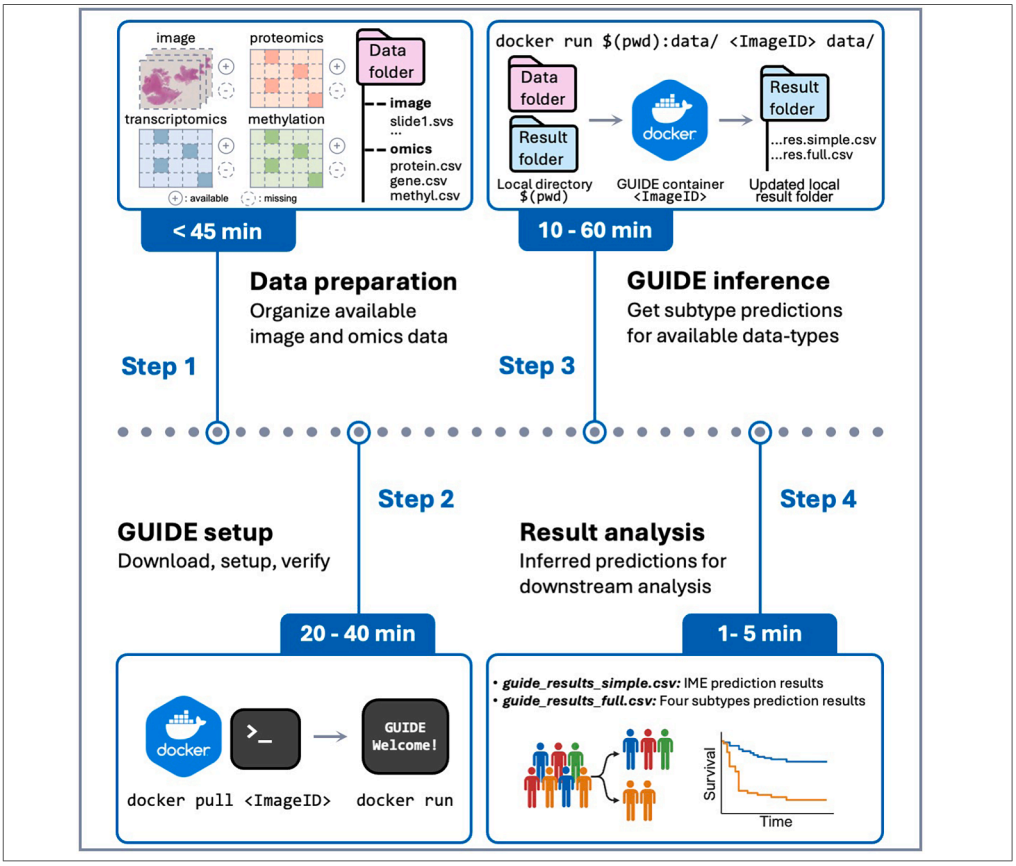
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for AI-powered classification of IDH-mutant astrocytoma using GUIDE



Accurate identification of prognostic cancer subtypes remains challenging using routine diagnostics alone. Here, we present a protocol to obtain isocitrate dehydrogenase (IDH)-mutant astrocytoma subtype predictions using the generic utility for immune/mesenchymal-enriched (IME) diagnostic estimation (GUIDE). We describe steps for preparation of image and omics inputs, setup of the GUIDE Docker container, execution of subtype inference, and retrieval of standardized output files. GUIDE integrates histopathology images and multi-omics data and enables robust and flexible subtype classification under variable data availability.

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

Jihong Tang, Yimeng Qiao, Yuyan Ruan, Jiguang Wang

jtangbd@connect.ust.hk (J.T.)
jgwang@ust.hk (J.W.)

Highlights

A protocol for classifying IDH-mutant astrocytomas using GUIDE

Instructions for preparing histopathology and multi-omics input data

Steps for running GUIDE subtype inference via a Docker container

Guidance on retrieving and interpreting binary and four-subtype outputs

Tang et al., STAR Protocols 7, 104512
June 19, 2026
https://doi.org/10.1016/j.xpro.2026.104512

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for AI-powered classification of IDH-mutant astrocytoma using GUIDE

Jihong Tang,[1,2,3,*] Yimeng Qiao,[1] Yuyan Ruan,[1] and Jiguang Wang[1,2,4,*]

[1]Division of Life Science, Department of Chemical and Biological Engineering, State Key Laboratory of Nervous System Disorders, The Hong Kong University of Science and Technology, Hong Kong SAR, China

[2]SIAT-HKUST Joint Laboratory of Cell Evolution and Digital Health, HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China

[3]Technical contact

[4]Lead contact

*Correspondence: jtangbd@connect.ust.hk (J.T.), jgwang@ust.hk (J.W.)
https://doi.org/10.1016/j.xpro.2026.104512

SUMMARY

Accurate identification of prognostic cancer subtypes remains challenging using routine diagnostics alone. Here, we present a protocol to obtain isocitrate dehydrogenase (IDH)-mutant astrocytoma subtype predictions using the generic utility for immune/mesenchymal-enriched (IME) diagnostic estimation (GUIDE). We describe steps for preparation of image and omics inputs, setup of the GUIDE Docker container, execution of subtype inference, and retrieval of standardized output files. GUIDE integrates histopathology images and multi-omics data and enables robust and flexible subtype classification under variable data availability.

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

BEFORE YOU BEGIN

Background on prognostic classification of IDH-mutant astrocytoma from multi-modal data

Prognostic classification of IDH-mutant astrocytomas can be formulated as a subtype-specific prediction problem based on molecular and histopathological features. Previous studies have demonstrated that immune/mesenchymal-enriched (IME) tumor programs underlie substantial clinical heterogeneity within IDH-mutant astrocytomas and are associated with adverse patient outcomes.[1] While histopathological diagnosis is mature and central to routine clinical practice, recent work has shown that integrating molecular features derived from transcriptomic, proteomic, and epigenetic profiling can substantially refine prognostic stratification.

Here, we introduce GUIDE (Generic Utility for IME Diagnostic Estimation), an AI-powered multimodal framework designed to identify prognostically relevant subtypes of IDH-mutant astrocytoma by integrating histopathology images with multi-omics data, including proteomics, transcriptomics, and DNA methylation profiles. The protocol described here details the steps for running a pretrained and containerized version of GUIDE to obtain subtype predictions under variable data availability. Additional methodological details and model development are described in the companion research article[1] and supporting documentation (Document S1).

Innovation

This protocol introduces GUIDE, a reproducible workflow for prognostic subtype classification of IDH-mutant astrocytomas using multi-modal data. GUIDE integrates histopathology images with proteomic, transcriptomic, and DNA methylation inputs within a single computational framework. The key advancement of GUIDE is its modular ensemble design, in which each available data

modality is analyzed independently and combined through a soft-voting strategy. This approach enables robust subtype prediction under variable data availability, addressing common limitations of single-modality classifiers in clinical and translational datasets.

The protocol further innovates by emphasizing usability and reproducibility. All analysis steps, trained models, and software dependencies are encapsulated within a Docker container, eliminating the need for manual installation or environment configuration. As a result, GUIDE can be executed consistently across operating systems and computational infrastructures using a small number of standardized commands.

By combining mature histopathological assessment with complementary molecular information in a containerized workflow, this protocol provides a practical extension of existing diagnostic and analytical methods. GUIDE enables reproducible, multi-modal prognostic classification of IDH-mutant astrocytomas and offers a generalizable framework that can be adapted to other tumor subtypes or integrated with additional data modalities in future applications.

Institutional permissions

To train the GUIDE GPU version provided in this protocol, we used the Chinese Glioma Genome Atlas (CGGA) tumor samples with informed consent under their local Institutional Review Boards.

System requirements

CPU mode:

Local memory: a minimum of 16 GB recommended.

Local storage: a minimum of 12.8 GB required.

GPU mode:

Minimum NVIDIA driver version: >=450.

Local memory: a minimum of 16 GB recommended.

Local storage: a minimum of 12.8 GB required.

Use CPU mode for making predictions with only molecular omics data; GPU mode for making predictions with imaging data.

Download and set up the GUIDE Docker container

Timing: <30 min

1. Download and launch Docker[2] client locally.

   Note: The steps presented here are tested with Docker version 20.10.17, build a89b842.

   a. Download Docker Desktop with Graphical User Interface (GUI) following the instructions on https://www.docker.com/products/docker-desktop/ or download Docker Engine without GUI following the instructions on https://docs.docker.com/engine/install/.
   b. For Docker Desktop users, launch Docker Desktop from the Applications folder (macOS) or Start Menu (Windows).
      i. Accept any required permissions when prompted.

A

```
jtangbd@dy024-181 ~ docker pull wanglabhkust/guide
Using default tag: latest
latest: Pulling from wanglabhkust/guide
96d54c3075c9: Pull complete
1d8f82780678: Pull complete
fb7423b4aec8: Pull complete
f8ae240d6263: Pull complete
81bb5702a96f: Pull complete
eaab28ffdfa4: Downloading [==============================>                    ]  852.1MB/1.303GB
7795b051ada1: Download complete
ae6c0f99e656: Download complete
4c93aa93344d: Download complete
e8d0f2816010: Downloading [========================================>          ]  755.4MB/862.3MB
429adc8a7ca4: Download complete
8216fe19863b: Download complete
60d482d14bc3: Downloading [===>                                               ]  27.83MB/404.1MB
d77d86e9eebb: Waiting
7c4ed5ea8688: Waiting
bc8b3340c6d4: Waiting
4f4fb700ef54: Waiting
```

B

```
jtangbd@dy024-181 ~ docker pull wanglabhkust/guide
Using default tag: latest
latest: Pulling from wanglabhkust/guide
96d54c3075c9: Pull complete
1d8f82780678: Pull complete
fb7423b4aec8: Pull complete
f8ae240d6263: Pull complete
81bb5702a96f: Pull complete
eaab28ffdfa4: Pull complete
7795b051ada1: Pull complete
ae6c0f99e656: Pull complete
4c93aa93344d: Pull complete
e8d0f2816010: Pull complete
429adc8a7ca4: Pull complete
8216fe19863b: Pull complete
60d482d14bc3: Pull complete
d77d86e9eebb: Pull complete
7c4ed5ea8688: Pull complete
bab4b55fdf74: Pull complete
4f4fb700ef54: Pull complete
Digest: sha256:a910a4e2944a0fd3045666ce11d12648149de80d67b7ef0d6ebe02c10c7d2c50
Status: Downloaded newer image for wanglabhkust/guide:latest
docker.io/wanglabhkust/guide:latest
```

Figure 1. Example Docker logs during GUIDE image download

(A) Representative console output showing the progress of downloading the GUIDE Docker image using the docker pull command, including layer-wise download status and progress indicators.

(B) Console output indicating successful completion of the image download, with all layers pulled and the final image ready for use.

ii. Wait until Docker Desktop completes initialization and displays a status indicating that Docker is running.

Note: For example, a green indicator or "Docker is running" message in the menu bar or system tray.

Note: Do not proceed until Docker is fully started.

c. For Docker Engine user, open a terminal and start the Docker daemon using the command appropriate for your operating system.

i. On most Linux systems, start the service with:

```
> sudo systemctl start docker
```

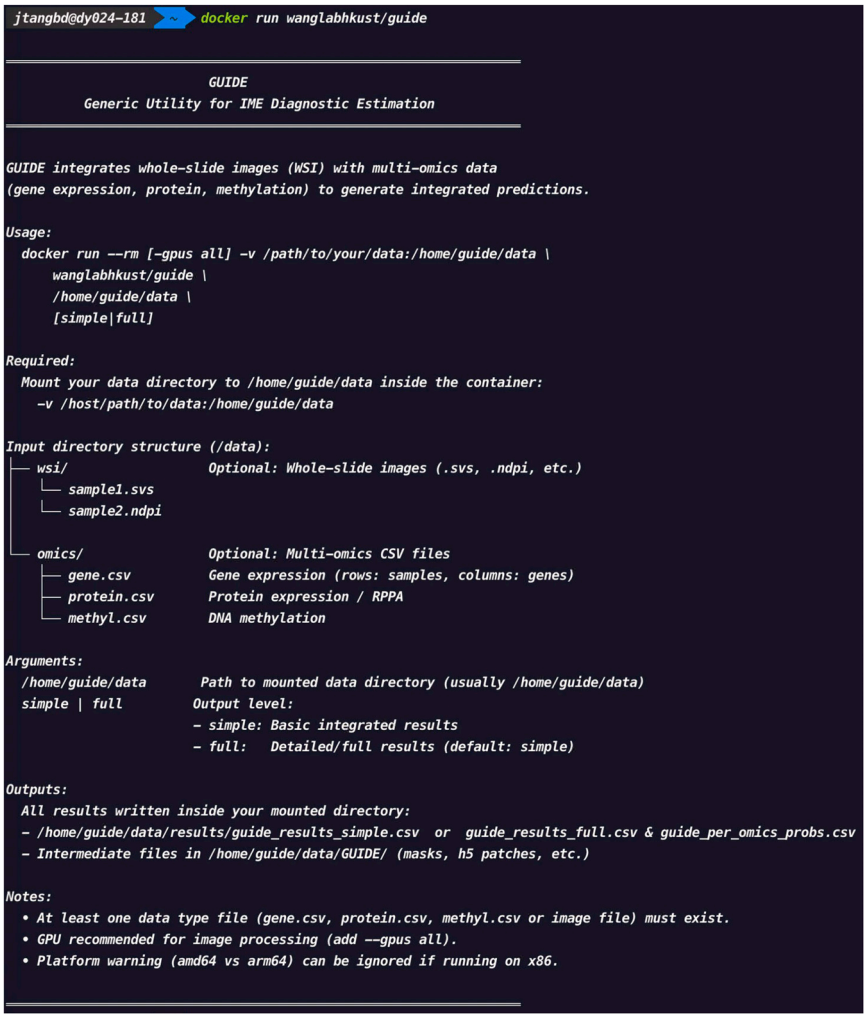


Figure 2. Example console output after launching the GUIDE Docker container

Representative terminal output displayed when running the GUIDE Docker container, showing the introductory message, usage instructions, required input directory structure, supported data modalities, available output modes (simple and full), and expected output file locations. This screen provides users with an overview of required inputs and command-line arguments before executing GUIDE inference.

ii. Verify that Docker daemon is running by executing:

```
> docker info
```

Note: Successful output confirms that Docker Engine is active and ready for use.

2. Obtain the GUIDE Docker image.

Note: The GUIDE image is 12.8 GB. This is the most time-consuming setup step depending on the user's available internet bandwidth.

a. From the terminal or command-line, navigate to the project root directory.
b. Pull the GUIDE Docker container using the command:

```
> docker pull wanglabhkust/guide
```

Note: Figure 1 shows example logs when the GUIDE image is downloading and when the download is finished.

Alternatives: For users without direct internet access, the GUIDE Docker image can be downloaded on a machine with internet connectivity, transferred to the target server, and installed manually. Load the Docker image from the provided archive file and create the container using the following command:

```
> docker load --input guide_image.tar.gz
```

3. Confirm that the Docker image has been loaded successfully.
   a. From the terminal or command-line, run the following command:
      `> docker images`
   b. Look for an entry under the column 'REPOSITORY' that says.
      `'wanglabhkust/guide'`

Note: The **<IMAGE ID>** for each entry will differ for each downloaded instance. Use this Image ID in the next step.

4. Launch a GUIDE Docker Container to understand the necessary inputs.
   a. From the terminal or command-line, run the following command:
      `> docker run -t <IMAGE ID>`

Alternatives: Run the following command:

```
> docker run wanglabhkust/guide
```

Note: A screenshot of the expected output is shown in Figure 2.

Prepare input files

Timing: <45 min

5. Organize input files in the required directory structure.
   a. Prepare all available input data before running GUIDE.
   b. Organize all input data in a local directory (/path/to/local/data) following the structure below:

```
/path/to/local/data/
├── wsi/
│   ├── slide1.svs
│   ├── slide2.ndpi
│   └── ...
└── omics/
├── gene.csv
├── protein.csv
└── methyl.csv
```

c. Store histopathology whole-slide images in the wsi/ subdirectory.
d. Store molecular data files in the omics/ subdirectory using the fixed file names gene.csv, protein.csv, and methyl.csv.

Note: All the four data modalities are optional; however, at least one modality (image, transcriptome, proteome or DNA methylome) must be provided for GUIDE inference.

Note: Currently, GUIDE supports whole-slide image files in .svs and .ndpi formats only.

6. Format omics input files correctly.
   a. Format each omics file such that each row corresponds to one sample, and each column corresponds to one molecular feature.
   b. Use the first column to specify sample IDs.

   Note: Sample IDs must be consistent across all provided omics files and, when applicable, should correspond to the associated whole-slide images.

   c. Use the first row to specify feature names (e.g., gene, protein, or CpG probe identifiers).

   Note: Feature lists for each omics are also included in Data S1.

7. Ensure that omics inputs are preprocessed prior to GUIDE execution.
   a. Ensure that gene expression, proteomics, and DNA methylation input files are fully preprocessed before running GUIDE.

   Note: GUIDE does not perform raw data normalization. Detailed preprocessing requirements are described in the data pre-processing section.

8. Verify input file formatting before proceeding.
   a. Confirm that all omics files are comma-separated (.csv) and contain no duplicated sample IDs.
   b. Confirm that missing values are either left blank or encoded as NA.
   c. Confirm that file and directory names exactly match the required structure.

   Note: Example input files are also provided in Data S2.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
guide_image.tar.gz (Docker image)	This article	https://doi.org/10.5281/zenodo.18217774
Molecular omics signatures	This article	Data S1
Example input data	This article	Data S2
Software and algorithms		
Docker (version 20.10.17)	N/A	https://www.docker.com
Other		
CPU Hardware for inference	This article	MacOS v15.6.1, 3.6 GHz 10-core intel core i9 processor, 64 GB memory
GPU Hardware for inference	This article	Ubuntu 20.04.6 LTS, Nvidia GeForce GTX 3090, NVIDIA Driver 470.256.02

STEP-BY-STEP METHOD DETAILS

Follow these steps to use the accompanying GUIDE tool to obtain prognostic subtype predictions for IDH-mutant astrocytoma samples. Before beginning, follow the instructions in the before you begin section to ensure that input data are correctly formatted and that the required software environment has been set up.

⚠ CRITICAL: GUIDE generates subtype predictions using the available input modalities, including histopathology images, gene expression, protein abundance, and DNA methylation data. At least one data modality must be provided for each sample. Prediction confidence and subtype resolution may vary depending on the combination of data types supplied.

Data preprocessing

Timing: <30 min

This major step ensures that all histopathology image and molecular omics inputs are correctly prepared, preprocessed, and internally consistent before running GUIDE. By confirming image readiness, validating preprocessing of gene expression, proteomics, and DNA methylation data, and verifying sample identifier consistency across modalities, this step establishes the required input state for reliable GUIDE inference. Completing these checks reduces execution errors and facilitates reproducible subtype prediction and downstream troubleshooting.

1. Confirm readiness of histopathology image input files. troubleshooting 4.
   a. Ensure that whole-slide images have been prepared and organized as described in the before you begin section.
   b. Confirm that supported whole-slide image formats (e.g., .svs or.ndpi) are used.
   c. Confirm that whole-slide images are scanned at 20× magnification as the preferred setting.

   Note: 40× magnification is also supported by the current version of GUIDE.

   d. Ensure that image filenames can be unambiguously associated with sample identifiers used in the omics input files, if applicable.

   Note: An image file named P_001.svs should correspond to the sample identifier P_001 used in the omics input tables.

   Note: GUIDE does not perform image format conversion or quality control. Image data are used as provided for downstream inference.

2. Confirm preprocessing and formatting of molecular omics input files. troubleshooting 1, troubleshooting 2.
   a. Ensure molecular input files (gene.csv, protein.csv, and/or methyl.csv) have been prepared according to the specifications described in the before you begin section.

   Note: Each file should contain unique sample identifiers in the first column and molecular feature names in the first row.

   b. Ensure all molecular input files have been correctly preprocessed.
      i. For gene expression data (gene.csv), provide values that have been log2 transformed, quantile normalized, and scaled according to the user's preprocessing pipeline.

ii. For proteomics data (protein.csv), provide protein abundance values that have been log2 transformed, quantile normalized, and scaled as continuous numeric measurements.
iii. For DNA methylation data (methyl.csv), provide normalized beta values as continuous numeric measurements.

Note: Gene expression and proteomics data are expected to be fully preprocessed prior to running GUIDE. No additional normalization or feature scaling is performed during protocol execution.

3. Validate consistency across input data modalities.
   a. Confirm that sample identifiers are consistent across all provided omics files.
   b. Confirm that missing values are encoded as NA or left blank where applicable.
   c. Confirm that duplicated sample identifiers are not present.
   d. Confirm that both image and omics data are organized as described in the section before you begin under the local directory:/path/to/local/data. This local directory will be mounted to run GUIDE.

⚠ CRITICAL: GUIDE assumes that gene expression and protein abundance inputs have been fully preprocessed prior to execution. Raw counts or unnormalized abundance values are not supported and may lead to unreliable subtype predictions.

Run GUIDE

Timing: 10–60 min

Note: Timing assumes a minimum of 16 GB RAM. GPU acceleration is required when histopathology image data are provided.

This major step executes GUIDE to generate prognostic subtype predictions for IDH-mutant astrocytoma samples. GUIDE can be run in CPU mode or GPU mode, depending on the available input data: GPU resources are required when histopathology whole-slide images are included, whereas CPU-only execution is sufficient when only molecular omics data are provided. GUIDE also supports two output modes: a simple mode that reports binary IME versus non-IME classification, and a full mode that reports probabilistic predictions for all four molecular subtypes.

4. Confirm readiness to run GUIDE. troubleshooting 8.
   a. Before proceeding, ensure that Docker is installed and the GUIDE Docker image (wanglabhkust/guide) has been downloaded as described in the Download and set up the GUIDE Docker container subsection under the before you begin section.
5. Run GUIDE in CPU-only mode (omics-only input).
   a. When only molecular omics data are provided, run GUIDE with the active docker container and `/path/to/local/data:`

```
docker run -v /path/to/local/data:/home/guide/data \
wanglabhkust/guide \
/home/guide/data \
[simple|full]
```

   b. Command-line arguments and recommended settings:
      i. `docker run:` starts a new container instance from the GUIDE image.

A

```
jtangbd@dy024-181 ~/Downloads/guide_test docker run \
          -v /Users/jtangbd/Downloads/guide_test:/home/guide/data \
          wanglabhkust/guide    \
          /home/guide/data \
          full
Data path: /home/guide/data
Output level: full
=== Device Check ===
GPU not available, using CPU
Selected device: cpu
=== Not Using Images ===
=== Starting Multi-Omics Integration ===

=== GUIDE Full Mode Final Results ===
          final_AFM  final_PPR  ...               used_omics  predicted_cluster
CGGA_2103   0.231979   0.621292  ...  gene,methylation,protein                PPR
CGGA_P109   0.289683   0.559896  ...             gene,protein                PPR
CGGA_P137   0.754459   0.180695  ...                     gene                AFM
CGGA_P151   0.409811   0.266529  ...                     gene                NEU
CGGA_P153   0.441429   0.229939  ...                     gene                IME
CGGA_P83    0.582260   0.417740  ...                  protein                AFM
CGGA_P84    0.367638   0.311500  ...      methylation,protein                NEU
CGGA_P87    0.221481   0.600371  ...      methylation,protein                PPR

[8 rows x 6 columns]

Main results saved to: /home/guide/data/results/guide_results_full.csv
Per-omics raw probabilities saved to: /home/guide/data/results/guide_results_full_omics_prob.csv
=== Complete Multi-Omics Integration ===
=== Results saved to results/guide_results_full.csv and results/guide_results_full_omics_prob.csv (detailed probabilities)===
```

B

```
(base) yimeng@DigMed21-Image2:~$ docker run --gpus all \
>           --shm-size 12G \
>           -v /mnt/disk/yimeng/projects/glioma/guide/data:/home/guide/data \
>           guide:latest \
>           /home/guide/data \
>           simple
Data path: /home/guide/data
Output level: simple
=== Device Check ===
GPU detected and available
Selected device: cuda
=== Using Images ===
Running in GPU mode with CUDA_VISIBLE_DEVICES=0
WSI files path: /home/guide/data/wsi
=== Starting Image Preprocessing ===
Found 2 files
Number of processes: 2
Number of training images: 2
Task #1: Process slide 0
Task #2: Process slide 1
Opening Slide #0: /home/guide/data/wsi/CGGA_2103.svs
Opening Slide #1: /home/guide/data/wsi/CGGA_P87.svs
=== Starting Image Prediction ===
2
Downloading: "https://download.pytorch.org/models/resnet18-f37072fd.pth" to /root/.cache/torch/hub/checkpoints/resnet18-f37072fd.pth
100%|██████████| 44.7M/44.7M [00:00<00:00, 113MB/s]
  0%|          | 0/2 [00:00<?, ?it/s]CGGA_2103.h5 already exists
CGGA_P87.h5 already exists
100%|██████████| 2/2 [00:00<00:00, 3344.74it/s]
=== Complete Image Prediction ===
=== Starting Multi-Omics Integration ===

=== GUIDE Simple Mode Final Results ===
          P_IDHm_IME                      used_omics  IME_positive
CGGA_2103    0.303355  gene,image,methylation,protein         False
CGGA_P109    0.440104             gene,protein         False
CGGA_P137    0.245541                     gene         False
CGGA_P151    0.354958                     gene         False
CGGA_P153    0.558571                     gene          True
CGGA_P83     0.213925                  protein         False
CGGA_P84     0.245784      methylation,protein         False
CGGA_P87     0.376966    image,methylation,protein         False

Results saved to: /home/guide/data/results/guide_results_simple.csv
=== Complete Multi-Omics Integration ===
=== Results saved to results/guide_results_simple.csv ===
```

Figure 3. Example console logs from successful GUIDE runs using different execution modes

(A) Representative terminal output from a successful GUIDE run executed in CPU-only mode using molecular omics data.

(B) Representative terminal output from a successful GUIDE run executed in GPU-enabled mode, required when histopathology whole-slide images are included.

ii. `-v /path/to/local/data:/home/guide/data`: mounts the user's local input directory (`/path/to/local/data`) into the container at `/home/guide/data`.

Note: /path/to/local/data should be replaced with the absolute path to the prepared input folder on the local machine. The container path (/home/guide/data) should remain unchanged.

iii. `wanglabhkust/guide`: the GUIDE Docker image name.

iv. `home/guide/data`: the input directory inside the container. GUIDE reads the required subfolders from this location.

Note: Keep this argument unchanged unless the container mount point is modified.

v. `[simple|full]`: selects the output mode. simple returns binary IME vs non-IME classification, whereas full returns four-subtype probability score outputs (AFM, PPR, IME, NEU) and modality-specific probability scores.

Note: CPU-only execution is not allowed when whole-slide image data are provided.

Note: The computational workflow underlying simple and full modes is identical. The difference lies only in the scope of output reporting. Runtime differences between the two modes are minimal and generally negligible relative to overall execution time.

6. Run GUIDE in GPU mode (required for image data). troubleshooting 3, troubleshooting 4, troubleshooting 7.
   a. When histopathology whole-slide images are included, run GUIDE with GPU support enabled and `/path/to/local/data`:

```
docker run --gpus all \
    --shm-size 12G \
    -v /path/to/local/data:/home/guide/data \
    wanglabhkust/guide \
    /home/guide/data \
    [simple|full]
```

   b. Additional command-line arguments compared to step 5:
      i. `--gpus all`: exposes all available GPU devices to the container. Required for image-based inference.
      ii. `--shm-size 12G`: increases shared memory allocation to improve stability when processing large whole-slide images.

Note: The value may be adjusted depending on available system memory.

Note: GPU acceleration substantially reduces runtime for image-based inference.

Note: If GPU mode is invoked while only molecular omics data are provided, GUIDE will execute normally using CPU resources. In this case, GPU acceleration is not utilized, and runtime is equivalent to CPU-only mode. GPU resources are only required when histopathology image inputs are included.

7. Run GUIDE using a specified GPU device (optional). troubleshooting 3, troubleshooting 7.
   a. To restrict GUIDE execution to a specific GPU (e.g., CUDA device 1), run GUIDE with CUDA_VISIBLE_DEVICES and `/path/to/local/data`:

```
docker run --gpus all \
    -e CUDA_VISIBLE_DEVICES=1 \
```

```
    --shm-size 12G \
    -v /path/to/local/data:/home/guide/data \
    wanglabhkust/guide \
    /home/guide/data \
    [simple|full]
```

b. Additional command-line arguments compared to step 6:
   i. `-e CUDA_VISIBLE_DEVICES=1`: restricts execution to a specified GPU device (e.g., device 1).

Note: Replace 1 with the desired CUDA device index.

8. Select output mode and retrieve results. troubleshooting 5, troubleshooting 8.
GUIDE generates results in one of two output modes, specified by the final argument in the command above.
   a. Simple mode [simple].
      Produces a binary classification indicating whether each sample is classified as IME or non-IME. Results are written to:
      `/path/to/local/data/results/guide_results_simple.csv`
   b. Full mode [full].
      Produces probabilistic predictions for all four molecular subtypes (AFM, PPR, IME, and NEU). Results are written to:
      `/path/to/local/data/results/guide_results_full.csv`
      `/path/to/local/data/results/guide_per_omics_probs.csv`

Note: Select the argument simple or full in the commands above to generate binary IME classification results or full four-subtype probability score outputs, respectively. Figure 3 shows example logs from successful GUIDE runs.

Pause point: Runtime depends on the number of samples, inclusion of image data, and available hardware. Omics-only analyses typically complete within minutes on CPU, whereas image-based analyses benefit substantially from GPU acceleration.

EXPECTED OUTCOMES

Upon successful execution of the protocol, GUIDE generates structured output files summarizing prognostic subtype predictions for each input sample. These files are written to a local results directory within the mounted data path and are produced consistently across CPU-only and GPU-enabled executions. The specific outputs depend on whether GUIDE is run in simple or full mode, but in all cases the protocol yields standardized, machine-readable tables suitable for downstream statistical analysis, cohort stratification, and visualization.

When GUIDE is run in simple mode, the protocol produces a single output file, `guide_results_simple.csv`, which reports binary immune/mesenchymal-enriched (IME) classification results for each sample. Each row corresponds to one sample, identified by the row name, which matches the sample identifiers used in the input data. The column P_IDHm_IME reports a continuous probability score between 0 and 1, representing the predicted likelihood that the sample belongs to the IME subtype. The column IME_positive provides a binary classification result derived from this probability score using a predefined decision threshold of 0.5, indicating whether the sample is classified as IME-positive or non-IME. The column used_omics records the data modalities that were available

A

	P_IDHm_IME	used_omics	IME_positive
CGGA_2103	0.363510415	gene,protein	FALSE
CGGA_P109	0.447755663	gene,protein	FALSE
CGGA_P137	0.246126616	gene	FALSE
CGGA_P151	0.409867503	gene	FALSE
CGGA_P153	0.576733323	gene	TRUE
CGGA_P83	0.213924745	protein	FALSE
CGGA_P84	0.099673602	protein	FALSE
CGGA_P87	0.259010519	protein	FALSE

B

	final_AFM	final_PPR	final_IME	final_NEU	used_omics	predicted_cluster
CGGA_2103	0.160429461	0.636489585	0.363510415	0.105567397	gene,protein	PPR
CGGA_P109	0.257852992	0.552244337	0.447755663	0.09213952	gene,protein	PPR
CGGA_P137	0.741979097	0.192941328	0.246126616	0.258020903	gene	AFM
CGGA_P151	0.455750811	0.324423619	0.409867503	0.544249189	gene	NEU
CGGA_P153	0.423266677	0.216169237	0.576733323	0.33673915	gene	IME
CGGA_P83	0.582259895	0.417740105	0.213924745	0.203333829	protein	AFM
CGGA_P84	0.320459087	0.240992921	0.099673602	0.679540913	protein	NEU
CGGA_P87	0.120513057	0.62958018	0.259010519	0.37041982	protein	PPR

C

	protein_AFM	protein_PPR	protein_IME	protein_NEU	gene_AFM	gene_PPR	gene_IME	gene_NEU	used_omics
CGGA_2103	0.192551274	0.589318342	0.410681658	0.0764697	0.096185836	0.730832072	0.269167928	0.163762792	gene,protein
CGGA_P109	0.262199573	0.519093043	0.480906957	0.074348693	0.249159829	0.618546925	0.381453075	0.127721175	gene,protein
CGGA_P137					0.741979097	0.192941328	0.246126616	0.258020903	gene
CGGA_P151					0.455750811	0.324423619	0.409867503	0.544249189	gene
CGGA_P153					0.423266677	0.216169237	0.576733323	0.33673915	gene
CGGA_P83	0.582259895	0.417740105	0.213924745	0.203333829					protein
CGGA_P84	0.320459087	0.240992921	0.099673602	0.679540913					protein
CGGA_P87	0.120513057	0.62958018	0.259010519	0.37041982					protein

Figure 4. Example GUIDE output files generated in simple and full modes

(A) Representative output from GUIDE executed in simple mode, showing the file guide_results_simple.csv. Each row corresponds to one sample, and columns report the predicted probability score of the immune/mesenchymal-enriched (IME) subtype (P_IDHm_IME), the resulting binary IME classification (IME_positive), and the data modalities used for prediction (used_omics).

(B) Representative output from GUIDE executed in full mode, showing the file guide_results_full.csv. For each sample, subtype-specific integrated sbility scores are reported for the four molecular subtypes (AFM, PPR, IME, and NEU), along with the final predicted subtype (predicted_cluster) and the contributing data modalities (used_omics).

(C) Representative output from the modality-specific probability score file guide_per_omics_probs.csv, generated in full mode. Subtype probability score estimates are reported separately for each available data modality using standardized column prefixes (e.g., gene_, protein_). These modality-specific probability scores support interpretability and quality control and are combined internally by GUIDE to generate the integrated predictions shown in panel B.

and used for prediction, allowing users to assess how data availability influenced the classification outcome. This output mode is intended for analyses requiring a simplified IME versus non-IME stratification. Figure 4A illustrates representative output files generated using GUIDE in simple mode.

When GUIDE is run in full mode, two complementary output files are generated. The first file, `guide_results_full.csv`, reports the final integrated subtype prediction for each sample across four molecular subtypes: AFM, PPR, IME, and NEU. For each sample, the columns final_AFM, final_PPR, final_IME, and final_NEU provide subtype-specific probability scores ranging from 0 to 1, reflecting the confidence of GUIDE in assigning the sample to each subtype after integrating all available data modalities. These subtype-specific probability scores are computed using a one-versus-maximum-other formulation and integrated across modalities using a weighted soft-voting strategy. As a result, the four subtype values are not constrained to sum to 1 and should be interpreted as relative confidence scores rather than a multinomial probability vector. The column predicted_cluster reports the final subtype assignment, defined as the subtype with the highest integrated probability score. As in simple mode, the used_omics column indicates which data types

contributed to the final prediction. Together, these columns provide a complete probabilistic and categorical description of the subtype classification for each sample.

The second file generated in full mode, `guide_per_omics_probs.csv`, provides modality-specific subtype probability score estimates and is intended to support interpretability and quality control. Each row corresponds to one sample, and subtype probability scores are reported separately for each available data modality. Columns prefixed with gene_ (such as gene_AFM, gene_PPR, gene_IME, and gene_NEU) report subtype probability scores inferred using gene expression data alone, whereas columns prefixed with protein_ report subtype probability scores inferred using proteomics data alone. The same naming convention is applied to image- and DNA methylation–derived predictions. If a particular data modality is not available for a given sample, the corresponding columns are left empty. The used_omics column summarizes which data modalities were present for that sample. These modality-specific probability scores are internally combined by GUIDE using a soft-voting strategy to generate the integrated probability scores reported in `guide_results_full.csv`. Figures 4B and 4C illustrates representative output files generated using GUIDE in full mode.

Overall, successful completion of this protocol yields reproducible and interpretable subtype prediction outputs. Researchers can expect GUIDE to generate consistent probability score estimates and subtype assignments given identical inputs, enabling validation, comparison across datasets, and downstream exploratory or confirmatory analyses.

LIMITATIONS

The current version of GUIDE provides prognostic subtype predictions specifically for IDH-mutant astrocytoma and has not been evaluated for other glioma entities or cancer types. The pre-trained models were developed using data from defined cohorts, and prediction performance may be reduced when applied to external datasets with differences in patient populations, tissue processing, sequencing platforms, or histopathology image acquisition protocols.

GUIDE requires gene expression and proteomics inputs to be fully preprocessed prior to execution and does not perform additional normalization or quality control during inference. For histopathology data, image-based analysis requires GPU-enabled execution; CPU-only mode is not supported for image inputs in the current version. In addition, GUIDE relies on Docker and command-line execution, which may limit accessibility for users without prior experience in computational workflows. Variability in data availability across modalities may also affect subtype confidence, particularly when predictions are generated from a single data type.

TROUBLESHOOTING

Problem 1

GUIDE fails to process the molecular input files because samples are provided as columns and features as rows (related to step-by-step method details, step 2).

Potential solution

GUIDE requires samples as rows and features as columns. Transpose the input matrix so that each row corresponds to one sample and each column corresponds to one molecular feature, as described in section data pre-processing, step 2. Ensure that the first column contains sample identifiers and the first row contains feature names before running GUIDE.

Problem 2

GUIDE completes execution but produces unexpected or biologically implausible subtype probability score patterns (related to step-by-step method details, step 2).

Potential solution

Verify that gene expression and proteomics inputs have been log2 transformed, quantile normalized and scaled prior to execution. GUIDE does not perform additional normalization internally. Confirm preprocessing steps as described in section data pre-processing, step 2. and re-run GUIDE.

Problem 3

GUIDE fails or terminates unexpectedly when histopathology image data are provided (related to step-by-step method details, step 6 and 7).

Potential solution

In the current version of GUIDE, CPU-only execution is not supported for image-based analysis. Re-run GUIDE using GPU-enabled execution as described in section run GUIDE, step 6 or step 7, and ensure that GPU resources are correctly exposed to Docker.

Problem 4

Histopathology images appear to be ignored during execution, or image-based results are missing (related to step-by-step method details, step 1 and 6).

Potential solution

Confirm that whole-slide images are stored in the correct directory and use supported formats (.svs or.ndpi), and that GPU execution is enabled. Refer to section data pre-processing, step 1 to verify image readiness and section run GUIDE, step 6 or step 7 to ensure GPU mode is used.

Problem 5

Some modality-specific probability score columns are empty in `guide_per_omics_probs.csv` (related to step-by-step method details, step 8 and expected outcomes).

Potential solution

This behavior is expected when a data modality is unavailable for a given sample. GUIDE reports modality-specific subtype probability scores only for modalities present in the input data. Check the used_omics column to confirm which data types were used for each sample, as described in section expected outcomes.

Problem 6

GUIDE completes successfully, but predicted subtypes differ substantially from known or expected subtype labels (related to step-by-step method details, step 9).

Potential solution

After confirming correct preprocessing (see problem 2), review the modality-specific outputs in `guide_per_omics_probs.csv` (full mode) to identify whether disagreement originates from a particular data modality. Because GUIDE integrates subtype probability scores using predefined weights, systematic bias or distribution shifts in one modality (e.g., due to batch or platform differences) may affect the final prediction.

Re-run GUIDE using selected modalities (e.g., excluding a suspected modality) and compare subtype assignments. Since GUIDE re-normalizes weights across available modalities, excluding an inconsistent modality may change the integrated result and help identify the source of discrepancy.

In addition, if the two highest subtype probability scores are very close (e.g., difference < 0.1), the sample may represent mixed subtype characteristics rather than a clear assignment to a single subtype. In such cases, the prediction should be interpreted cautiously as reflecting potential biological heterogeneity.

If discrepancies persist across modality combinations, this may reflect broader differences between the new dataset and the cohorts used to develop GUIDE. In such cases, results should be interpreted within the context of dataset-specific characteristics. GUIDE does not support retraining or parameter modification in the current release.

Problem 7

GUIDE execution is slow or terminates with memory-related errors (related to step-by-step method details, step 6 and 7).

Potential solution

Ensure that the system meets the recommended hardware requirements and that sufficient shared memory is allocated when running Docker (e.g., –shm-size 12G). Refer to section run GUIDE, steps 6 and step 7 for recommended execution settings.

Problem 8

Users encounter difficulty executing GUIDE commands or retrieving output files (related to step-by-step method details, step 4 and 8).

Potential solution

Confirm that Docker is installed and the GUIDE image has been downloaded as described in the Download and set up the GUIDE Docker container subsection under the before you begin section. Verify command syntax carefully and ensure that the local data directory is correctly mounted, as described in section run GUIDE, step 4 and step 8.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. Jiguang Wang (jgwang@ust.hk).

Technical contact

Technical questions on executing this protocol should be directed to and will be answered by the technical contact, Dr. Jihong Tang (jtangbd@connect.ust.hk).

Materials availability

This study did not generate new unique reagents or materials.

Data and code availability

- All original code has been deposited at GitHub and is publicly available at https://github.com/WangLabHKUST/GUIDE.
- A version of record of the GUIDE code repository has been archived at Zenodo and is publicly available at https://doi.org/10.5281/zenodo.19317032.
- The GUIDE Docker image has been deposited at Zenodo and is publicly available at https://doi.org/10.5281/zenodo.18217774.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon reasonable request.

ACKNOWLEDGMENTS

This work was supported by the NSFC/RGC Collaborative Research Scheme (no. 82261160578 to Jiang laboratory and no. CRS_HKUST605/22 to Wang laboratory), the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0525300), the Guangdong Science and Technology Department (no. 2025A0505000039), the RGC grants (R6003–22, C6040-24G, 16101021, 16102724, and T12-705/24R), the ITC grants (ITCPD/17-9 and MHP/004/19), and the CAS Croucher Funding Scheme (CAS24SC03). J.W. is also supported by the Padma Harilela Professorship and the Sinovac Fellowship Program (Z1541). The authors would like to thank all contributors to the Chinese Glioma Genome Atlas and the Asian Glioma Genome Atlas.

AUTHOR CONTRIBUTIONS

J.W. supervised the study, provided conceptual guidance, and secured funding. J.T. conceptualized the protocol, developed and implemented the GUIDE computational framework, and performed data analysis. Y.Q. contributed to Docker

workflow development and testing. Y.R. developed the histopathology image analysis module and assisted with method optimization and validation. Members of the J.W. laboratory assisted with testing and benchmarking the GUIDE framework. J.W. and J.T. wrote the manuscript, which was then revised and proofread by all authors.

DECLARATION OF INTERESTS

J.W., J.T., and Y.R. have filed a CN patent on GUIDE (application number 202511218061.X), filed on August 28, 2025.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this protocol, the authors used DeepSeek-R1 to polish the manuscript for clarity, coherence, and conciseness. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

SUPPLEMENTAL INFORMATION

Supplementary data related to this article can be found online at https://doi.org/10.1016/j.xpro.2026.104512.

REFERENCES

1. Tang, J., Fan, W., Ruan, Y., Liu, X., Qiu, F., Feng, J., Huang, G., Yan, M., Wang, H., Mu, Q., et al. (2025). Protein-based classification reveals an immune-hot subtype in IDH mutant astrocytoma with worse prognosis. Cancer Cell *43*, 2136–2155.e14. https://doi.org/10.1016/j.ccell.2025.08.006.

2. Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. Linux J. *239*, 2.