlines recommended procedures for processing these datasets.

For the RNA sequencing (RNA-seq) platform, raw sequencing data are trimmed using Trimmomatic,[15] reads are aligned to the reference genome using HISAT2.[16] Gene counts and RPKM values

are quantified using StringTie.[17] All expression data are log2-transformed prior to downstream analysis.

Model installation and usage

Timing: ~10 min

This section describes the complete setup and execution process for training and evaluating the BPformer model.

2. Installation
   a. Create a conda virtual environment and activate it.

```
>conda create -n bpformer python=3.8.13

>git clone https://github.com/xmuyulab/BPformer.git

>cd BPformer

>conda activate bpformer
```

   b. Install PyTorch and torchvision following the official instructions (https://pytorch.org/), e.g.

```
>conda install pytorch==2.4.1 torchvision==0.19.1 torchaudio==2.4.1 pytorch-cuda=11.8 -c pytorch -c nvidia
```

   c. Install a set of required python packages, including scikit_learn, einops, numpy, pandas and tqdm.

```
>pip install -r requirements.txt
```

Note: Different versions of CUDA correspond to different package versions, which can be found at https://pytorch.org/get-started/previous-versions/. Here, we also provide an integrated method for creating and configuring all conda environment.

```
>git clone https://github.com/xmuyulab/BPformer.git

>cd BPformer

>conda env create -f environment.yml
```

3. Preparation of input data.
   a. Download the RNA-seq data and put them under the path 'RNAseq/Raw'. We provide our experiment dataset on: https://zenodo.org/records/14863210. Here, we provide an example.

```
>cd RNAseq/Raw

>wget --content-disposition https://zenodo.org/records/14863210/files/best_weight.pth?download=1
```

   b. Transfer the input RNA-seq data into required format (genes are arranged in biological pathways). The attributes of the column in the input file include 'gene name' and 'label' of CUP.

```
>cd ../../

>python RNAseq/Raw_to_KEGG.py
```

Note: Please ensure a stable internet connection when downloading. The download speed will be faster when the network connection is stable and reliable. Different datasets may use different gene naming conventions. To facilitate subsequent data processing steps and identify common genes, it is necessary to convert all genes to gene symbols.

4. Model Training and Inference.
   a. Use our provided training datasets (data_train.pkl), which include TCGA and ICGC primary tumors (TCGA-RNA-p.pkl and ICGC-RNA-p.pkl).

```
>python train.py
```

   b. Evaluate BPformer on independent test datasets, including primary tumors from GEO (GEO-RNA-p.pkl), metastatic sites of metastatic tumors from GEO (GEO-RNA-m.pkl), and the primary sites of metastatic tumors from GEO, TCGA, and ICGC (GEO-RNA-p-m.pkl, TCGA-RNA-p-m.pkl, and ICGC-RNA-p-m.pkl). Our trained model file is saved in best_weight.pth, which can be downloaded in https://zenodo.org/records/14863210.

```
>python test.py
```

Usage of the website

Timing: 5–120 s

This section introduces the web-based BPformer interface for cancer type prediction and result visualization.

5. Go to the dataset download page by clicking on "DATASETS" (Figure 1) in the top navigation bar or "DATASETS" in "Our Services" (Figure 2).
6. Click the "RNA" button to choose the data type. Click the "⬇" icon in the Link to download the corresponding datasets (Figure 3). Click the "‹" button to go to the previous page, click the "›" button to go to the next page, or enter the destination page number in the input box and click the "GO" button to go directly (Figure 3).
7. Go to the Model page by clicking on "ANALYSIS" in the top navigation bar or "ANALYSIS" in "Our Services" or "Try Now" in "BPformer" (Figure 4).
8. Select "Use Example Data" or "Use Your Data" mode in "Input Mode" (Figure 5) as needed. Due to computer resource limitations, "Use Your Data" mode may run slower if the amount of data is large, you can click "download code" to deploy the codes to run (Figure 5).
9. If "Use Example Data" is selected, select Data Types (mRNA) in "Parameter Setting", select Data Sources if the data type is mRNA and then select Data Input (Figure 6). BPformer provides four examples, among which example 1-3 include single sample data and example 4 include multi samples. You can click "view all" to view the details of the samples, and then click "submit" button to submit.
10. If you choose "Use Your Data", you need to select the Data Types (mRNA) in "Parameter Setting", select Data Sources and then upload the Input Data (Data Input), you can click "view input examples" to see the input example. Click "upload file" to select the file you want to upload, and then click "submit" to submit (Figure 7).

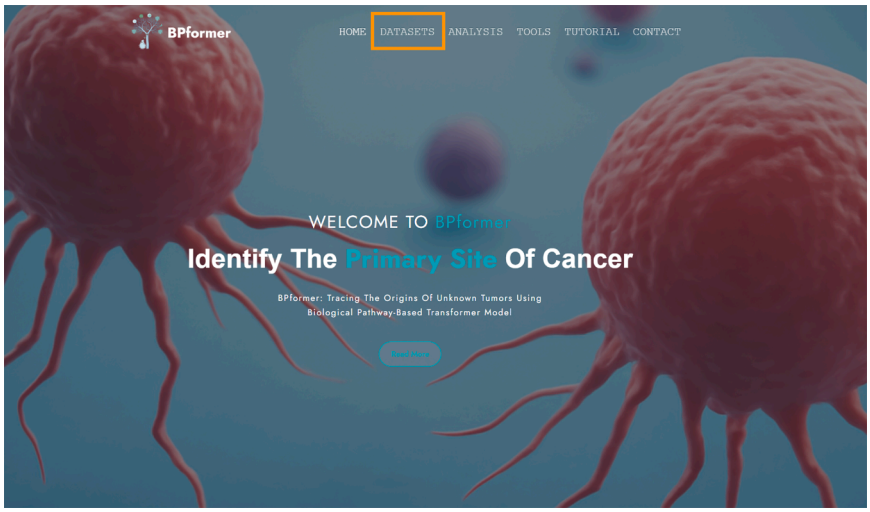


Figure 1. Main image of the BPformer webpage

11. After submission, the final prediction results of each sample will be displayed in "Diagnosis Result", but only the results of the first 12 samples will be displayed. You can click "⬇" next to "Diagnosis Result" to view the prediction results of all samples. The top 5 prediction results of each sample, including the type name (type) and the probability (score), will be displayed in the "Cancer Type Score". Similarly, only the top 5 prediction results of the first 12 samples will be displayed. You can click "⬇" next to the table to view the Top 5 prediction results for all samples. Click on the entry in the table to view the probability of 32 cancer types of the selected sample in the graphical box below. Click the "⬇" next to the bar graph to save the image (Figure 8).
12. Click on "TOOLS" in the top navigation bar or "TOOLS" in "Our Services" to go to the code download page. Select the code you want to download and click the "GitHub >" button to download it (Figure 9).

EXPECTED OUTCOMES

This protocol presents a deep learning tool that integrates a Transformer model with KEGG pathways to accurately identify the primary sites for CUPs, alongside a user-friendly web server.

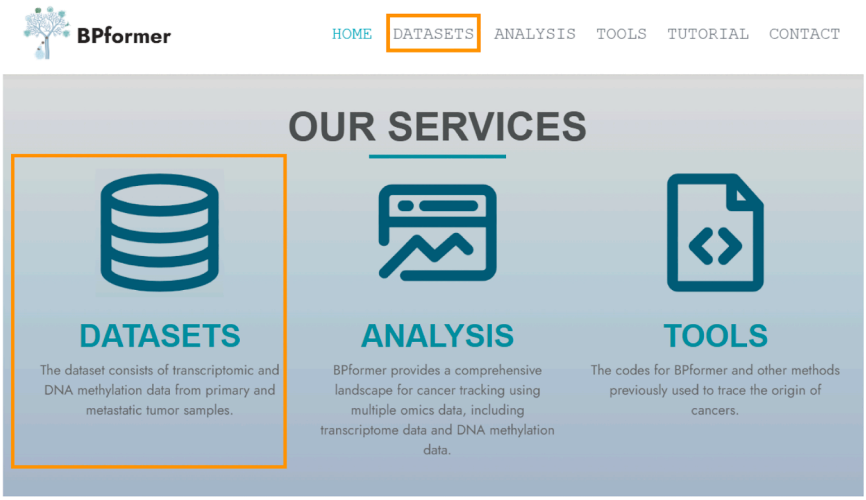


Figure 2. Function selection interface

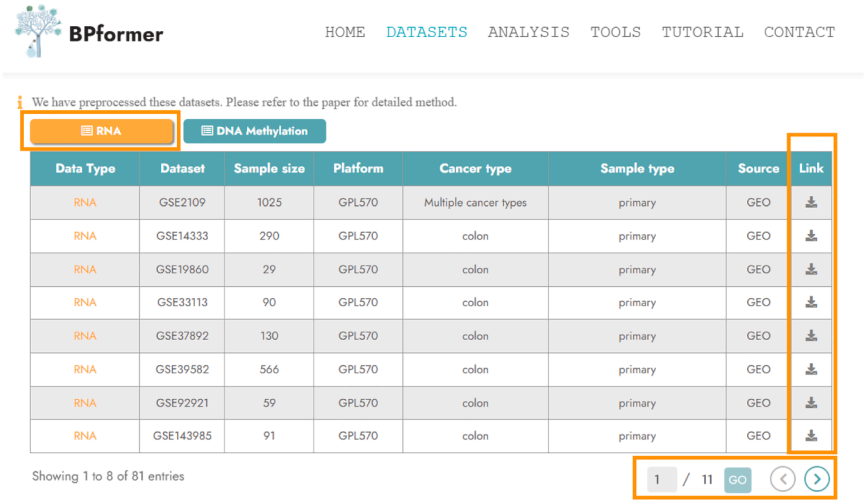


Figure 3. Choose data type, browse, and download

Preliminary studies have shown that the BPformer model outperforms other algorithms in terms of overall F1-score, precision, and recall, while also demonstrating strong cross-platform compatibility and robustness (Figure 10). The tool provides access to 178 datasets including 13,990 samples that were used in the study and can be downloaded directly from the portal. Users can access the web interface to make predictions on their own data. Additionally, users can download the codes and checkpoints for the BPformer method to trace origins for CUPs or use the website to run predictions directly. In summary, this protocol supports comprehensive research on CUP by providing both large datasets and models.

LIMITATIONS

There are several limitations to the BPformer platform

Limitations in data type: BPformer was evaluated exclusively on transcriptomes. However, its architecture is adaptable to other omics data types, such as DNA methylation and proteome. Incorporating these additional data types could further enhance the performance of the model.

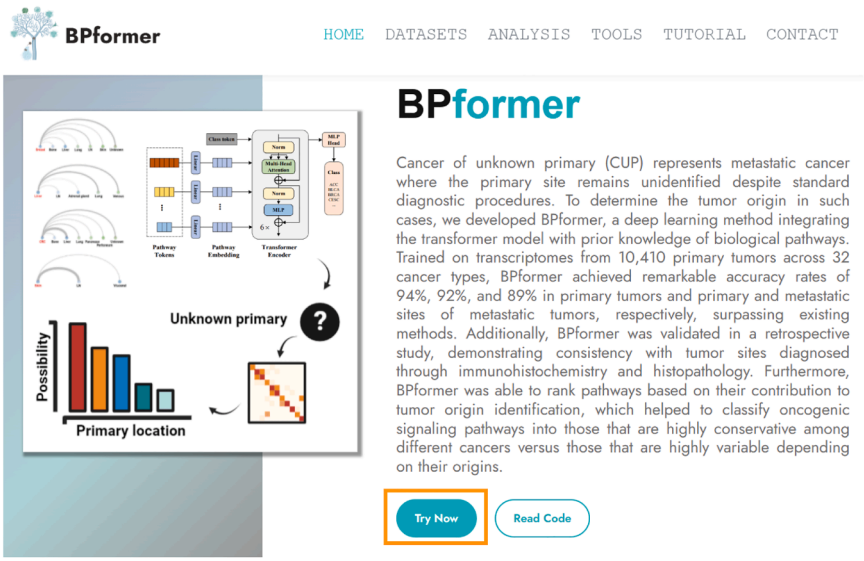


Figure 4. Model overview and description

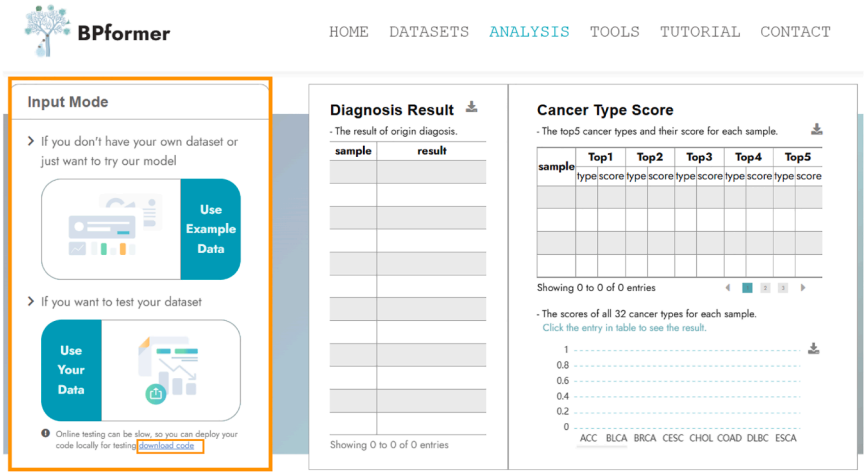


Figure 5. Select data input mode

Limitations in biological function: BPformer currently relies on pathways exclusively from the KEGG database. Integrating additional biological pathway data could enhance the model's performance and provide deeper insights into cancer-specific biological processes associated with various origins. This, in turn, could facilitate the identification of potential therapeutic targets.

Limitations in computational resources: A key limitation of BPformer is its high memory consumption. The design of pathway embeddings aims to mitigate memory overflow issues that arise when processing long input sequences in the transformer module. Specifically, to address the issue of long sequences, Eric Nguyen and colleagues developed HyenaDNA. HyenaDNA is a powerful and efficient solution that, for the first time in genomics, utilizes in-context learning based on implicit convolutions. It is capable of processing longer DNA sequences while preserving single nucleotide-level precision, thereby surpassing traditional Transformer models in a variety of genomics tasks.[18] Addressing this limitation could enhance the model's scalability and overall usability.

TROUBLESHOOTING

Problem 1

The RNA-seq data cannot be used due to formatting issues (related to Step 3).

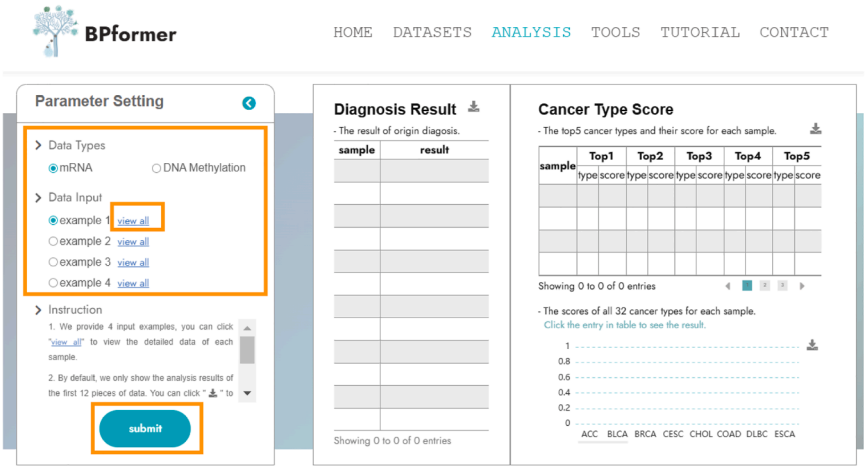


Figure 6. Demonstration with example data

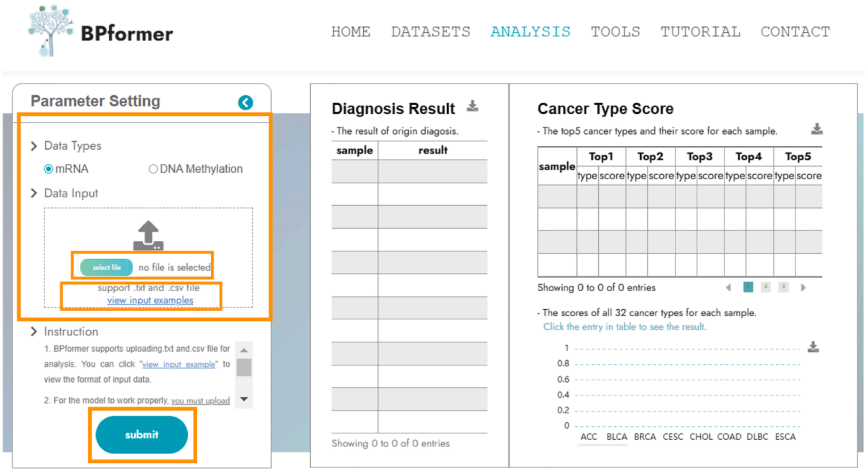


Figure 7. Demonstration with user-provided data

Potential solution

If issues arise with data input due to formatting discrepancies, they are likely caused by variations in gene nomenclature across different datasets. Therefore, it is essential to standardize the gene names in the input data. Gene names should be converted to Gene Symbol format, as demonstrated in the provided dataset. Additionally, gene names must be standardized according to the format used in the KEGG_Pathway_information.csv file.

Problem 2

The weights and test data cannot be found (related to Step 4).

Potential solution

The data and models have been uploaded to Zenodo; however, some users may encounter difficulties in accessing these files. To access the datasets and models stored on the cloud, users can directly download the weights and data through this link: https://zenodo.org/records/14863210.

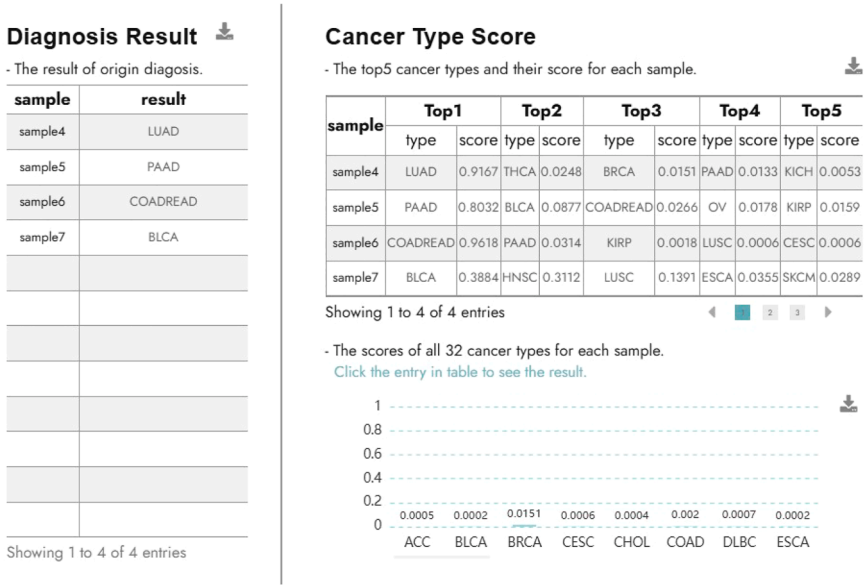


Figure 8. Presentation of BPformer output results

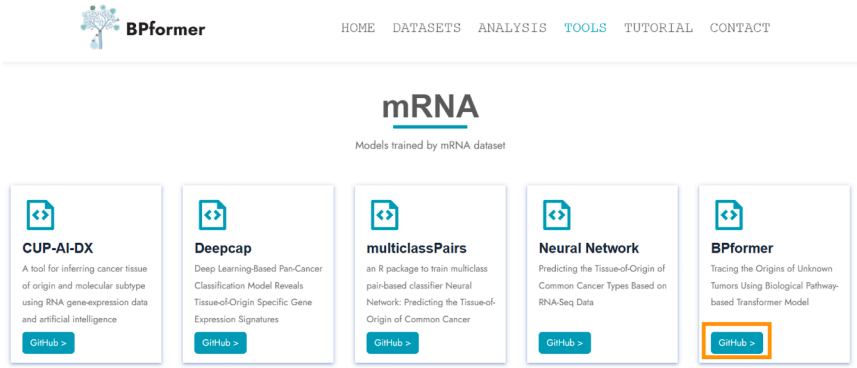


Figure 9. Web links for available CUPs prediction tools

Problem 3

BPformer model is unavailable on the web platform (related to Step 8, 9 and 10).

Potential solution

If the BPformer model is unavailable on the web platform, users should execute the model locally. The procedure involves downloading the pre-trained model weights from https://zenodo.org/records/14863210 and conducting testing on the dataset.

Problem 4

Unable to Upload Due to Excessive Sample Size (related to Step 10).

Potential solution

If the file upload is incomplete, resulting in a status of "In Progress" or "Upload of input dataset failed," the issue is likely due to the web platform being optimized for small sample sizes. For larger sample sizes, users are advised to refer to the solution outlined in Problem 3, which involves running the BPformer model code locally.

Problem 5

No data available for download (related to Step 11).

Potential solution

If no data is available for download when clicking "Diagnosis Result," "Cancer Type Score," or "The scores of all 32 cancer types for each sample," please ensure that the data submission step has been completed before attempting to download.

Problem 6

A CUDA error occurred when running BPformer (cuda error: invalid device ordinal).

Potential solution

First, check your CUDA version. If the CUDA version is 11.8, follow the steps above to install the dependencies. If not, see the official instructions (https://pytorch.org/get-started/previous-versions/) to install PyTorch and torchvision and update the corresponding dependency versions.

Problem 7

A GLIBC vision error occurred when running BPformer (version 'GLIBC_2.27' not found).

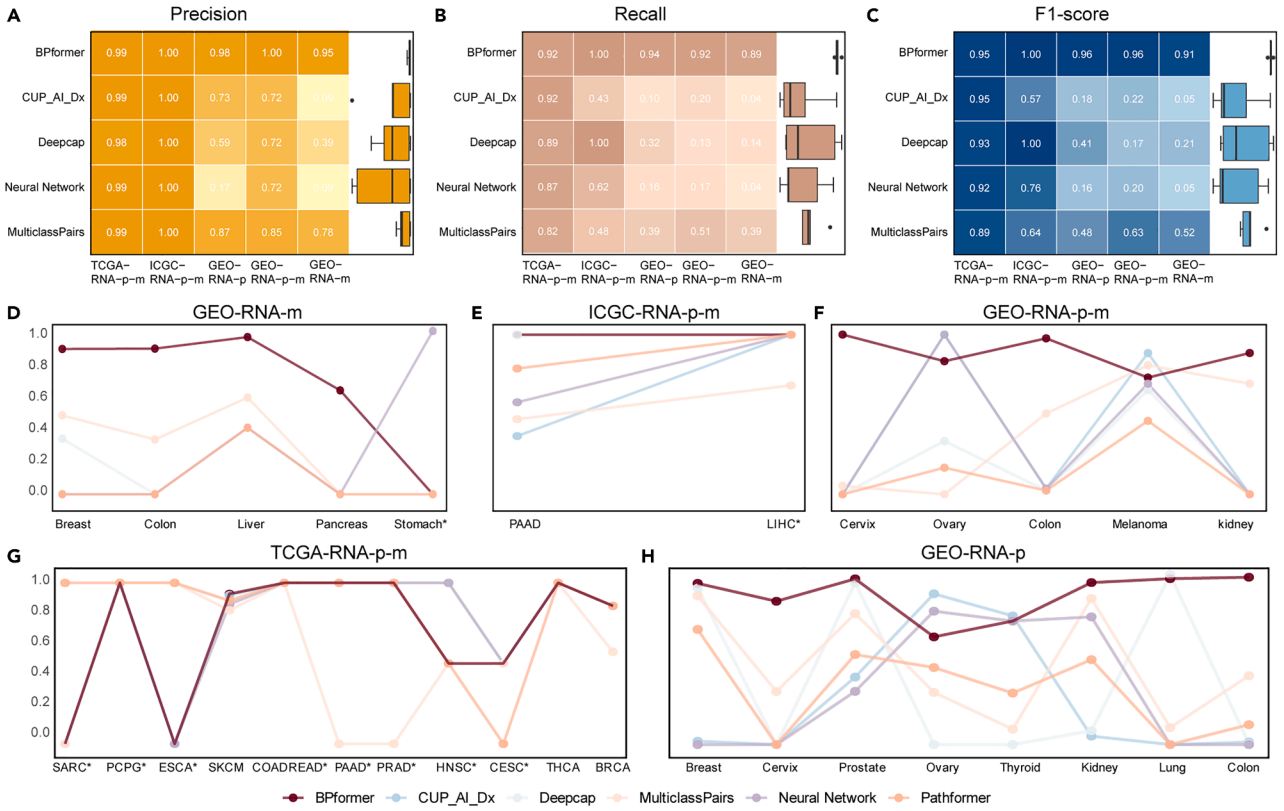


Figure 10. The performance of the BPformer model

(A–C) Heatmaps and boxplots illustrating comparisons of precision, recall, and F1-score across six distinct methods evaluated on five independent test datasets.

(D–H) Comparative analyses of recall for different methods across multiple cancer types within each independent test dataset. Cancer types with sample sizes fewer than five are denoted by an asterisk (*). Data in panels A-C and D-H were reused with permission from Xie et al.[1]

Potential solution

It is necessary to utilize the corresponding Linux systems and GLIBC libraries to ensure compatibility with the target application or dependencies. The GLIBC version is 2.35(Ubuntu GLIBC 2.35-0ubuntu3.9). The Ubuntu version is 22.04.4.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Rongshan Yu (rsyu@xmu.edu.cn).

Technical contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Rongshan Yu (rsyu@xmu.edu.cn).

Materials availability

There are no newly generated materials associated with this protocol.

Data and code availability

Raw data derived from CUP samples have been deposited at Genome Sequence Archive database: HRA005449. Processed data for CUP samples are available at https://zenodo.org/records/11179119. All datasets of primary tumors, primary or metastatic sites of metastatic tumors used in this study are available at https://zenodo.org/records/11174764. BPformer original code has been deposited at Zenodo and is publicly available as of the date of publication at https://zenodo.org/records/11174895. DOI is also listed in the key resources table. Additionally, source code is available

at https://github.com/xmuyulab/BPformer. Other comparative algorithms are listed in the key resources table. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank Peixuan Huang for her assistance in proofreading and refining the text of this paper. This work was supported by the National Key R&D Program of China (grant number 2024YFF1206801 and 2024YFF1206800); Natural Science Foundation of Xiamen, China (grant number 3502Z202471014); National Natural Science Foundation of China (grant number 82002529); and the Wang Deyao Outstanding Graduate Scholarship Program of Xiamen University (grant number 0660-T040001).

AUTHOR CONTRIBUTIONS

J.X., T.Q., R.Y., and M.T. conceived the study. J.X., T.Q., and Y.C. performed the data curation and analysis. T.Q. took charge of web content conceptualization and figure drafting. J.X. and T.Q. wrote the original draft. R.Y., M.T., and C.Y. revised the manuscript. All authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Xie, J., Chen, Y., Luo, S., Yang, W., Lin, Y., Wang, L., Ding, X., Tong, M., and Yu, R. (2024). Tracing unknown tumor origins with a biological-pathway-based transformer model. Cell Rep. Methods *4*, 100797. https://doi.org/10.1016/j.crmeth.2024.100797.

2. Guan, X. (2015). Cancer metastases: challenges and opportunities. Acta Pharm. Sin. B *5*, 402–418. https://doi.org/10.1016/j.apsb.2015.07.005.

3. Zhang, Z., Lu, Y., Vosoughi, S., Levy, J.J., Christensen, B.C., and Salas, L.A. (2023). HiTAIC: hierarchical tumor artificial intelligence classifier traces tissue of origin and tumor type in primary and metastasized tumors using DNA methylation. NAR Cancer *5*, zcad017. https://doi.org/10.1093/narcan/zcad017.

4. Lee, M.S., and Sanoff, H.K. (2020). Cancer of unknown primary. BMJ *371*, m4050. https://doi.org/10.1136/bmj.m4050.

5. Zhao, Y., Pan, Z., Namburi, S., Pattison, A., Posner, A., Balachander, S., Paisie, C.A., Reddi, H.V., Rueter, J., Gill, A.J., et al. (2020). CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine *61*, 103030. https://doi.org/10.1016/j.ebiom.2020.103030.

6. Qaseem, A., Usman, N., Jayaraj, J.S., Janapala, R.N., and Kashif, T. (2019). Cancer of Unknown Primary: A Review on Clinical Guidelines in the Development and Targeted Management of Patients with the Unknown Primary Site. Cureus *11*, e5552. https://doi.org/10.7759/cureus.5552.

7. Massard, C., Loriot, Y., and Fizazi, K. (2011). Carcinomas of an unknown primary origin–diagnosis and treatment. Nat. Rev. Clin. Oncol. *8*, 701–710. https://doi.org/10.1038/nrclinonc.2011.158.

8. Niazi, M.K.K., Parwani, A.V., and Gurcan, M.N. (2019). Digital pathology and artificial intelligence. Lancet Oncol. *20*, e253–e261. https://doi.org/10.1016/S1470-2045(19)30154-8.

9. Weiss, L.M., Chu, P., Schroeder, B.E., Singh, V., Zhang, Y., Erlander, M.G., and Schnabel, C.A. (2013). Blinded comparator study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors. J. Mol. Diagn. *15*, 263–269. https://doi.org/10.1016/j.jmoldx.2012.10.001.

10. Tothill, R.W., Shi, F., Paiman, L., Bedo, J., Kowalczyk, A., Mileshkin, L., Buela, E., Klupacs, R., Bowtell, D., and Byron, K. (2015). Development and validation of a gene expression tumour classifier for cancer of unknown primary. Pathology (Phila., Pa.) *47*, 7–12. https://doi.org/10.1097/PAT.0000000000000194.

11. Handorf, C.R., Kulkarni, A., Grenert, J.P., Weiss, L.M., Rogers, W.M., Kim, O.S., Monzon, F.A., Halks-Miller, M., Anderson, G.G., Walker, M.G., et al. (2013). A multicenter study directly comparing the diagnostic accuracy of gene expression profiling and immunohistochemistry for primary site identification in metastatic tumors. Am. J. Surg. Pathol. *37*, 1067–1075. https://doi.org/10.1097/PAS.0b013e31828309c4.

12. Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068. https://doi.org/10.1038/nature07385.

13. International Cancer Genome Consortium, Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer genome projects. Nature *464*, 993–998. https://doi.org/10.1038/nature08987.

14. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. *41*, D991–D995. https://doi.org/10.1093/nar/gks1193.

15. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

16. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

17. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295. https://doi.org/10.1038/nbt.3122.

18. Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., et al. (2023). HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.15794.