nature communications

Article

https://doi.org/10.1038/s41467-025-57430-4

DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis

Received: 11 March 2024

Accepted: 24 February 2025

Published online: 06 March 2025

Check for updates

Wei Liu[1,9], Jun Li[1,9], Yitao Tang[1,2], Yining Zhao[1], Chaozhong Liu[3], Meiyi Song[4], Zhenlin Ju[1], Shwetha V. Kumar[1], Yiling Lu[5], Rehan Akbani[1], Gordon B. Mills[6] & Han Liang[1,2,7,8] ✉

Functional proteomics provides critical insights into cancer mechanisms, facilitating the discovery of novel biomarkers and therapeutic targets. We have developed a comprehensive cancer functional proteomics resource using reverse phase protein arrays, incorporating data from nearly 8000 patient samples from The Cancer Genome Atlas and approximately 900 samples from the Cancer Cell Line Encyclopedia. Our dataset includes a curated panel of nearly 500 high-quality antibodies, covering all major cancer hallmark pathways. To enhance the accessibility and analytic power of this resource, we introduce DrBioRight 2.0 (https://drbioright.org), an intuitive bioinformatic platform powered by state-of-the-art large language models. DrBioRight enables researchers to explore protein-centric cancer omics data, perform advanced analyses, visualize results, and engage in interactive discussions using natural language. By streamlining complex proteogenomic analyses, this tool accelerates the translation of large-scale functional proteomics data into meaningful biomedical insights.

Over the last decade, remarkable progress has been achieved in the generation of cancer omics data, particularly at the DNA and RNA levels in patient tumors. Landmark initiatives such as The Cancer Genome Atlas (TCGA)[1] and the Cancer Cell Line Encyclopedia (CCLE)[2] have played pivotal roles in this transformative era. Despite these strides, a critical gap persists in our understanding of the translational and post-translational landscape of human cancers, especially across many cancer lineages. To fill this critical gap, reverse phase protein arrays (RPPAs) provide a powerful platform for large-scale functional proteomics data of cancer samples in a sensitive, high-throughput, cost-effective manner[3–5]. Previously, we used this platform to profile approximately 8000 samples from TCGA patient tumors and 900 samples from CCLE cell lines, focusing on over 200 clinically relevant protein markers[2,6]. To facilitate a broad community to capitalize on these data, we built a user-friendly data portal, TCPA[7–9], for exploring the data in a rich context.

However, two notable challenges limit the immediate utility of TCPA. First, the previous RPPA data have limited coverage of protein markers (~200 only). Second, the data portal only provides several predefined analytic modules, with little flexibility for user-defined analyses. To address these challenges, we have recently expanded our RPPA protein panel to approximately 500 high-quality antibodies[10]. This expansion has enabled the development of a comprehensive, high-quality pan-cancer functional proteomics compendium, termed

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Houston, Houston, TX, USA. [3]Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA. [4]Brown University, Providence, RI, USA. [5]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [6]Knight Cancer Institute and Cell, Developmental and Cancer Biology, Oregon Health & Science University, Portland, OR, USA. [7]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [8]Institute for Data Science in Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [9]These authors contributed equally: Wei Liu, Jun Li. ✉e-mail: hliang1@mdanderson.org

RPPA500, integrating data from both TCGA and CCLE samples. Alongside with our expanded proteomic dataset, here we introduce DrBioRight 2.0 (https://drbioright.org), a cutting-edge chatbot powered by large language models (LLMs). This tool is designed to lower technical barriers, enabling seamless analysis of complex omics data. Users with diverse backgrounds can easily access, analyze, and visualize data seamlessly through intuitive natural language queries.

Results

Using the well-established data processing pipeline[6,11] and following the guidelines established in the community[3], our RPPA500 compendium encompasses a total of 9000 samples, comprising both patient tumor and cancer cell line samples[10]. The TCGA cohort dataset includes protein expression profiles from 7828 patient tumors across 32 distinct cancer types (Fig. 1). Predominant tissue types in this dataset include breast (BRCA, $n = 881$), kidney (KIRC/KIRP/KICH, $n = 756$), and lung (LUAD/LUSC, $n = 693$). The CCLE cohort dataset covers 878 cancer cell lines, with lung, blood, lymphocyte, and colorectal lineages, each over 50 distinct cell lines (Fig. 1). Most of these cell lines have parallel functional data such as gene dependency, metastatic potential, and drug sensitivity data[12–19]. The final RPPA500 protein set contains 447 protein markers, including 357 total proteins and 90 post-translationally modified (PTM) proteins (e.g., phosphorylated proteins), and it is highly enriched in therapeutic targets and biomarkers (Supplementary Data 1). To underscore the expanded coverage of cancer-related pathways, we aligned the protein markers with hallmark gene sets[20]. Our RPPA500 protein panel comprehensively covers all 50 hallmark gene sets (Supplementary Fig. 1), including a robust coverage for apoptosis ($n = 43$), PI3K-Akt-mTOR signaling ($n = 34$), estrogen response ($n = 32$), hypoxia ($n = 31$), IL6-JAK-STAT3 signaling ($n = 31$), apical junction ($n = 29$), interferon response ($n = 26$), EMT ($n = 18$), G2M checkpoint ($n = 18$), P53 pathway ($n = 17$), KRAS signaling ($n = 12$), and DNA repair ($n = 7$). Compared to our previous protein panel[6], there is a significant increase of 115% in the number of total proteins and a 67% increase in the number of PTM proteins across these gene sets, highlighting a substantially increased capacity to comprehend cancer biology at the protein level.

Recent breakthroughs in LLM-based generative AI have ushered in a transformative era for data analytics[21–23]. In this study, we have developed a new LLM-based chatbot, DrBioRight 2.0, empowered with natural language processing, enabling users to explore, analyze, and visualize the above RPPA data intuitively and intelligently (Fig. 1). Specifically, we first generated a unified multi-omics dataset with standardization and normalization of patient clinical data, molecular profiling data at DNA, RNA, and RPPA500-based protein levels, as well as cell line phenotypic datasets. Collectively, over 1 billion data values were curated and restructured under the HDF5 format in a No-SQL database hosted on an I/O efficient cloud-based server. Addressing the long-standing challenge of non-standard protein annotation, we thoroughly reviewed protein markers and cross-referenced them with external databases to comprehensively annotate proteins at individual, pathway, functional, and disease levels. This detailed annotation facilitates user-friendly analysis of data with biologically driven questions. DrBioRight has several features that are not available in conventional analytics platforms, including natural language understanding, transparency and reproducibility, and user friendliness. These features are supported by several key cutting-edge techniques: (i) Chat UI: a real-time conversational-based chatting interface; (ii) Prompts: highly customizable LLM-oriented domain-knowledge-specific prompts; (iii) LLMs: LLM-empowered generative AI; (iv) Code generation: seamless code-generation-correction cycle; (v) Plugins: deep-nested interactive plugins provide a unique suite of tools for enhanced effective data visualization and analysis, such as interactive clustering heatmaps[24].

To demonstrate its utility, we present an illustrative example where users can easily query, “Please generate a heatmap for protein expression data of the current dataset.” In response, DrBioRight dynamically processes the data and calls the corresponding heatmap plugin to generate an interactive heatmap (Fig. 2A). Similar to other interactive plugins we have implemented, the heatmap plugin can efficiently handle large datasets. It offers a comprehensive global overview along with numerous features (such as selection, zoom in/out, searching, 2D/3D scatter plots, pathway mapping, and linking to external resources) to facilitate effective data exploration. For a more detailed analysis, users can further ask, “Could you please show me the correlation between AKT2PS474 and IL6 expression?” DrBioRight then extracts the data, performs the corresponding statistical analysis, and presents the results in a clear scatter plot. Leveraging the same dataset, users can conduct a survival analysis by inquiring about the correlation between a protein and the patient survival time, followed by

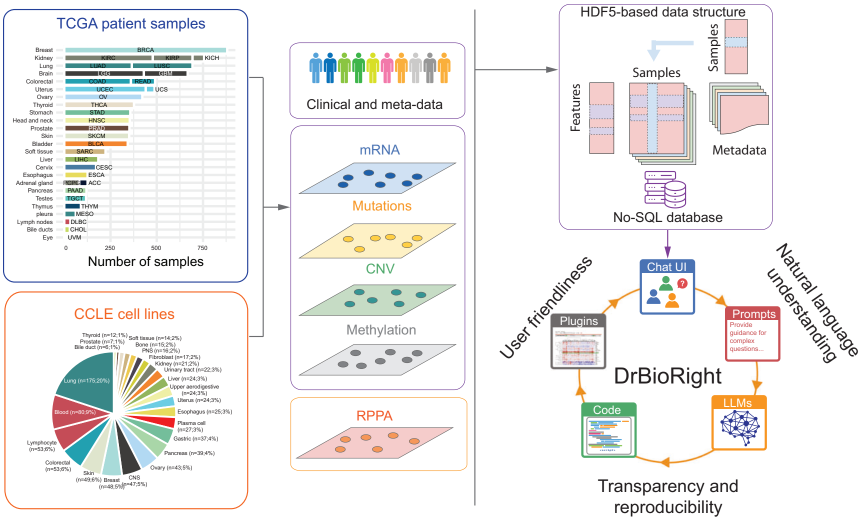
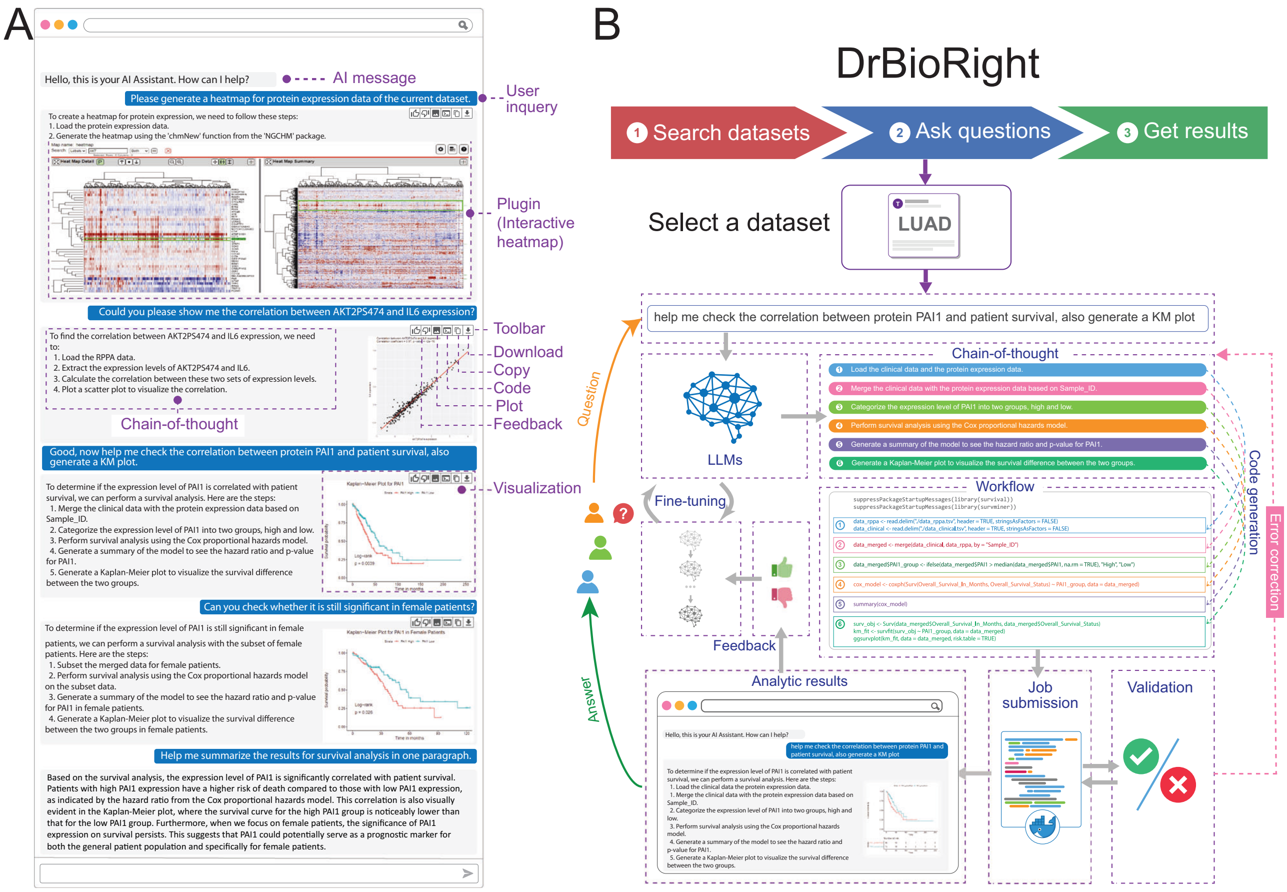


Fig. 1 | Overview of the data integration workflow and key innovations in DrBioRight 2.0. A workflow illustrates the complete process of data integration and technical features in DrBioRight 2.0, covering several key stages: starting with the sample cohorts, attributing data types, detailing the HDF5-based data format structure, and introducing three features. The features are supported by five cutting-edge technological components: conversational AI, domain-specific prompts, large language models (LLMs), automatic code generation, and interactive smart visualization. Source data are provided as a Source Data file.



Hello, this is your AI Assistant. How can I help?

Please generate a heatmap for protein expression data of the current dataset.

To create a heatmap for protein expression, we need to follow these steps:
1. Load the protein expression data.
2. Generate the heatmap using the 'chmNew' function from the 'NGCHM' package.

Could you please show me the correlation between AKT2PS474 and IL6 expression?

To find the correlation between AKT2PS474 and IL6 expression, we need to:
1. Load the RPPA data.
2. Extract the expression levels of AKT2PS474 and IL6.
3. Calculate the correlation between these two sets of expression levels.
4. Plot a scatter plot to visualize the correlation.

Good, now help me check the correlation between protein PAI1 and patient survival, also generate a KM plot.

To determine if the expression level of PAI1 is correlated with patient survival, we can perform a survival analysis. Here are the steps:
1. Merge the clinical data with the protein expression data based on Sample_ID.
2. Categorize the expression level of PAI1 into two groups, high and low.
3. Perform survival analysis using the Cox proportional hazards model.
4. Generate a summary of the model to see the hazard ratio and p-value for PAI1.
5. Generate a Kaplan-Meier plot to visualize the survival difference between the two groups.

Can you check whether it is still significant in female patients?

To determine if the expression level of PAI1 is still significant in female patients, we can perform a survival analysis with the subset of female patients. Here are the steps:
1. Subset the merged data for female patients.
2. Perform survival analysis using the Cox proportional hazards model on the subset data.
3. Generate a summary of the model to see the hazard ratio and p-value for PAI1 in female patients.
4. Generate a Kaplan-Meier plot to visualize the survival difference between the two groups in female patients.

Help me summarize the results for survival analysis in one paragraph.

Based on the survival analysis, the expression level of PAI1 is significantly correlated with patient survival. Patients with high PAI1 expression have a higher risk of death compared to those with low PAI1 expression, as indicated by the hazard ratio from the Cox proportional hazards model. This correlation is also visually evident in the Kaplan-Meier plot, where the survival curve for the high PAI1 group is noticeably lower than that for the low PAI1 group. Furthermore, when we focus on female patients, the significance of PAI1 expression on survival persists. This suggests that PAI1 could potentially serve as a prognostic marker for both the general patient population and specifically for female patients.

help me check the correlation between protein PAI1 and patient survival, also generate a KM plot

1. Load the clinical data and the protein expression data.
2. Merge the clinical data with the protein expression data based on Sample_ID.
3. Categorize the expression level of PAI1 into two groups, high and low.
4. Perform survival analysis using the Cox proportional hazards model.
5. Generate a summary of the model to see the hazard ratio and p-value for PAI1.
6. Generate a Kaplan-Meier plot to visualize the survival difference between the two groups.

```
suppressPackageStartupMessages(library(survival))
suppressPackageStartupMessages(library(survminer))

① data_rppa <- read.delim("./data_rppa.tsv", header = TRUE, stringsAsFactors = FALSE)
  data_clinical <- read.delim("./data_clinical.tsv", header = TRUE, stringsAsFactors = FALSE)
② data_merged <- merge(data_clinical, data_rppa, by = "Sample_ID")
③ data_merged$PAI1_group <- ifelse(data_merged$PAI1 > median(data_merged$PAI1, na.rm = TRUE), "High", "Low")
④ cox_model <- coxph(Surv(Overall_Survival_In_Months, Overall_Survival_Status) ~ PAI1_group, data = data_merged)
⑤ summary(cox_model)
⑥ surv_obj <- Surv(data_merged$Overall_Survival_In_Months, data_merged$Overall_Survival_Status)
  km_fit <- survfit(surv_obj ~ PAI1_group, data = data_merged)
  ggsurvplot(km_fit, data = data_merged, risk.table = TRUE)
```

Fig. 2 | **Overview of the DrBioRight 2.0 platform. A** Snapshot of the DrBioRight user interface. **B** A diagram illustrating the detailed design chart of the chatbot ecosystem.

visualization through Kaplan-Meier plots. In contrast to the previous analytic modules at TCPA, DrBioRight distinguishes itself by offering versatile analyses, including customizable interactions with the chatbot. For instance, after performing a survival analysis across all the samples in the full cohort, users can further investigate specific associations within male or female patients or change the colors in a plot. Another noteworthy feature of DrBioRight is its seamless transition between analytics-driven and general questions. As depicted in Fig. 2A, users can request the chatbot to summarize the results. Moreover, DrBioRight allows users to download the corresponding project report in an R markdown file and run it in RStudio locally to reproduce the analysis (Supplementary Fig. 2A). These features collectively position DrBioRight as a highly convenient analytic tool, providing unparalleled flexibility and customization in data analysis.

The system architecture of DrBioRight 2.0 comprises three integral components: (i) a No-SQL database, (ii) a back-end LLM-powered analytics module, and (iii) an interactive chat interface (Fig. 2B). To start an analysis, a user simply begins by selecting a disease (e.g., lung adenocarcinoma [LUAD]). Then, the chatbot automatically links relevant multi-omics data to the user's project space, making it ready for querying and analysis. The back-end LLMs will predict user's intent, distinguishing between general inquiries and questions requiring code generation or bioinformatics analysis. DrBioRight outputs a logical flow based on a chain-of-thought approach to enhance user understanding. In the back end, LLMs generate text-based answers or programming scripts on the fly. Before submission to the job queue, the platform reviews and validates codes, autonomously correcting common errors like missing libraries or incompatible package versions. Following successful result generation, the user-friendly chat interface displays the outcomes. For ongoing improvements, we integrate a rating function that allows users to evaluate analytic results, and the user feedback together with the expert manual evaluations will then guide iterative refinements to fine-tune LLMs through the reinforcement learning from human feedback (RLHF)[25–28].

To maximize the performance of DrBioRight 2.0, we have implemented cutting-edge techniques to enhance the LLMs (Fig. 3A). Overall, we incorporated a multi-agent workflow to build hierarchical agent teams using a graph architecture (Supplementary Fig. 2B). This framework can better organize the multi-agent system and streamline the development process (Methods). Each team consists of one or more agents or tools. For example, the multi-omics data analysis team uses a heatmap to provide a dataset overview and a survival analysis tool to link proteins with patient survival data. A correlation analysis tool performs association analyses between features including protein expression, mutations, and clinical variables. A supervisor routes team-specific questions to appropriate tools for task execution and analytic results. Each agent is powered by a model coupled with task specific prompts. These prompts include a mini knowledge base on our RPPA500 data, a summary of our meta-data, and general analysis information. To fine-tune LLMs, we curated and standardized thousands of user queries through expert review, creating both training and test datasets. Using the training dataset, we performed model fine-tuning through three steps: (i) initial supervised fine-tuning. The base model was initially fine-tuned using prompt and response pairs to learn domain-specific contexts. (ii) based on the fine-tuned model, we developed an evaluation system to allow domain experts to rank the AI responses (Supplementary Fig. 3). The evaluation datasets were further used to train a reward model. (iii) the optimization step was performed by the PPO (proximal policy optimization) trainer from Hugging Face. To evaluate its performance, we tested our platform using an independent test set of

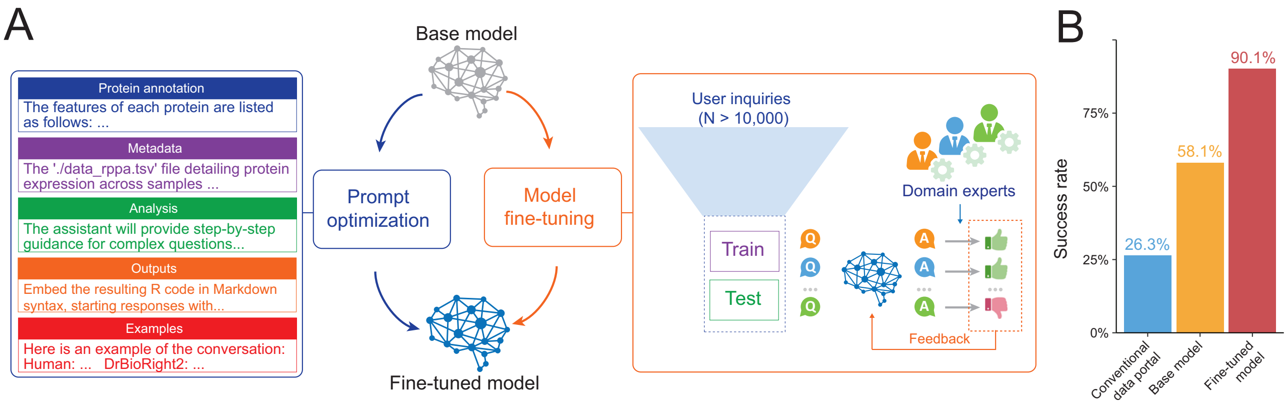

A
Protein annotation
The features of each protein are listed as follows: ...
Metadata
The './data_rppa.tsv' file detailing protein expression across samples ...
Analysis
The assistant will provide step-by-step guidance for complex questions...
Outputs
Embed the resulting R code in Markdown syntax, starting responses with...
Examples
Here is an example of the conversation: Human: ... DrBioRight2: ...
Base model
Prompt optimization
Model fine-tuning
Fine-tuned model
User inquiries (N > 10,000)
Train
Test
Q
A
Domain experts
Feedback
B
Success rate
0%
25%
50%
75%
26.3%
58.1%
90.1%
Conventional data portal
Base model
Fine-tuned model

Fig. 3 | Overview of the finetuning process and model evaluation. A Overview of the finetuning process applied to the foundational large language models. **B** Model evaluation.

queries not used in the fine-tuning process. Only 26% of the questions could be addressed by our classic TCPA platform (Fig. 3B), highlighting a major need for a versatile and customizable tool for such analyses. We then test the same questions using GPT-4 and achieved a 58% success rate, underscoring the limitations of a general LLM in addressing domain-specific questions through natural language-based data analytics. However, when employing the fine-tuned models under the graph-based workflow using LangGraph on the same set of questions, we achieved an impressive 90% success rate (Methods). This emphasizes the impact of incorporating domain-specific knowledge, fine-tuning process, and multi-agent workflow.

Discussion

DrBioRight 2.0 represents a major advancement for researchers engaging with cancer proteomics data, achieving three key milestones. First, it broadens the protein space for the most commonly used cohorts of cancer patients and cell lines, providing a unique and valuable resource for biomedical researchers. Second, the LLM-empowered chatbot, DrBioRight, offers an intuitive, versatile, and highly customizable platform, effectively lowering entry barriers and enabling researchers from diverse backgrounds to analyze data efficiently without extensive domain knowledge. Third, the deep integration between the data resource and the LLMs significantly amplifies the utility of such a resource. This integration not only increases data accessibility but also features a learn-by-use design, accelerates the user-developer feedback loop, and offers enhanced customization options. In contrast to traditional tools that often entail substantial efforts for integration and harmonization during development and iteration, DrBioRight adeptly addresses these challenges with its unique combination of a comprehensive data resource and advanced LLMs. We anticipate that similar efforts like DrBioRight will spearhead a paradigm shift in the next generation of data analysis and sharing platforms, ultimately fostering a comprehensive ecosystem tailored for biomedical researchers.

Methods

Compilation of RPPA500 dataset

Antibody validation was in accordance with the established RPPA pipeline, as outlined in our prior works[3,4,6]. The entire process of generating RPPA data from cell line and patient tumor samples from the CCLE and TCGA projects was described in our recent study[10]. Level 1 RPPA data were from the images using ArrayPro software, primarily consisting of protein signal intensities. Level 2 data were obtained from Level 1 data through a curve fitting analysis using the SuperCurve algorithm[29]. Level 3 data were obtained from Level 2 data through median-centering normalization. Finally, to ensure the data quality and consistency, we generated Level 4 data by applying a replicate-based normalization method to Level 3 data. All the subsequent analyses and input data for DrBioRight are based on Level 4 data. The proteins were associated with pathways by aligning their corresponding gene identifiers with the member genes of the gene sets obtained from the Human Molecular Signature Database (MSigDB[20]), and the network visualization was based on the RCy3[30] package.

Curation and preprocessing of other data

We integrated all molecular, functional, and clinical data and the corresponding metadata from TCGA (https://portal.gdc.cancer.gov) and CCLE (https://depmap.org/portal/), subsequently converting them into HDF5 format. This conversion facilitates efficient and effective real-time data extraction and analysis. Additionally, we comprehensively annotated the protein markers within the RPPA500 dataset to enable users to search and analyze the protein markers effectively.

Platform architecture

The DrBioRight 2.0 platform consists of two major components: (i) a client web interface and (ii) a backend server system. Specifically, the front-end web interface was built based on React (https://react.dev) and MUI. The backend system includes (i) Graph-based agentic workflows (LangGraph and LangChain), which provides multi-agents framework to process, evaluate, and analyze user requests. (ii) LLM APIs generate responses based on different types of requests, including text/code generation (e.g., OpenAI/GPT series and Llama 3 models), and sentence classification (injection attack detection models); (iii) Code execution environment (including customized packages, such as DrBioRight's survival, network, and report generation packages), and (iv) No-SQL database (MongoDB) for user and data management.

Training and test data

We curated a collection of >10,000 user queries including the feedback from >250 DrBioRight unique users, rigorous internal testing, and contributions from domain experts. This dataset formed the core of our training and test sets. To prevent potential overfitting and ensure robustness, the training and test sets were obtained from different user pools. This approach ensures that the model does not simply memorize user-specific patterns.

Prompt engineering

In the fine-tuning process, we utilized the training set in conjunction with custom-designed prompts. These prompts were specifically crafted to align with the tasks related to those different analyses. Specifically, on the top level, we used prompts to facilitate routing the original user queries to their related analytic agents. For each agent, the prompt defines how it processes and analyzes the user queries and outputs the final responses. To optimize each agent-specific prompt, we first designed multiple initial prompts and tested against same set of user queries. After manual evaluation from experts, we chose those prompts with the best performance as the final agent prompts. All the selected prompts are version-controlled and deposited in our prompt repository. Thus, the prompts were improved through an iterative approach.

Model finetuning and alignment

The overall platform architecture allows us to utilize multiple models during the entire workflow including both public and internally fine-tuned models. For example, to obtain initial training and evaluation datasets, we utilized OpenAI GPT4/4o and Llama 3 models to generate responses for user questions. To fine-tune models, we used python libraries from Hugging Face. Specifically, models were initially finetuned based on SFTTrainer for supervised fine-tuning. The reward model was trained based on expert scored prompt-response pairs (chosen vs rejected). Based on the reward model and initial fine-tuned model, we further performed optimization by PPO Trainer. The model was then detoxified by a second round of PPO training with a toxicity evaluation model facebook/roberta-hate-speech-dynabench-r4-target model.

Model evaluation

Model evaluation was performed routinely before each public release. The evaluation samples were selected from a user pool independent from that of the training set. We developed an evaluation pipeline to automatically submit user queries to DrBioRight and generate the response reports in a PDF format. Based on the reports, our experts manually reviewed the content. To estimate the successful rate for our classic TCPA platform, we consider whether those analyses can be performed by the modules available on our classic TCPA platform. The user queries not covered by the classic TCPA include general conversations, new, and customized analysis (e.g., changing colors, labels, and choosing user defined sample cohorts). To assess the successful rate for OpenAI/GPT-4o, for a fair comparison, we provided it with a system-level prompt describing all the information of the associated data types and their meta-data.

Defense strategies

To minimize security vulnerabilities, we implemented several defense strategies: (i) Input sanitization. To enhance security against potential injection attacks, a prompt injection identification model from Hugging Face was integrated (protectai/deberta-v3-base-prompt-injection-v2). We tested its performance by challenging our platform with >100 injection attack prompts, all of which were successfully identified. For example, when a user attempted to inject code to delete all the system files, our platform successfully detected and blocked the malicious command. (ii) Rate limit. DrBioRight monitors query frequency per user to prevent load attacks. DrBioRight would pause for a user if he/she submits concurrent queries in a pre-defined short time interval. (iii) Environment isolation. This effort ensures that all code executions occur in an isolated environment under a non-root user account, thereby minimizing the impact of harmful code and safeguarding other system components.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the RPPA data are available for download from the website (https://drbioright.org/resources/). Additionally, the complete molecular datasets can be downloaded from TCGA (https://portal.gdc.cancer.gov) and DepMap (https://depmap.org/portal/). Source data are provided with this paper.

Code availability

The compiled software and detailed description of the code's functionality is available at https://drbioright.org. Supplementary Data 2 provides the list of key modules/packages used in this study.

References

1. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet* **45**, 1113–1120 (2013).
2. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
3. Akbani, R. et al. Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: a workshop report: the RPPA (Reverse Phase Protein Array) society. *Mol. Cell Proteom.* **13**, 1625–1643 (2014).
4. Hennessy, B. T. et al. A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clin. Proteom.* **6**, 129–151 (2010).
5. Paweletz, C. P. et al. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**, 1981–1989 (2001).
6. Akbani, R. et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.* **5**, 3887 (2014).
7. Chen, M. M. et al. TCPA v3.0: an integrative platform to explore the pan-cancer analysis of functional proteomic data. *Mol. Cell Proteom.* **18**, S15–S25 (2019).
8. Li, J. et al. TCPA: a resource for cancer functional proteomics data. *Nat. Methods* **10**, 1046–1047 (2013).
9. Li, J. et al. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell* **31**, 225–239 (2017).
10. Li, J. et al. A protein expression atlas on tissue samples and cell lines from cancer patients provides insights into tumor heterogeneity and dependencies. *Nat. Cancer* **5**, 1579–1595 (2024).
11. Siwak, D. R., Li, J., Akbani, R., Liang, H. & Lu, Y. Analytical Platforms 3: processing samples via the RPPA pipeline to generate large-scale data for clinical studies. *Adv. Exp. Med. Biol.* **1188**, 113–147 (2019).
12. Garnett, M. J. et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
13. Basu, A. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
14. Corsello, S. M. et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **1**, 235–248 (2020).
15. Dempster, J. M. et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biol.* **22**, 343 (2021).
16. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
17. Jin, X. et al. A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020).
18. Seashore-Ludlow, B. et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).
19. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).
20. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

21. Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems*. 30 (2017).
22. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
23. Achiam, J. et al. Gpt-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
24. Broom, B. M. et al. A galaxy implementation of next-generation clustered heatmaps for interactive exploration of molecular profiling data. *Cancer Res.* **77**, e23–e26 (2017).
25. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
26. Nakano, R. et al. Webgpt: browser-assisted question-answering with human feedback. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2112.09332 (2021).
27. Ziegler, D. M. et al. Fine-tuning language models from human preferences. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1909.08593 (2019).
28. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1707.06347 (2017).
29. Hu, J. et al. Non-parametric quantification of protein lysate arrays. *Bioinformatics* **23**, 1986–1994 (2007).
30. Gustavsen, J. A., Pai, S., Isserlin, R., Demchak, B. & Pico, A. R. RCy3: network biology using Cytoscape from within R. *F1000Res.* **8**, 1774 (2019).

Acknowledgements

This study was supported by the National Institutes of Health (U24CA264128 to H.L., R.A., and G.B.M., U01CA253472 and U01CA281902 to H.L. and G.B.M., R01CA251150 to H.L., P50CA221703 and P50CA281701 to H.L., U24CA210950 and U24CA210949 to R.A. and G.B.M., R50CA221675 to Y.L., P50CA70907 to J.L., and P30CA016672). This study was also supported by the University Cancer Foundation via the Institutional Research Grant program to J.L. and the Barnhart Family Distinguished Professorship in Targeted Therapies to H.L. at the University of Texas MD Anderson Cancer Center.

Author contributions

J.L. and H.L. conceived of the project. W.L., J.L., Y.T., Y.Z., C.L., M.S., Z.J., S.V.K., Y.L., R.A., G.B.M., and H.L. contributed to the data analysis and result discussion. J.L. and H.L. wrote the paper with input from other authors. H.L. supervised the whole project.

Competing interests

R.A. is a bioinformatics consultant for the University of Houston. G.B.M. is on the scientific advisory board and/or a consultant for Amphista, Astex, AstraZeneca, BlueDot, Chrysallis Biotechnology, Ellipses Pharma, GSK, ImmunoMET, Infinity, Ionis, Leapfrog Bio, Lilly, Medacorp, Nanostring, Nuvectis, PDX Pharmaceuticals, Qureator, Roche, Signalchem Lifesciences, Tarveda, Turbine, and Zentalis Pharmaceuticals. G.B.M. has stock or receives income from BlueDot, Catena Pharmaceuticals, ImmunoMet, Nuvectis, SignalChem, Tarveda, and Turbine. G.B.M. has licensed HRD assay to Myriad Genetics. G.B.M. is an inventor on a patent (U.S. Patent No. 10501777) titled "Simultaneous quantification of a plurality of proteins in a user-defined region or a cross-sectioned tissue," licensed to NanoString Technologies, Inc., which is unrelated to this study. G.B.M. also receives income for sponsored research from AstraZeneca. H.L. is a shareholder and advisor for Precision Scientific. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57430-4.

Correspondence and requests for materials should be addressed to Han Liang.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.