HONeYBEE: Enabling Scalable Multimodal AI in Oncology Through Foundation Model–Driven Embeddings

Aakash Tripathi[1,3*†], Asim Waqas[1,2†], Matthew B. Schabath[2], Yasin Yilmaz[3], Ghulam Rasool[1,3]

[1]Department of Machine Learning, Moffitt Cancer Center & Research Institute, 12902 USF Magnolia Drive, Tampa, 33620, Florida, USA.
[2]Departments of Cancer Epidemiology, Moffitt Cancer Center & Research Institute, 12902 USF Magnolia Drive, Tampa, 33620, Florida, USA.
[3]Department of Electrical Engineering, University of South Florida, 4202 E Fowler Ave, Tampa, 33620, Florida, USA.

*Corresponding author(s). E-mail(s): aakash.tripathi@moffitt.org; Contributing authors: asim.waqas@moffitt.org; matthew.schabath@moffitt.org; yasiny@usf.edu; ghulam.rasool@moffitt.org;
†These authors contributed equally to this work.

Abstract

The increasing availability of multimodal biomedical data has created new opportunities for AI-driven oncology. However, integrating these heterogeneous data sources into unified, high-quality representations remains a critical challenge. We introduce **H**armonized **ON**colog**Y** **B**iomedical **E**mbedding **E**ncoder (HONeYBEE), a modular, open-source framework that preprocesses diverse oncology datasets—including clinical notes, pathology slides, radiology scans, and molecular profiles—to generate consistent, high-dimensional data representations (also known as embeddings) using domain-specific foundation models. HONeYBEE simplifies downstream applications such as prognosis estimation, cancer subtype classification, and patient retrieval through a unified API. Using multimodal data from over 11,400 patients in The Cancer Genome Atlas (TCGA), we show that HONeYBEE's integrated representations improve cancer-type separability in t-SNE space and enhance survival prediction. Multimodal embeddings achieved

superior performance across all survival models (C-index: CoxPH = 0.6587, RSF = 0.6252, DeepSurv = 0.6621), outperforming single-modality baselines. Fine-tuned clinical embeddings further improved classification accuracy by 14.8% compared to pretrained models. By releasing all representations, models, and pipelines, HONeYBEE enables reproducible, scalable, and clinically meaningful AI workflows for precision oncology.

Keywords: Multi-modal, Representation Learning, Computational Pathology, Foundation Models, Precision Medicine

1 Introduction

Recent advances in computational oncology have been fueled by the increasing digitization of diverse biomedical data—ranging from clinical documentation and radiological imaging to digital pathology slides and high-dimensional molecular profiles [6–11]. This wealth of multimodal data offers unprecedented opportunities to improve patient stratification, predict treatment response, and model disease progression [7, 8, 11]. In parallel, the adaptation of deep learning techniques from computer vision and natural language processing has enabled powerful solutions in these domains [8, 12]. However, despite these gains, a fundamental challenge persists: the lack of robust and generalizable methods for integrating heterogeneous data sources into unified representations that reflect the full complexity of cancer biology [13].

Despite the availability of biomedical data in oncology, much of it remains underutilized due to its fragmentation across modalities—clinical narratives, radiologic imaging, pathology slides, and molecular profiles are often analyzed in isolation [13, 14]. Recent advances in self-supervised learning have enabled the development of foundation models (FMs) that extract meaningful representations from large-scale unlabeled datasets [6, 15, 16]. These models, pretrained on text, images, or omics data, learn latent features that capture the domain-specific structure and can be transferred to downstream tasks such as cancer classification or survival prediction. However, most FMs are designed to operate within a single or dual modality, missing the opportunity to leverage the complementary information embedded across diverse data types [17]. In oncology, while data availability is growing, a more critical bottleneck lies in the lack of standardized, scalable tools to unify these modalities into cohesive, high-quality representations.

We hypothesize that integrating FM-derived representations from multiple modalities can yield richer and more clinically informative embeddings than single-modality approaches. Rather than relying solely on increased model scale or parameter count, we propose that the information content across diverse biomedical data types provides a powerful and orthogonal path for improving predictive performance in oncology. To test this hypothesis, we present HONeYBEE, an open-source framework for multimodal representation learning and integration in cancer (https://lab-rasool.github.io/HoneyBee/). **H**armonized **ON**colog**Y** **B**iomedical

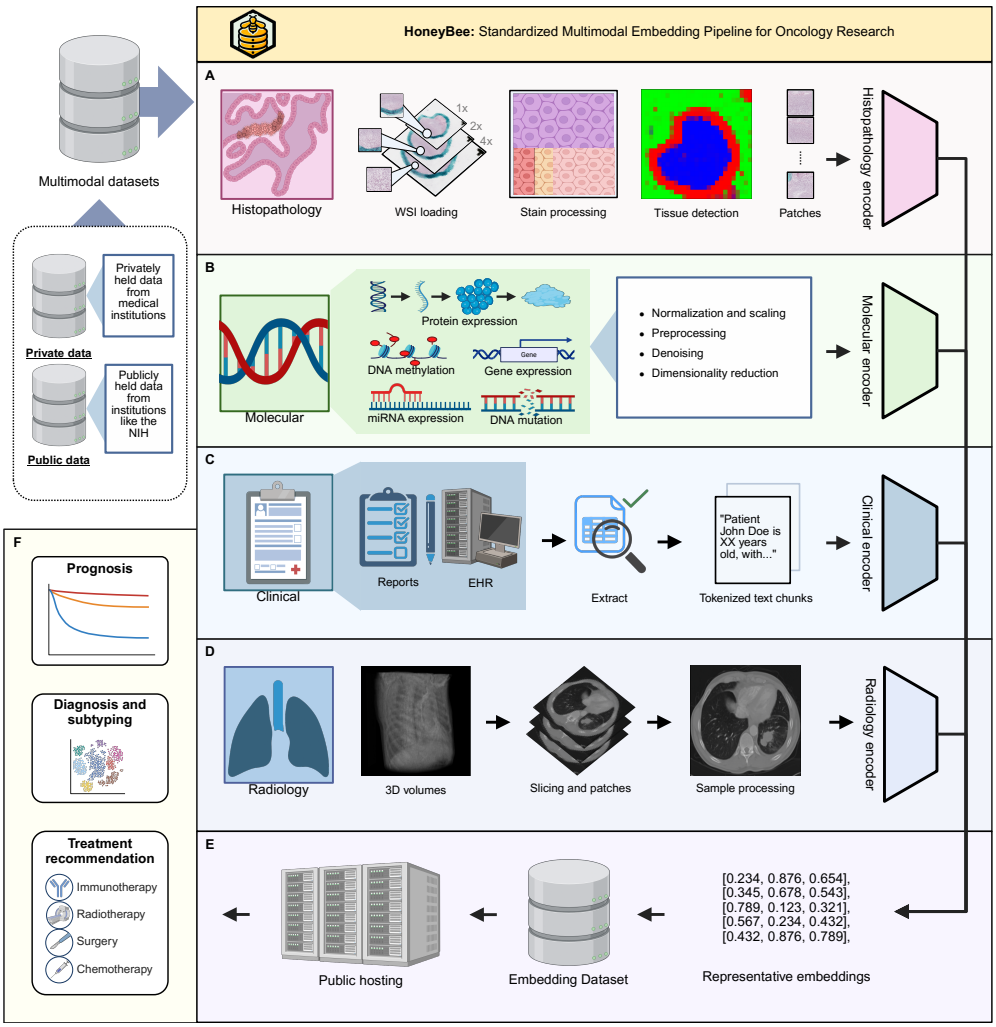


Fig. 1 Overview of the HONeYBEE framework. The HONeYBEE framework enables scalable integration and embedding of diverse oncology data types to support AI model development. Multimodal datasets from public repositories, such as the NCI CRDC [1], including the PDC [2], GDC [3], IDC [4], and TCIA [5], as well as private institutional datasets can be processed through dedicated modality-specific pipelines. **(A)** Histopathology images are processed via WSI loading, stain normalization, tissue detection, and patch extraction prior to the extraction of embeddings. **(B)** Molecular data undergo cleaning and normalization across multiple omics layers, including DNA methylation, gene expression, and somatic mutations. **(C)** Clinical text, including pathology and radiology reports and EHR, is extracted and tokenized into coherent text segments for downstream model processing. **(D)** Radiology images are prepared using volumetric slicing, patch extraction, and modality-specific preprocessing optimized for downstream tasks. **(E)** Embeddings, fixed-length feature vectors that capture the most informative patterns from each data modality, are generated using FMs. HONeYBEE supports both pretrained, publicly available, and custom FMs via the Hugging Face ecosystem, allowing researchers to tailor models to specific use cases. All extracted embeddings and processed datasets are publicly released to promote reproducibility, sharing, and external validation. **(F)** Finally, these multimodal embeddings serve as a standardized input for downstream oncology tasks, including prognosis estimation, cancer subtype classification, treatment decision support, and broader AI model development.

Embedding **E**ncoder or HONeYBEE harmonizes clinical, radiologic, pathologic, and molecular data using domain-specific FMs, and produces unified patient embeddings optimized for downstream machine learning and AI tasks such as cancer subtype classification, survival prediction, and patient similarity retrieval.

While numerous models, algorithms, and data processing pipelines have been proposed for analyzing clinical, imaging, and molecular data, most existing tools are designed for single-modality workflows and lack the flexibility to support integrated, end-to-end pipelines [6, 18]. Published methods are often implemented as isolated codebases with rigid dependencies, domain-specific interfaces, and limited extensibility, which complicates reproducibility and hinders multimodal experimentation [19]. Moreover, the absence of standardized pipelines for data harmonization and embedding extraction increases the technical burden on researchers and slows the development of clinically meaningful models [20].

HONeYBEE enables oncology researchers to generate high-quality, unified embeddings from diverse data modalities using domain-specific FMs [13, 21–25]. The framework supports the structured integration of clinical notes, radiology images, pathology slides, and molecular profiles, with dedicated preprocessing and representation pipelines for each modality. Through a simplified API and modular design, users can easily incorporate new FMs, compare embedding strategies, and deploy embeddings for tasks such as cancer subtype classification, survival prediction, patient similarity retrieval, and retrieval-augmented generation. By abstracting away the complexities of data harmonization, model configuration, and format translation, HONeYBEE empowers researchers to focus on hypothesis-driven analysis while ensuring reproducibility and scalability across studies.

In summary, HONeYBEE is a unified, open-source framework designed to overcome the fragmentation in multimodal oncology analysis. It seamlessly integrates data from various sources including the NCI Cancer Research Data Commons (CRDC), Proteomics Data Commons (PDC), Genomic Data Commons (GDC), Imaging Data Commons (IDC), and The Cancer Imaging Archive (TCIA). An overview of the framework's architecture and supported modalities is shown in Fig. 1. HONeYBEE features a simplified API, standardized pipelines, and support for clinical, radiologic, pathologic, and molecular data, enabling scalable and reproducible workflows without managing the complexities of data harmonization or modality-specific preprocessing. The framework includes pretrained FMs, supports adding new FMs, dedicated pipelines for each modality, and compatibility with widely adopted platforms such as PyTorch, Hugging Face, and FAISS [26, 27]. Unlike existing tools that focus on single modalities or require extensive customization, HONeYBEE supports flexible model deployment, seamless cross-modality integration, and minimal-code implementation of state-of-the-art methods [28–30]. In the sections that follow, we evaluate HONeYBEE's performance on diverse oncology tasks, clustering, survival prediction, classification, and patient retrieval, using the The Cancer Genome Atlas (TCGA) dataset, demonstrating its effectiveness as a generalizable platform for multimodal cancer research.

2 Results

We developed and evaluated HONeYBEE, an open-source framework for large-scale multimodal data integration in oncology. HONeYBEE generates unified, high-quality embeddings from diverse data types, including clinical narratives, pathology whole-slide images (WSIs), radiology images, and molecular profiles, to support downstream applications in precision oncology. To assess the utility of these multimodal embeddings, we conducted a systematic evaluation across three dimensions: (1) their effectiveness in improving cancer type characterization through modality integration, (2) their generalizability across critical machine learning tasks including classification, retrieval, and survival prediction, and (3) the impact of parameter-efficient fine-tuning (PEFT) on enhancing task-specific performance with minimal computational overhead.

Our analysis leveraged multimodal patient-level data from TCGA, over 11,400 patients and 33 distinct cancer types [14]. Data modalities included clinical text, pathology reports and images, radiology scans, and molecular profiles, including gene expression, DNA methylation, somatic mutations, miRNA expression, and protein expression. The data were curated and preprocessed through the Multimodal Integration of Oncology Data System (MINDS) framework to harmonize heterogeneous sources [13].

Across all experiments, multimodal embeddings consistently outperformed single-modality baselines in cancer-type classification, survival prediction, and similarity-based retrieval. Furthermore, targeted fine-tuning of the embedding models significantly improved performance in oncology-specific tasks, underscoring the value of modular adaptation strategies. The following subsections present detailed evaluations of each dimension, highlighting the role of multimodal integration in improving both clinical insight and model utility.

2.1 Modality-Specific Embedding Generation

To construct a comprehensive and clinically meaningful representation of oncology patients, HONeYBEE generates modality-specific embeddings across four primary data types: clinical text, pathology images, radiology scans, and molecular profiles. Each modality is processed using specialized FMs selected for domain relevance, computational efficiency, and proven downstream performance.

The clinical text was encoded using GatorTron-medium, a 3.9-billion-parameter transformer model pre-trained on large biomedical corpora [22]. GatorTron demonstrated robust performance on medical NLP tasks, such as patient age prediction, and was selected for its scalability and adaptability to diverse clinical narratives. Pathology images were processed with the UNI model, a vision transformer trained on whole-slide pathology images. UNI generates $1,024$-dimensional patch-level embeddings and outperformed larger counterparts in classification and retrieval tasks, offering an efficient and scalable solution for computational pathology[21]. Radiology scans were embedded using RadImageNet, a convolutional neural network pre-trained on over four million medical images across CT, MRI, and PET modalities. RadImageNet yielded superior performance in image retrieval tasks and

was favored for its robustness and domain specificity compared to alternatives such as REMEDIS [23, 25]. Molecular data, including gene expression, DNA methylation, somatic mutations, miRNA, and protein expression, was processed using SeNMo, a self-normalizing deep learning encoder designed for high-dimensional multi-omics data [24]. SeNMo produces biologically meaningful, compressed representations that preserve critical molecular variation while reducing dimensionality.

All embeddings were generated from harmonized oncology data using standardized, modality-specific pipelines within the HoneyBee framework. The resulting patient-level feature vectors, along with associated metadata, have been publicly released via Hugging Face repositories including TCGA (https://huggingface.co/datasets/Lab-Rasool/TCGA), CGCI (https://huggingface.co/datasets/Lab-Rasool/CGCI), FM (https://huggingface.co/datasets/Lab-Rasool/FM), CPTAC (https://huggingface.co/datasets/Lab-Rasool/CPTAC), and TARGET (https://huggingface.co/datasets/Lab-Rasool/TARGET), promoting reproducibility, model benchmarking, and clinical translation.

2.2 Multimodal Integration Improves Cancer Type Characterization

To evaluate the impact of multimodal integration on disease characterization, we examined the clustering behavior of patient embeddings derived from different data modalities. We selected a subset of 1,105 patients from TCGA for whom complete data were available across four modalities: clinical text, pathology reports, radiology images, and molecular profiles. This subset included several cancer types, with kidney renal clear cell carcinoma (KIRC, $n = 267$), ovarian cancer (OV, $n = 141$), and breast cancer (BRCA, $n = 139$) as the largest cohorts.

We visualized the high-dimensional embedding spaces using t-distributed stochastic neighbor embedding (t-SNE) to assess how effectively each modality distinguished cancer types (Fig. 2A-D). Clinical embeddings showed relatively strong separation of certain cancers, particularly KIRC and OV. Molecular embeddings displayed a structured but overlapping manifold. Pathology and radiology embeddings yielded more diffuse clusters, reflecting weaker cancer-type separation in isolation. In contrast, the multimodal embedding space (Fig. 2E), formed by concatenating embeddings across all four modalities, yielded notably improved separation. The integrated embeddings produced compact, well-defined clusters with sharper boundaries between cancer types, most prominently for KIRC, BRCA, and OV. These findings underscore the synergistic value of combining complementary data modalities for enhanced phenotypic differentiation.

To quantify these observations, we computed silhouette scores and normalized mutual information (NMI) for each modality. Clinical embeddings achieved the highest silhouette score among single modalities (0.0186), but multimodal embeddings outperformed all individual modalities in overall cluster structure. The multimodal NMI score (0.4908) exceeded that of clinical (0.4391), molecular (0.2861), pathology (0.2882), and radiology (0.1308) embeddings, representing an 11.8% improvement over the best single modality and a 275% increase over radiology.

The greatest gains were observed for KIRC. In the radiology-only space, KIRC samples overlapped significantly with other cancer types (mean inter-cluster distance = 153.29). In the multimodal space, KIRC formed a distinct cluster with improved cohesion (intra-cluster distance = 36.22) and greater separation (inter-cluster distance = 168.74), yielding a 10.1% increase in inter-cluster distance.

These results highlight the value of multimodal representation learning in preserving biologically and clinically meaningful structures. By integrating heterogeneous data types, the HONeYBEE framework enables more accurate and interpretable patient-level stratification, thus laying the foundation for improved diagnosis, prognosis, and treatment selection in oncology.

2.2.1 Survival Prediction with Multimodal Embeddings

To evaluate the clinical utility of HONeYBEE's multimodal embeddings, we assessed their ability to predict patient survival outcomes. Using a cohort of 1,105 patients with complete data across clinical, pathology, radiology, and molecular modalities, we trained three widely used survival models: Cox proportional hazards (CoxPH), Random Survival Forests (RSF), and DeepSurv [31]. Each model was trained using both single-modality embeddings and concatenated multimodal embeddings. Predictive performance was measured using the concordance index (C-index), which quantifies the agreement between predicted risk scores and observed survival times.

Across all models, multimodal embeddings consistently outperformed those derived from any single modality (Fig. 2F). Specifically, the multimodal embeddings achieved the highest C-index scores: 0.6587 (CoxPH), 0.6252 (RSF), and 0.6621 (DeepSurv). Among individual modalities, clinical embeddings performed best for CoxPH (0.6368) and RSF (0.6235), while pathology embeddings yielded the second-best results for DeepSurv (0.6337). Radiology embeddings consistently underperformed across all models, with C-indices as low as 0.5209 for DeepSurv.

To determine whether these performance differences were statistically significant, we conducted bootstrap analysis with 1,000 stratified resamples. Multimodal embeddings maintained consistently superior performance across all models: CoxPH (95% CI: $0.5931 - 0.7211$), RSF (95% CI: $0.5511 - 0.6994$), and DeepSurv (95% CI: $0.5860 - 0.7130$). Notably, multimodal embeddings outperformed radiology embeddings in DeepSurv with statistical significance (C-index difference = 0.1330, 95% CI: $0.0314 - 0.2346$, $p = 0.01$). While multimodal embeddings also outperformed other single modalities, those improvements did not reach statistical significance ($p > 0.05$). Five-fold cross-validation further confirmed the robustness of these results. The multimodal approach maintained superior average C-index (mean $\pm$ SD) values (CoxPH = 0.5910 ± 0.0542; RSF = 0.5909 ± 0.0504), indicating stable performance across resampling and model variations.

2.3 Embedding Evaluation on Downstream Tasks

To assess the generalizability and practical utility of HONeYBEE embeddings, we evaluated their performance across two critical downstream tasks: cancer type classification and similarity-based retrieval. These benchmarks were conducted using the

full TCGA dataset of 11,478 patients, expanding beyond the subset used in earlier analyses. Unlike prior experiments that required complete data across all modalities, this evaluation leveraged all available samples for each modality independently. This design reflects real-world clinical scenarios where patients may have incomplete or variable diagnostic data based on their clinical presentation, disease stage, or treatment setting.

The dataset encompassed 33 cancer types and exhibited varying levels of modality availability: 11,428 patients with clinical text, 13,804 with molecular data, 11,195 with pathology reports, and 11,229 with radiological images. This heterogeneity allowed us to evaluate the robustness of modality-specific embeddings under practical constraints and assess their effectiveness when deployed independently or in combination.

By benchmarking performance across diverse tasks and modalities, we aimed to determine whether the embeddings generated by HONeYBEE retain discriminative value beyond survival prediction, specifically in settings that demand accurate patient classification or efficient retrieval of similar cases for reference, decision support, or cohort discovery.

2.3.1 Cancer Type Classification

To assess the effectiveness of HONeYBEE embeddings for cancer-type classification, we trained a random forest classifier using modality-specific embeddings from the full TCGA dataset. Patient embeddings were generated independently for clinical notes, pathology reports, radiology images, and molecular profiles. Each dataset was randomly split into training (80%) and test (20%) sets and a random forest model with 100 estimators was trained on the training embeddings to predict cancer types.

As shown in 3A, classification performance varied substantially across modalities. Embeddings from clinical text achieved the highest accuracy (90.21% $\pm$ 0.53%), substantially outperforming embeddings derived from radiology (57.62% $\pm$ 0.74%), molecular data (56.73% $\pm$ 0.79%), and pathology reports (49.32% $\pm$ 0.72%). The confusion matrix for the clinical embedding classifier Fig. 3B) illustrates the model's strong discriminative ability across a broad range of cancer types, with minimal overlap in misclassification patterns.

To ensure robustness and minimize bias from data partitioning, we repeated the experiment 10 times using different random seeds. For each modality, we computed the mean accuracy and standard deviation across these runs to produce performance estimates with associated uncertainty. These results highlight the strong discriminative capacity of clinical embeddings and demonstrate that even without multimodal integration, modality-specific representations can provide valuable signals for large-scale cancer-type classification.

2.3.2 Similarity-Based Retrieval

To further evaluate the semantic quality of HONeYBEE embeddings, we conducted a similarity-based retrieval benchmark using FAISS, a library for efficient similarity

search and clustering of dense vectors [26]. For each modality, we computed precision@*k*, a metric that quantifies the proportion of patients with the same cancer type among the k nearest neighbors in embedding space. High precision@*k* values indicate that the embeddings effectively capture underlying phenotypic similarity relevant to cancer classification and cohort matching.

As shown in 3C, retrieval performance varied notably across modalities. Clinical embeddings achieved the highest precision@*10* (80.30% $\pm$ 26.44%), followed by pathology (46.10% $\pm$ 34.81%), radiology (35.90% $\pm$ 30.73%), and molecular embeddings (22.40% $\pm$ 23.67%). These results mirror trends observed in the classification task and highlight the discriminative strength of clinical text-derived embeddings.

To ensure robustness, we conducted 100 independent retrieval trials per modality, each using a randomly selected patient embedding as the query. The resulting distribution of precision scores, visualized in 3D, reveals that clinical embeddings not only perform better on average but also exhibit greater consistency across different cancer types. In contrast, the broader variance observed in other modalities reflects less reliable retrieval performance, suggesting weaker clustering and noisier embedding spaces.

Together, these findings reinforce the value of clinical embeddings for patient similarity tasks and suggest that HONeYBEE's modality-specific and multimodal representations can support scalable, retrieval-based applications such as clinical decision support and cohort discovery.

2.4 Parameter-Efficient Fine-Tuning (PEFT) Enhances Embedding Quality

To further enhance HONeYBEE's utility for domain-specific applications, we investigated the impact of PEFT, a lightweight fine-tuning strategy that has gained traction in recent large language model research, on the quality of clinical text embeddings [32, 33]. This analysis focused on the clinical modality, which demonstrated the strongest performance in our classification and retrieval benchmarks. We evaluated two widely used language models—GatorTron-medium and BERT-base—and compared their pre-trained and fine-tuned variants using downstream cancer-type classification tasks [22, 34].

To visualize the embedding structure, we used t-SNE to project the high-dimensional embeddings into two dimensions. As shown in Fig. 3, the fine-tuned versions of both GatorTron and BERT produced more distinct and separable clusters by TCGA cancer type, as indicated by clearer boundaries between colored clusters representing project IDs. In contrast, the pre-trained embeddings showed greater overlap and less structure, suggesting that fine-tuning enables the models to better capture task-relevant features in oncology narratives.

We employed an adapter-based PEFT approach in which low-rank adapter layers were inserted into the query, key, and value projection layers of the transformer architecture. Only these lightweight adapter modules were updated during fine-tuning, while the core parameters of the pre-trained model remained frozen.

This strategy drastically reduces the number of trainable parameters while retaining strong performance, illustrating HONeYBEE's compatibility with efficient and scalable adaptation methods.

To visualize the embedding structure, we used t-SNE to project the high-dimensional embeddings into two dimensions. As shown in Fig. 4, the fine-tuned versions of both GatorTron and BERT produced more distinct and separable clusters by TCGA cancer type, as indicated by clearer boundaries between colored clusters representing project IDs. In contrast, the pre-trained embeddings showed greater overlap and less structure, suggesting that fine-tuning enables the models to better capture task-relevant features in oncology narratives.

3 Discussion

In this study, we introduced HONeYBEE, a modular, open-source framework designed to accelerate the development of multimodal AI in oncology by generating high-quality, unified representations from diverse biomedical data types. By leveraging domain-specific FMs [6] for clinical text [22], pathology images [21], radiology scans [25], and molecular profiles [24], HONeYBEE produces patient-level embeddings suitable for a wide range of downstream tasks. Using data from over 11,400 patients in TCGA, we demonstrated that multimodal embeddings consistently outperformed single-modality baselines across clustering, classification, survival prediction, and retrieval tasks. These findings highlight the power of integrated representation learning for multimodal oncology and establish HONeYBEE as a flexible and scalable platform for building generalizable AI solutions in cancer research.

The performance improvements observed across tasks are biologically and clinically grounded. Each data modality captures distinct, complementary aspects of cancer: clinical narratives reflect the patient history and comorbidities, molecular profiles reveal genomic drivers and expression patterns, and imaging modalities, both radiology and pathology, offer spatial, morphological, and phenotypic information. By unifying these signals into a shared embedding space, HONeYBEE enables more comprehensive modeling of the disease state [35–38]. This multimodal integration enhanced cancer-type clustering (e.g., a 10.1% improvement in inter-cluster distance for KIRC) and yielded higher normalized mutual information (NMI $= 0.4908$), an 11.8% gain over the best single modality. Similarly, across all survival models, multimodal embeddings achieved the highest concordance indices, demonstrating their clinical utility for outcome prediction and stratification.

In addition to integrating multiple modalities, HONeYBEE supports the fine-tuning of FMs for domain-specific tasks using parameter-efficient approaches. Our results demonstrate that this lightweight adaptation strategy yields significant performance gains: for instance, fine-tuning GatorTron with adapters improved cancer-type classification accuracy by 14.8% compared to the pretrained baseline [22]. This approach requires minimal computational overhead, making it feasible for use in settings with limited resources. By supporting plug-and-play fine-tuning for each data

type, HONeYBEE enables researchers to tailor general-purpose models to oncology-specific applications without retraining AI models from scratch, thereby lowering both technical and infrastructural barriers to effective AI model customization.

Compared to existing tools, HONeYBEE offers a unified and extensible framework for multimodal oncology data integration. While domain-specific platforms such as OHDSI [39], PyDicom [40], OpenSlide [41], and TIAToolbox [42] focus on specific modalities and require researchers to manually coordinate data formats, interfaces, and metadata across systems. This fragmentation creates substantial technical overhead, particularly when harmonizing cross-modal metadata or managing inconsistent APIs. HONeYBEE resolves these challenges by abstracting complex processing pipelines through a unified API and modular encoder system [13]. It also directly includes pretrained FMs across all modalities, reducing the setup burden typically associated with deep learning in oncology. Unlike plugin-based systems that support model integration only at the surface level, HONeYBEE enables full representation learning and embedding generation with minimal configuration. This positions it as a more complete solution for developing, benchmarking, and deploying multimodal AI models in oncology.

HONeYBEE's ability to generate shared, high-quality representations from multimodal oncology data has broad implications for both research and clinical applications. In research contexts, these embeddings can accelerate biomarker discovery, facilitate cross-modal hypothesis generation, and serve as compact, task-agnostic features for downstream machine learning and AI models [6]. In clinical settings, embedding-based workflows enable use cases such as cohort identification, patient similarity search, and decision support—particularly valuable in complex or rare disease presentations. Our retrieval experiments illustrate the feasibility of these approaches, with clinically relevant matches identified using vector-based similarity search. Notably, because embeddings can be computed locally at each site using the HONeYBEE pipeline, federated learning becomes a natural extension [43]. This enables multi-institutional model development without the need to share raw data, reducing computational burden and minimizing risks of data leakage. By releasing all processed embeddings and metadata through the Hugging Face Datasets Hub, HONeYBEE makes these capabilities immediately accessible, while integration with vector databases like FAISS and Annoy [26, 27] supports scalable deployment in retrieval-augmented generation (RAG), clustering, or risk stratification pipelines.

HONeYBEE's core strength lies in its unified and extensible architecture that simplifies the integration of multimodal oncology data. Unlike traditional toolkits that require managing fragmented pipelines, HONeYBEE provides standardized, modality-specific preprocessing and embedding generation using pretrained FMs. Its plug-and-play design allows researchers to swap or fine-tune encoders (embedding extractors) without modifying the underlying infrastructure, supporting rapid experimentation and benchmarking across clinical text, radiology, pathology, and molecular data. The framework is designed for scalability; that is, embeddings are precomputed and stored in a consistent format, enabling downstream applications such as classification, survival modeling, and retrieval to be run with minimal overhead. Critically, HONeYBEE supports unimodal and partially available datasets,

making it robust to real-world clinical variability. All processed embeddings and models are publicly released via the Hugging Face Datasets Hub and are compatible with vector search frameworks like FAISS and Annoy [26, 27], supporting real-time similarity search and retrieval-augmented applications. Together, these strengths make HONeYBEE a practical and reproducible foundation for developing, evaluating, and deploying multimodal AI models in precision oncology.

While HONeYBEE demonstrates strong performance across multiple benchmarks, several limitations should be acknowledged. Our evaluation was conducted on TCGA, a research-focused dataset that may not fully reflect the complexity of real-world clinical settings. Although HONeYBEE is compatible with other datasets, including those from CRDC repositories, the generalizability of embeddings should be further validated across diverse institutions and populations. Additionally, while our current analyses focus on cancer classification, clustering, and survival prediction, future work could expand to longitudinal modeling, treatment response prediction, and interpretability of learned embeddings. Continued development will also explore integrating additional modalities, such as clinical time series and functional imaging, and leveraging task-specific fine-tuning to further enhance clinical utility.

4 Methods

HONeYBEE is a modular, open-source framework designed to enable scalable, reproducible, and standardized preprocessing of multimodal oncology data for downstream AI application development, machine learning, and representation learning tasks. The framework provides dedicated processing pipelines for four primary biomedical data modalities: whole slide histopathology images, radiology imaging, molecular profiles, and clinical text. Each modality-specific pipeline incorporates state-of-the-art preprocessing techniques tailored to the data type and is designed to generate high-quality, fixed-length data representations (embeddings) using domain-specific FMs. These embeddings are stored in a structured and accessible format to support classification, retrieval, survival analysis, and other downstream applications.

The HONeYBEE architecture supports the integration of both public and institutional datasets and emphasizes interoperability through standardized APIs and metadata handling. Pretrained and custom models are supported via Hugging Face, and outputs are organized for compatibility with modern AI pipelines. Fig. 1 provides an overview of the framework's architecture, highlighting the sequence of data ingestion, preprocessing, embedding generation, and downstream application.

4.1 Pathology Image Processing

Whole Slide Images (WSIs) present unique computational challenges due to their extreme size, multi-resolution pyramid structure, and vendor-specific file formats. HONeYBEE addresses these challenges through an integrated processing pipeline specifically designed for computational pathology workflows. Our implementation

leverages GPU acceleration to efficiently handle WSI loading, preprocessing, tissue detection, normalization, and feature extraction, as illustrated in Fig. 5A.

4.1.1 WSI Loading and Data Management

HONeYBEE utilizes CuImage [44] for efficient loading and handling of WSIs, supporting multiple vendor-specific formats, including Aperio SVS, Philips TIFF, and generic tiled multi-resolution RGB TIFF files. CuImage provides GPU-accelerated I/O capabilities through CUDA [44], enabling fast loading of specific regions of interest while managing memory constraints inherent in working with gigapixel-scale images. The library supports various compression schemes, including JPEG, JPEG2000, LZW, and Deflate, providing flexibility in handling different WSI formats while maintaining processing speed.

The framework implements metadata handling to preserve critical image properties and structure. This includes maintaining physical spacing information in micrometers, coordinate system specifications, and vendor-specific metadata such as objective magnification and microns-per-pixel ratios. For formats like Aperio SVS, detailed scan parameters, and ICC color profiles are preserved, allowing for accurate spatial measurements and color reproduction in downstream analysis tasks.

HONeYBEE's multi-resolution pyramid management allows efficient access to WSI data at different magnification levels. The system maintains resolution information, including dimensions, downsampling factors, and tile specifications for each level. This enables memory-efficient implementation of various analysis strategies, from rapid whole-slide thumbnail generation to high-resolution region analysis. The framework implements on-demand loading of specific regions at desired resolution levels, supporting both CPU and GPU memory targets through device specification. When available, the system can leverage NVIDIA GPUDirect Storage for direct data transfer from storage to GPU memory, further reducing I/O bottlenecks [45].

Memory efficiency is achieved through intelligent region reading strategies that maintain coordinate consistency across resolution levels. Users can specify locations in base-resolution coordinates while accessing data at any pyramid level, simplifying the development of multi-scale analysis pipelines. The implementation includes automatic memory management through caching strategies, optimizing performance while preventing memory overflow when working with multiple large WSIs simultaneously.

4.1.2 Tissue Detection and Segmentation

Effective tissue detection and patch extraction are essential preprocessing steps for computational pathology analysis. As shown in Fig. 5, HONeYBEE implements two distinct approaches for tissue detection: a classical threshold-based method using Otsu's algorithm [46] and a deep learning-based approach utilizing a pretrained DenseNet Slidl model [47].

The Otsu-based method operates on the assumption that tissue regions exhibit distinct intensity characteristics compared to the background. After converting the image to grayscale, the algorithm automatically determines an optimal threshold

that maximizes the variance between tissue and background classes. As shown in Fig. 5, this approach generates a gradient magnitude map and binary tissue mask, followed by grid-based patch extraction. While computationally efficient, this method works best with images that have good contrast between tissue and background regions.

The deep learning-based approach employs a pretrained DenseNet model that classifies image patches into three categories: tissue, background, and noise. This model, originally trained by Slidl [47], provides more robust tissue detection and can identify artifacts such as pen markings that should be excluded from analysis. Fig. 5 demonstrates the model's capability to distinguish between actual tissue regions, background, and artifacts like pen markings, enabling more precise patch extraction. However, even this sophisticated approach can face challenges with certain types of slides.

Following tissue detection, HONeYBEE implements an intelligent patch extraction strategy that maximizes the coverage of relevant tissue regions while minimizing computational overhead. The framework divides detected tissue regions into standardized patches with configurable size parameters (typically 256×256 or 512×512 pixels). Patches are filtered based on tissue content thresholds to ensure meaningful analysis, and their coordinates are maintained relative to the original WSI for spatial reference. The extracted patches can be efficiently loaded into memory as needed, enabling scalable processing of large whole slide images while maintaining spatial context for downstream analysis tasks.

4.1.3 Stain Normalization

Histopathological image appearance varies significantly due to differences in tissue preparation, scanner specifications, and laboratory protocols. These variations negatively impact both visual interpretation and computational analysis. HONeYBEE addresses this challenge by implementing three state-of-the-art stain normalization methods, as illustrated in Fig. 5B:

1. Reinhard Normalization: This technique performs color normalization by matching statistical properties of the source image to a target image in LAB color space. While computationally efficient, it does not explicitly account for specific hematoxylin and eosin stain characteristics [48].
2. Macenko Normalization: This method estimates stain vectors through singular value decomposition in optical density space, enabling separation of hematoxylin and eosin contributions before normalization. This approach better preserves relative staining patterns while standardizing overall appearance [49].
3. Vahadane Normalization: Using sparse non-negative matrix factorization, this technique decomposes images into stain density maps, offering robust stain separation even with varying tissue characteristics or additional stain presence [50].

All three normalization methods in HONeYBEE are GPU-accelerated, providing significant performance improvements over CPU-based alternatives (average 8-12×

speedup on typical hardware configurations). Our implementation maintains consistency in normalization results across different image scales through automated parameter adjustment and handles memory management for large WSIs through efficient tiling strategies.

4.1.4 Stain Separation

Beyond normalization, quantitative tissue analysis often requires the isolation of individual stain components. HONeYBEE implements color deconvolution methods to decompose H&E stained images into constituent components based on the Beer-Lambert law of light absorption, as shown in Fig. 5A.

Our implementation converts RGB images to Hematoxylin-Eosin-DAB (HED) color space through calibrated deconvolution matrices that account for the characteristic absorption spectra of each stain. This transformation enables separate analysis of (i) the hematoxylin component, Highlighting nuclear structures; (ii) the eosin component, Emphasizing cytoplasmic and stromal features; and (iii) the DAB component, Isolating immunohistochemical signals when present.

The framework provides bidirectional conversion between RGB and HED color spaces, allowing researchers to analyze separated stain channels and reconstruct normalized images with modified stain characteristics. These operations are GPU-accelerated and process large image regions through automated tiling strategies to maintain memory efficiency.

4.2 Radiology Imaging Processing

Medical imaging plays a central role in oncology for diagnosis, staging, treatment planning, and response assessment. HONeYBEE implements comprehensive processing pipelines for major radiological modalities, including Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET). Each modality presents unique characteristics that require specialized preprocessing techniques while maintaining standardized outputs for downstream multimodal integration.

4.2.1 Radiological Data Management

HONeYBEE provides support for standard medical imaging formats through integration with industry-standard Digital Imaging and Communications in Medicine (DICOM) protocols and specialized neuroimaging formats such as Neuroimaging Informatics Technology Initiative (NIfTI) [51, 52]. Our implementation preserves metadata that influences image interpretation and quantitative analysis. For CT imaging, this includes Hounsfield unit (HU) calibration parameters, exposure settings (kVp, mAs, dose), slice thickness, reconstruction kernel information, and scanner-specific calibration factors. MRI sequence metadata preservation encompasses pulse sequence parameters (TR/TE values), field strength, coil configuration, contrast administration details, and acquisition plane and timing information. PET

studies metadata retention includes radiopharmaceutical type and injection parameters, standardized uptake value (SUV) conversion factors, attenuation correction methods, and decay correction timing information [53].

The framework implements memory-efficient loading strategies through lazy evaluation, allowing researchers to work with large volumetric datasets without excessive memory requirements [54]. This is particularly important for multi-sequence MRI studies or time-series imaging data that can exceed several gigabytes per patient. The implementation supports both single-file and multi-file DICOM series, automatically handling series organization, temporal sequencing, and multi-frame images [55, 56].

4.2.2 Anatomical Segmentation and Masking

Isolating relevant anatomical structures and regions of interest forms a critical preprocessing step for targeted analysis of oncological imaging. HONeYBEE incorporates both classical algorithms and deep learning-based approaches for automated segmentation across different imaging modalities [57, 58]. For CT segmentation, the framework offers automated lung segmentation using threshold-based approaches with morphological refinement, multi-organ segmentation through pre-trained U-Net architectures optimized for thoracic, abdominal, and pelvic regions, automated lung nodule detection using 3D convolutional networks with sensitivity exceeding 85% for nodules larger than 4mm, and bone and soft tissue separation using HU-based thresholding with anatomical constraints [59].

MRI segmentation capabilities include brain extraction tools optimized for different MRI sequences (T1, T2, FLAIR), multi-parametric tumor segmentation integrating information across multiple sequences, specialized sequence-specific segmentation models for contrast-enhanced regions, and atlas-based segmentation for standardized anatomical referencing [60]. For PET-based segmentation, the framework implements metabolic volume delineation using fixed and adaptive SUV thresholding, gradient-based segmentation for heterogeneous uptake regions, joint PET/CT segmentation leveraging complementary information from both modalities, and automated detection of hypermetabolic regions with statistical deviation analysis [58].

All segmentation methods are implemented with GPU acceleration when applicable, providing significant performance improvements over CPU-based alternatives [61, 62]. The framework maintains original image coordinates and transformation matrices, ensuring accurate spatial registration between original images and segmentation masks.

4.2.3 Denoising and Artifact Reduction

Image noise and artifacts can significantly impact both visual interpretation and computational analysis of medical images. HONeYBEE implements multiple denoising strategies optimized for specific imaging modalities and noise characteristics. CT denoising approaches include non-local means filtering with optimized parameters for different dose levels, deep learning-based denoising through convolutional

neural networks trained on paired low-dose/high-dose CT images, edge-preserving bilateral filtering with automatic parameter tuning, and structural fidelity metrics to ensure critical diagnostic features remain uncompromised [63].

MRI denoising implementations feature Rician noise models specifically designed for magnitude MRI data, non-local means filtering adapted for different sequence types, wavelet-based denoising with soft thresholding, and specialized approaches for diffusion-weighted imaging addressing unique noise characteristics [64]. PET denoising capabilities encompass sinogram-space denoising for raw PET data when available, image-space denoising with SUV preservation guarantees, temporal filtering for dynamic PET acquisitions, and joint PET/CT denoising leveraging structural information from CT [65].

Our quantitative evaluations demonstrate that these denoising methods achieve an average 40-60% reduction in noise levels while maintaining structural similarity indices (SSIM) above 0.92 compared to reference images. Performance metrics for each method are continuously benchmarked across public datasets to ensure optimal parameter selection [66].

4.2.4 Spatial Standardization and Resampling

Variability in spatial resolution and slice thickness across different scanners and protocols necessitates standardization for consistent analysis [57][61]. HONeYBEE implements comprehensive resampling capabilities with modality-specific optimization. Resolution standardization includes isotropic voxel resampling with configurable target dimensions, support for both downsampling and upsampling with appropriate filtering, preservation of spatial relationships and anatomical proportions, and multiple interpolation methods including nearest neighbor, linear, B-spline, and Lanczos [54, 67].

CT-specific resampling techniques focus on Hounsfield unit preservation during interpolation, thin-slice to thick-slice conversion with appropriate averaging, axial to coronal/sagittal reformatting with consistent spatial referencing, and partial volume effect compensation for quantitative analysis [68]. MRI-specific resampling addresses multi-sequence alignment to establish spatial correspondence, motion correction between sequences acquired at different timepoints, distortion correction for echo-planar imaging sequences, and field inhomogeneity compensation for accurate spatial registration [69]. PET-specific resampling includes SUV-preserving interpolation methods, PET/CT alignment with motion compensation, resolution recovery for quantitative accuracy, and partial volume effect correction for small lesions [70].

Our implementation tracks and propagates spatial transformation matrices through all processing steps, enabling accurate coordinate mapping between original and processed images. This spatial consistency is critical for multimodal integration and region-of-interest analysis [71].

4.2.5 Intensity Normalization

Intensity standardization addresses variability in signal intensities across different scanners, protocols, and patient characteristics. HONeYBEE implements modality-specific normalization techniques that preserve quantitative relationships while enabling consistent analysis. CT intensity normalization includes Hounsfield unit verification and calibration, predefined window/level settings for different anatomical regions (lung, soft tissue, bone), adaptive histogram equalization for enhanced contrast in selected regions, and intensity outlier handling for metal artifacts and beam hardening [57, 72].

MRI intensity normalization encompasses Z-score normalization on a per-sequence basis, histogram matching to standardized references, bias field correction for signal inhomogeneity, tissue-specific intensity normalization using segmentation priors, and cross-sequence intensity harmonization for multi-parametric analysis [73, 74]. PET intensity standardization features SUV calculation with support for different normalization factors (body weight, lean body mass, body surface area), time-decay correction for quantitative accuracy, scanner-specific calibration factor application, and cross-scanner harmonization for multi-center studies [75].

The framework maintains both raw and normalized versions of the imaging data, allowing researchers to access original quantitative values when needed while providing standardized inputs for machine learning algorithms and AI model development [76].

4.3 Molecular Data Processing

Molecular data refers to biological information derived from molecules within cells, including DNA, RNA, proteins, and metabolites. Molecular data is widely used in biomedical research, personalized medicine, and multi-omics analyses to understand disease mechanisms, predict outcomes, and develop targeted therapies. It provides insights into cellular functions, genetic variations, and biochemical processes. Common types of molecular data include genomic data (DNA sequences, mutations, variations), transcriptomic data (RNA expressions), proteomic data (protein expressions, modifications, interactions), epigenomic data (DNA methylation, histone modifications), and metabolomic data. Although rich in information, using molecular data for analysis presents several challenges across multiple stages, from data acquisition to interpretation. These challenges include varying complexity across modalities, '*big P-small N*' (very high dimensionality features and low sample size), uncommon standardization, integration, missingness, experimental noise, and high computational and storage demands [77]. Addressing these challenges requires robust statistical methods, computational advancements, standardized protocols, and interdisciplinary collaboration. HONeYBEE framework incorporates minimum essential pre-processing steps to ensure the integrity of the data while minimizing the complexity in handling the heterogeneous molecular data modalities. The molecular data processing pipeline of HONeYBEE involves five data modalities including DNA methylation, gene expression, protein expression, DNA mutation,

and miRNA expression data. This pipeline ensures data harmonization, dimensionality reduction, and the generation of contextually enriched embeddings for downstream machine learning and AI applications.

4.3.1 Data Acquisition and Preprocessing

HONeYBEE utilizes publicly available datasets from repositories such as TCGA and University of California, Santa Cruz (UCSC) Xena portal to curate molecular data. The raw datasets undergo a series of preprocessing steps as given below:

1. **Normalization and Scaling:** Data modalities are normalized using modality-specific techniques, such as calculating the beta value for DNA methylation and log transformation for RNA sequencing data, ensuring consistency across samples. All the data modalities were normalized and scaled individually as part of SeNMo preprocessing pipeline [24].
2. **Drop Constant, Duplicate, and Colinear Features:** Dropping constant, duplicate, and colinear features is an essential preprocessing step in molecular data analysis to enhance model performance and efficiency. Constant and quasi-constant features with 99.8% same values have no discriminative power and were removed. Duplicate features, often resulting from an up-regulation pathway of genes or noisy data, can bias models and were dropped to keep one copy of the features. Colinear features, identified by low variance, can have low information content and may be dropped to attend to more informative features. We removed these low variance ($\approx$ 0.25) features to simplify the dataset, reduce dimensionality, and improve the robustness and interpretability of downstream machine learning models and AI applications.
3. **Remove Low-expression Genes:** To reduce the high number of features in gene expression data with Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values ranging from 0 to 12, low-expression genes were removed. Since high expression levels are more biologically relevant, only genes with expression values above 7 (equivalent to 127 FPKM) were kept. This filtering reduced the dataset genes with expression values ranging from 7 to 12.
4. **Remove NaNs and Handling Missing Data:** The removal of NaNs from molecular data is a critical preprocessing step to counter experimental errors, incomplete measurements, or data acquisition limitations. We removed the features having NaNs across all samples. Following the removal of NaNs, we imputed the missing values using the mean of samples within each modality.

4.3.2 Feature Selection and Integration

The preprocessing pipeline produces a standardized feature set for each molecular modality. These feature sets are then unified into a multimodal feature matrix to facilitate pan-cancer analyses. Key steps include:

1. **Union of Features:** Features are aggregated across cancer types within each modality, resulting in all available features for all patients in the data. To create full-length features for all samples, zero imputation is employed, which has been shown to minimize noise and enhance model performance [78, 79].

2. **Elimination of Redundancies:** Duplicate and collinear features are identified and removed to optimize the feature matrix.
3. **Multi-Modal Integration:** Features from all modalities are combined into a single matrix using union operations, resulting in a comprehensive multimodal representation having all five molecular features across all samples for all 33 cancer types from TCGA [14].
4. **Clinical Features Integration:** SeNMo framework employs patient covariates that have been found to be informative in prognosis and improve treatment responses [24]. These covariates were integrated with the molecular data as features including age, gender, race, and cancer stage of patients. The categorical features were converted to numerical values for integration. The final feature vector for each patient consisted of five sets of molecular features (DNA methylation, gene expression, protein expression, DNA mutation, miRNA expression) and four clinical features (age, gender, race, cancer stage).

4.4 Clinical Data Processing

Clinical data in oncology presents unique challenges due to its heterogeneous nature, spanning structured data elements (laboratory values, vital signs, medications) and unstructured narrative text (clinical notes, radiology reports, pathology reports). HONeYBEE implements processing pipelines to handle both data types while maintaining semantic relationships for clinical analysis.

4.4.1 Text Extraction and Document Processing

The extraction of clinical text from various document formats represents a critical preprocessing step in clinical data analysis. HONeYBEE supports multiple input formats including PDF, scanned images, and electronic health record (EHR) exports. The framework implements a multi-stage processing pipeline with modality-specific optimizations: For digitized documents, the framework utilizes a multi-stage optical character recognition (OCR) pipeline utilizing Tesseract OCR with specialized medical dictionaries to improve recognition accuracy of clinical terminology [80]. The pipeline includes image preprocessing (deskewing, noise reduction, binarization), layout analysis to preserve document structure, and post-processing with medical terminology verification to identify and correct common OCR errors in clinical contexts [81].

For structured EHR data in tabular formats, HONeYBEE implements a conversion pipeline that transforms discrete data elements into standardized key-value pair representations. This process preserves the semantic relationships inherent in the original data structure while enabling unified processing alongside unstructured text [82]. Laboratory results, vital signs, medication orders, and other structured elements are converted into a consistent text format (e.g., "blood_pressure: 120/80 mmHg", "white_blood_cells: 7.2 K/µL") that can be processed by language models while maintaining clinical meaning and temporal relationships [82].

Document structure analysis is performed to identify and preserve the hierarchical organization of clinical reports, maintaining relationships between sections,

subsections, and individual data elements. The framework implements specialized handlers for common clinical document types, including operative reports, pathology reports, and consultation notes, each with specific rules for extracting structured information from semi-structured text [80]. The framework implements quality control measures to flag problematic documents and provides confidence scores for extracted text segments, enabling downstream processes to account for potential extraction errors [81].

4.4.2 Tokenization and Language Model Integration

Following successful text extraction, HONeYBEE implements tokenization through integration with the Hugging Face Transformers library, enabling compatibility with state-of-the-art clinical language models. The framework supports multiple pretrained tokenizers optimized for biomedical text, including BioClinicalBERT, PubMedBERT, GatorTron, and ClinicalT5, ensuring appropriate handling of domain-specific vocabulary and clinical narratives [83, 84]. The tokenization pipeline processes extracted clinical text through several stages:

1. Preliminary cleaning to standardize whitespace, handle special characters, and normalize common clinical abbreviations
2. Text segmentation into appropriate units (sentences, paragraphs, sections) based on document structure
3. Subword tokenization using model-specific strategies (WordPiece, Byte-Pair Encoding, SentencePiece)
4. Special token handling for model-specific requirements (e.g., [CLS], [SEP], [MASK])
5. Sequence length management with configurable strategies for truncation and sliding windows

To accommodate long clinical documents that exceed typical model input token sequence lengths (generally 512-1024 tokens), HONeYBEE implements multiple strategies: (i) sliding window tokenization with configurable overlap percentages (typically 10-20%), (ii) hierarchical document segmentation preserving section boundaries, (iii) important segment identification using clinical term density heuristics, and (iv) document summarization for extremely long texts using extractive methods.

The framework's tokenization implementation is optimized for batch processing, enabling efficient handling of large clinical document collections while maintaining memory efficiency through dynamic batch sizing based on available computational resources [85].

4.4.3 Clinical Entity Recognition and Normalization

Identification and standardization of clinical concepts are essential for structured analysis of medical narratives. HONeYBEE implements a comprehensive clinical entity recognition pipeline utilizing both rule-based and deep learning approaches[84, 86].

Our implementation includes integration with medical ontologies and terminologies including SNOMED CT, RxNorm, LOINC, and ICD-O-3, enabling normalization of extracted clinical entities to standardized concept identifiers. This normalization process addresses challenges such as synonymy (multiple terms for the same concept), abbreviation expansion based on context, and disambiguation of terms with multiple potential meanings[86].

For cancer-specific entity recognition, the framework implements specialized models for extracting and normalizing[87]: (i) tumor characteristics (histology, grade, size, invasiveness), (ii) staging information (TNM parameters, stage groupings), (iii) biomarker status (receptor expression, molecular alterations), (iv) treatment details (surgical procedures, chemotherapy regimens, radiation protocols), and (iv) response assessment (RECIST criteria, clinical response categories).

Temporal information extraction identifies and normalizes dates, durations, and relative time expressions within clinical narratives. This enables the construction of longitudinal patient timelines with precise sequencing of diagnostic, treatment, and follow-up events critical for oncology research[88, 89].

The extracted and normalized entities maintain provenance links to their source documents and original text spans, facilitating verification and quality assessment of automated extractions. Performance metrics for entity recognition are continuously benchmarked across public datasets[84, 86].

4.5 Embedding Generation and Storage

4.5.1 Generation of Embeddings

Each preprocessed data sample is passed through pre-trained FM, which produces a fixed-length embedding vector that varies in shape depending on the model architecture. HONeYBEE utilizes GPU acceleration and distributed computing, when available, to efficiently generate embeddings for large-scale datasets. The generated embeddings, along with associated metadata, are stored in a structured format to facilitate downstream tasks such as similarity search, clustering, and AI and machine learning model training. HONeYBEE employs efficient data compression and indexing techniques to optimize the storage and retrieval of high-dimensional embeddings. As an example, the processed molecular data is passed through the SeNMo model, a self-normalizing deep learning encoder, to generate latent feature vectors (embeddings). These embeddings capture critical biological patterns and are suitable for tasks such as survival analysis, cancer subtype classification, and biomarker discovery.

4.5.2 Data storage and Accessibility

The generated embeddings and tabular data are stored using the Hugging Face datasets library, which provides a standardized interface for data access and processing [90]. The datasets are organized into a structured format, containing the embeddings, metadata, and labels (if available). PyTorch DataLoaders are employed

to efficiently load and iterate over the datasets during model training and evaluation, handling tasks such as batching, shuffling, and parallel processing. Additionally, HONeYBEE datasets can be integrated into vector databases such as Faiss and Annoy [26, 27] to enable fast similarity search, nearest neighbor retrieval, and clustering on the high-dimensional embedding vectors. These databases are optimized for efficient querying and retrieval of embeddings based on similarity metrics such as Euclidean distance or cosine similarity [91]. By deploying vector databases, researchers can quickly identify similar samples, perform data exploration, and retrieve relevant subsets based on embedding similarity, facilitating various downstream tasks such as retrieval augmented generation (RAG) [92]. The structured storage and accessibility components of HONeYBEE ensure that the generated embeddings and associated data are readily available for researchers and practitioners to use in their AI and machine learning pipelines and downstream applications.

5 Conclusion

The HONeYBEE framework addresses a critical need in computational oncology: the scalable, standardized generation of high-quality data representations across clinical, radiological, pathological, and molecular domains. By leveraging domain-specific FMs and providing reproducible pipelines, HONeYBEE enables the creation of unified, multimodal embeddings that consistently outperform single-modality approaches across clustering, survival prediction, and classification tasks. Our evaluation using over 11,400 TCGA patients demonstrates clear advantages of multimodal integration, including improved cancer-type discrimination and a 5–10% gain in survival model concordance indices. Fine-tuning clinical language models further enhanced performance, highlighting the value of parameter-efficient adaptation for domain-specific tasks. Future development will focus on incorporating additional modalities—such as time series and functional imaging—expanding support for specialized FMs, and validating performance across more diverse, real-world patient cohorts. Interpretability and clinical trust remain essential next steps to foster translational adoption. While TCGA offers rich multimodal data, inherent cohort biases underscore the importance of generalizability testing. By releasing all embeddings, models, and processing tools as part of an open-source ecosystem, HONeYBEE lowers barriers to entry and promotes reproducibility, transparency, and collaboration in AI for oncology. Ultimately, the framework lays the groundwork for robust, scalable AI pipelines that support precision oncology at scale.

Declarations

Funding

This research was supported by NSF Awards 2234836 and 2234468 and NAIRR pilot funding.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

This research utilized publicly available data from The Cancer Genome Atlas (TCGA), which has been previously collected with appropriate ethical approvals and patient consent. No additional human subjects were involved in this study, and therefore, no additional ethics approval was required.

Data availability

The datasets generated during the current study are available in the Hugging Face repository, https://huggingface.co/datasets/Lab-Rasool/TCGA. The raw data used in this study was obtained from The Cancer Genome Atlas (TCGA) through the MINDS framework [13]. A comprehensive collection of the datasets and models used in this study is available at https://huggingface.co/collections/Lab-Rasool/honeybee-66b9f15e8b908c4eeeec9cb4.

Code availability

The source code for the HONeYBEE framework is publicly available on GitHub at https://github.com/lab-rasool/HoneyBee under an open-source license. Additionally, the MINDS codebase used for data extraction is available at https://github.com/lab-rasool/MINDS. The HONeYBEE paper preprint is available at https://arxiv.org/abs/2405.07460.

Authors' contributions

AT and AW contributed equally to this work. AT led the development of the multimodal embedding framework, designed and implemented the neural network architectures, and conducted computational experiments. AW led the molecular processing and organization efforts and contributed to the molecular model development. MBS provided clinical expertise, validated the clinical relevance of the findings, and contributed to the interpretation of results. YY provided guidance on algorithm design and statistical analysis. GR conceived and supervised the study, secured funding, and provided overall direction. All authors contributed to the manuscript writing and revision.

References

[1] Hinkson, Izumi V. and Davidsen, Tanja M. and Klemm, Juli D. and Chandramouliswaran, Ishwar and Kerlavage, Anthony R. and Kibbe, Warren A.: A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. Frontiers in Cell and Developmental Biology **5** (2017) https://doi.org/10.3389/fcell.2017.00083

[2] Thangudu, Ratna Rajesh and Rudnick, Paul A. and Holck, Michael and Singhal, Deepak and MacCoss, Michael J. and Edwards, Nathan J. and Ketchum, Karen A. and Kinsinger, Christopher R. and Kim, Erika and Basu, Anand: Abstract LB-242: Proteomic Data Commons: A resource for proteogenomic analysis. Cancer Research **80**(16_Supplement), 242 (2020) https://doi.org/10.1158/1538-7445.AM2020-LB-242

[3] Grossman, Robert L and Heath, Allison P and Ferretti, Vincent and Varmus, Harold E and Lowy, Douglas R and Kibbe, Warren A and Staudt, Louis M: Toward a shared vision for cancer genomic data. New England Journal of Medicine **375**(12), 1109–1112 (2016)

[4] A. Fedorov and W. Longabaugh and D. Pot and D. Clunie and S. Pieper and R. Lewis and H. Aerts and A. Homeyer and M. Herrmann and U. Wagner and T. Pihl and K. Farahani and R. Kikinis: NCI Imaging Data Commons. International Journal of Radiation Oncology*Biology*Physics **111**(3, Supplement), 101 (2021) https://doi.org/10.1016/j.ijrobp.2021.07.495 . 2021 Proceedings of the ASTRO 63rd Annual Meeting

[5] Clark, Kenneth and Vendt, Bruce and Smith, Kirk and Freymann, John and Kirby, Justin and Koppel, Paul and Moore, Stephen and Phillips, Stanley and Maffitt, David and Pringle, Michael and others: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging **26**, 1045–1057 (2013)

[6] Boehm, Kevin M and Khosravi, Pegah and Vanguri, Rami and Gao, Jianjiong and Shah, Sohrab P: Harnessing multimodal data integration to advance precision oncology. Nature Reviews Cancer **22**(2), 114–126 (2022)

[7] Paverd, Hania and Zormpas-Petridis, Konstantinos and Clayton, Hannah and Burge, Sarah and Crispin-Ortuzar, Mireia: Radiology and multi-scale data integration for precision oncology. NPJ Precision Oncology **8**(1), 158 (2024)

[8] Vanguri, Rami S. and Luo, Jia and Aukerman, Andrew T. and Egger, Jacklynn V. and Fong, Christopher J. and Horvat, Natally and Pagano, Andrew and Araujo-Filho, Jose de Arimateia Batista and Geneslaw, Luke and Rizvi, Hira and Sosa, Ramon and Boehm, Kevin M. and Yang, Soo-Ryum and Bodd, Francis M. and Ventura, Katia and Hollmann, Travis J. and Ginsberg, Michelle S. and Gao, Jianjiong and Vanguri, Rami and Hellmann, Matthew D. and Sauter, Jennifer L. and Shah, Sohrab P.: Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. Nature Cancer **3**(10), 1151–1164 (2022) https://doi.org/10.1038/s43018-022-00416-8

[9] Waqas, Asim and Tripathi, Aakash and Ramachandran, Ravi P and Stewart, Paul A and Rasool, Ghulam: Multimodal data integration for oncology in the era of deep neural networks: a review. Frontiers in Artificial Intelligence **7**,

1408843 (2024)

[10] Waqas, Asim and Naveed, Javeria and Shahnawaz, Warda and Asghar, Shoaib and Bui, Marilyn M and Rasool, Ghulam: Digital pathology and multimodal learning on oncology data. BJR— Artificial Intelligence **1**(1), 014 (2024)

[11] Boehm, K.M., Aherne, E.A., Ellenson, L., Nikolovski, I., Alghamdi, M., Vázquez-García, I., Zamarin, D., Long Roche, K., Liu, Y., Patel, D., Aukerman, A., Pasha, A., Rose, D., Selenica, P., Causa Andrieu, P.I., Fong, C., Capanu, M., Reis-Filho, J.S., Vanguri, R., Veeraraghavan, H., Gangai, N., Sosa, R., Leung, S., McPherson, A., Gao, J., Lakhman, Y., Shah, S.P.: Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature Cancer **3**(6), 723–733 (2022) https://doi.org/10.1038/s43018-022-00388-9

[12] Lipkova, Jana and Chen, Richard J and Chen, Bowen and Lu, Ming Y and Barbieri, Matteo and Shao, Daniel and Vaidya, Anurag J and Chen, Chengkuan and Zhuang, Luoting and Williamson, Drew FK and others: Artificial intelligence for multimodal data integration in oncology. Cancer cell **40**(10), 1095–1110 (2022)

[13] Tripathi, Aakash and Waqas, Asim and Venkatesan, Kavya and Yilmaz, Yasin and Rasool, Ghulam: Building Flexible, Scalable, and Machine Learning-Ready Multimodal Oncology Datasets. Sensors **24**(5) (2024) https://doi.org/10.3390/s24051634

[14] Weinstein, John N and Collisson, Eric A and Mills, Gordon B and Shaw, Kenna R and Ozenberger, Brad A and Ellrott, Kyle and Shmulevich, Ilya and Sander, Chris and Stuart, Joshua M: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)

[15] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You,

J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the Opportunities and Risks of Foundation Models (2022). https://arxiv.org/abs/2108.07258

[16] Waqas, Asim and Bui, Marilyn M and Glassy, Eric F and El Naqa, Issam and Borkowski, Piotr and Borkowski, Andrew A and Rasool, Ghulam: Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. Laboratory Investigation, 100255 (2023)

[17] Hartsock, Iryna and Rasool, Ghulam: Vision-language models for medical report generation and visual question answering: a review. Frontiers in Artificial Intelligence **7** (2024) https://doi.org/10.3389/frai.2024.1430984

[18] Steyaert, Sandra and Pizurica, Marija and Nagaraj, Divya and Khandelwal, Priya and Hernandez-Boussard, Tina and Gentles, Andrew J and Gevaert, Olivier: Multimodal data fusion for cancer biomarker discovery with deep learning. Nature machine intelligence **5**(4), 351–362 (2023)

[19] Nicora, Giovanna and Vitali, Francesca and Dagliati, Arianna and Geifman, Nophar and Bellazzi, Riccardo: Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Frontiers in oncology **10**, 1030 (2020)

[20] Cooper, Michael and Ji, Zongliang and Krishnan, Rahul G: Machine learning in computational histopathology: Challenges and opportunities. Genes, Chromosomes and Cancer **62**(9), 540–556 (2023)

[21] Chen, Richard J and Ding, Tong and Lu, Ming Y and Williamson, Drew FK and Jaume, Guillaume and Song, Andrew H and Chen, Bowen and Zhang, Andrew and Shao, Daniel and Shaban, Muhammad and others: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)

[22] Yang, Xi and Chen, Aokun and PourNejatian, Nima and Shin, Hoo Chang and Smith, Kaleb E and Parisien, Christopher and Compas, Colin and Martin, Cheryl and Costa, Anthony B and Flores, Mona G and others: A large language model for electronic health records. NPJ digital medicine **5**(1), 194 (2022)

[23] Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., *et al.*: Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nature Biomedical Engineering **7**(6), 756–779 (2023)

[24] Waqas, Asim and Tripathi, Aakash and Ahmed, Sabeen and Mukund, Ashwin and Farooq, Hamza and Schabath, Matthew B and Stewart, Paul and Naeini, Mia and Rasool, Ghulam: SeNMo: A Self-Normalizing Deep Learning Model for Enhanced Multi-Omics Data Analysis in Oncology. arXiv preprint

arXiv:2405.08226 (2024)

[25] Mei, Xueyan and Liu, Zelong and Robson, Philip M and Marinelli, Brett and Huang, Mingqian and Doshi, Amish and Jacobi, Adam and Cao, Chendi and Link, Katherine E and Yang, Thomas and others: RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), 210315 (2022)

[26] Johnson, Jeff and Douze, Matthijs and Jégou, Hervé: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)

[27] Bernhardsson, Erik: Annoy: Approximate Nearest Neighbors in C++/Python. Version 1.13.0 (2018). https://github.com/spotify/annoy

[28] Soenksen, Luis R and Ma, Yu and Zeng, Cynthia and Boussioux, Leonard and Villalobos Carballo, Kimberly and Na, Liangyuan and Wiberg, Holly M and Li, Michael L and Fuentes, Ignacio and Bertsimas, Dimitris: Integrated multimodal artificial intelligence framework for healthcare applications. NPJ digital medicine **5**(1), 149 (2022)

[29] Pahud de Mortanges, Aurélie and Luo, Haozhe and Shu, Shelley Zixin and Kamath, Amith and Suter, Yannick and Shelan, Mohamed and Pöllinger, Alexander and Reyes, Mauricio: Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. NPJ digital medicine **7**(1), 195 (2024)

[30] Chang, Jee Suk and Kim, Hyunwook and Baek, Eun Sil and Choi, Jeong Eun and Lim, Joon Seok and Kim, Jin Sung and Shin, Sang Joon: Continuous multimodal data supply chain and expandable clinical decision support for oncology. npj Digital Medicine **8**(1), 128 (2025)

[31] Katzman, Jared L and Shaham, Uri and Cloninger, Alexander and Bates, Jonathan and Jiang, Tingting and Kluger, Yuval: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology **18**, 1–12 (2018)

[32] Kim, Jeonghoon and Lee, Jung Hyun and Kim, Sungdong and Park, Joonsuk and Yoo, Kang Min and Kwon, Se Jung and Lee, Dongsoo: Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. Advances in Neural Information Processing Systems **36** (2024)

[33] Dettmers, Tim and Pagnoni, Artidoro and Holtzman, Ari and Zettlemoyer, Luke: Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems **36** (2024)

[34] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers), pp. 4171–4186 (2019)

[35] Azuaje, Francisco: Artificial intelligence for precision oncology: beyond patient stratification. NPJ precision oncology **3**(1), 6 (2019)

[36] Ovchinnikova, Katja and Born, Jannis and Chouvardas, Panagiotis and Rapsomaniki, Marianna and Kruithof-de Julio, Marianna: Overcoming limitations in current measures of drug response may enable AI-driven precision oncology. NPJ precision oncology **8**(1), 95 (2024)

[37] Truhn, Daniel and Eckardt, Jan-Niklas and Ferber, Dyke and Kather, Jakob Nikolas: Large language models and multimodal foundation models for precision oncology. NPJ Precision Oncology **8**(1), 72 (2024)

[38] Fang, Chao and Xu, Dong and Su, Jing and Dry, Jonathan R. and Linghu, Bolan: DeePaN: deep patient graph convolutional network integrating clinico-genomic evidence to stratify lung cancers for immunotherapy. npj Digital Medicine **4**(1), 14 (2021) https://doi.org/10.1038/s41746-021-00381-z

[39] Reich, Christian and Ostropolets, Anna and Ryan, Patrick and Rijnbeek, Peter and Schuemie, Martijn and Davydov, Alexander and Dymshyts, Dmitry and Hripcsak, George: OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. Journal of the American Medical Informatics Association **31**(3), 583–590 (2024)

[40] Mason, Darcy and Suever, J and Vanessasaurus, I and Orfanos, Dimitri Papadopoulos and Lemaitre, Guillaume and Panchal, Aditya and Rothberg, Alex and Herrmann, Markus D and Massich, Joan and Kerns, James and others: pydicom/pydicom: pydicom 2.3. 1. Published online March **29** (2022)

[41] Goode, Adam and Gilbert, Benjamin and Harkes, Jan and Jukic, Drazen and Satyanarayanan, Mahadev: OpenSlide: A vendor-neutral software foundation for digital pathology. Journal of pathology informatics **4**(1), 27 (2013)

[42] Pocock, Johnathan and Graham, Simon and Vu, Quoc Dang and Jahanifar, Mostafa and Deshpande, Srijay and Hadjigeorghiou, Giorgos and Shephard, Adam and Bashir, Raja Muhammad Saad and Bilal, Mohsin and Lu, Wenqi and others: TIAToolbox as an end-to-end library for advanced tissue image analytics. Communications medicine **2**(1), 120 (2022)

[43] Nikolas Koutsoubis and Asim Waqas and Yasin Yilmaz and Ravi P. Ramachandran and Matthew Schabath and Ghulam Rasool: Future-Proofing Medical Imaging with Privacy-Preserving Federated Learning and Uncertainty Quantification: A Review (2024). https://arxiv.org/abs/2409.16340

[44] RAPIDS AI: cuCIM - RAPIDS GPU-accelerated image processing library (2025). https://github.com/rapidsai/cucim

[45] NVIDIA Corporation: GPUDirect Storage (2025). https://developer.nvidia.com/gpudirect-storage

[46] Otsu, Nobuyuki and others: A threshold selection method from gray-level histograms. Automatica **11**(285-296), 23–27 (1975)

[47] Berman, A.G., Orchard, W.R., Gehrung, M., Markowetz, F.: SliDL: A toolbox for processing whole-slide images in deep learning. PloS One **18**(8), 0289499 (2023)

[48] Reinhard, Erik and Adhikhmin, Michael and Gooch, Bruce and Shirley, Peter: Color transfer between images. IEEE Computer graphics and applications **21**(5), 34–41 (2001)

[49] Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110 (2009). https://doi.org/10.1109/ISBI.2009.5193250

[50] Vahadane, Abhishek and Peng, Tingying and Sethi, Amit and Albarqouni, Shadi and Wang, Lichao and Baust, Maximilian and Steiger, Katja and Schlitter, Anna Melissa and Esposito, Irene and Navab, Nassir: Structure-preserving color normalization and sparse stain separation for histological images. IEEE transactions on medical imaging **35**(8), 1962–1971 (2016)

[51] Elhadad, Ahmed and Jamjoom, Mona and Abulkasim, Hussein: Reduction of NIFTI files storage and compression to facilitate telemedicine services based on quantization hiding of downsampling approach. Scientific Reports **14**(1), 5168 (2024)

[52] Levitas, Daniel and Hayashi, Soichi and Vinci-Booher, Sophia and Heinsfeld, Anibal and Bhatia, Dheeraj and Lee, Nicholas and Galassi, Anthony and Niso, Guiomar and Pestilli, Franco: ezBIDS: Guided standardization of neuroimaging data interoperable with major data archives and platforms. Scientific data **11**(1), 179 (2024)

[53] Larobina, Michele: Thirty Years of the DICOM Standard. Tomography **9**(5), 1829–1838 (2023) https://doi.org/10.3390/tomography9050145

[54] Schäfer, R., Nicke, T., Höfener, H., Lange, A., Merhof, D., Feuerhake, F., Schulz, V., Lotz, J., Kiessling, F.: Overcoming data scarcity in biomedical imaging with a foundational multi-task model. Nature Computational Science **4**(7), 495–509 (2024)

[55] Wulms, Niklas and Eppe, Sven and Dehghan-Nayyeri, Mahboobeh and Streeter, Adam J and Bonberg, Nadine and Berger, Klaus and Sundermann, Benedikt and Minnerup, Heike: The R package for DICOM to brain imaging data structure conversion. Scientific data **10**(1), 673 (2023)

[56] Dhapola, Parashar and Rodhe, Johan and Olofzon, Rasmus and Bonald, Thomas and Erlandsson, Eva and Soneji, Shamit and Karlsson, Göran: Scarf enables a highly memory-efficient analysis of large-scale single-cell genomics data. Nature communications **13**(1), 4616 (2022)

[57] Ma, Jun and He, Yuting and Li, Feifei and Han, Lin and You, Chenyu and Wang, Bo: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)

[58] Jin, Shangzhu and Yu, Sheng and Peng, Jun and Wang, Hongyi and Zhao, Yan: A novel medical image segmentation approach by using multi-branch segmentation network based on local and global information synchronous learning. Scientific Reports **13**(1), 6762 (2023)

[59] Primakov, Sergey P and Ibrahim, Abdalla and van Timmeren, Janita E and Wu, Guangyao and Keek, Simon A and Beuque, Manon and Granzier, Renée WY and Lavrova, Elizaveta and Scrivener, Madeleine and Sanduleanu, Sebastian and others: Automated detection and segmentation of non-small cell lung cancer computed tomography images. Nature communications **13**(1), 3423 (2022)

[60] Akinci D'Antonoli, Tugba and Berger, Lucas K and Indrakanti, Ashraya K and Vishwanathan, Nathan and Weiss, Jakob and Jung, Matthias and Berkarda, Zeynep and Rau, Alexander and Reisert, Marco and Küstner, Thomas and others: TotalSegmentator MRI: Robust Sequence-independent Segmentation of Multiple Anatomic Structures in MRI. Radiology **314**(2), 241613 (2025)

[61] Reska, Daniel and Kretowski, Marek: GPU-accelerated lung CT segmentation based on level sets and texture analysis. Scientific Reports **14**(1), 1444 (2024)

[62] Zhu, Bo and Liu, Jeremiah Z and Cauley, Stephen F and Rosen, Bruce R and Rosen, Matthew S: Image reconstruction by domain-transform manifold learning. Nature **555**(7697), 487–492 (2018)

[63] Qu, Hepeng and Liu, Kun and Zhang, Lina: Research on improved black widow algorithm for medical image denoising. Scientific Reports **14**(1), 2514 (2024)

[64] Koonjoo, Neha and Zhu, Bo and Bagnall, G Cody and Bhutto, D and Rosen, MS: Boosting the signal-to-noise of low-field MRI with deep learning image reconstruction. Scientific reports **11**(1), 8248 (2021)

[65] Lu, Peng and Oetjen, Karolyn A and Bender, Diane E and Ruzinova, Marianna B and Fisher, Daniel AC and Shim, Kevin G and Pachynski, Russell K and Brennen, W Nathaniel and Oh, Stephen T and Link, Daniel C and others: IMC-Denoise: a content aware denoising pipeline to enhance Imaging Mass Cytometry. Nature communications **14**(1), 1601 (2023)

[66] Sanaei, Behnoush and Faghihi, Reza and Arabi, Hossein: Employing Multiple Low-Dose PET Images (at Different Dose Levels) as Prior Knowledge to Predict Standard-Dose PET Images. Journal of Digital Imaging **36**(4), 1588–1596 (2023)

[67] Lee, Hyunjong and Seo, Sujin and Won, Sungho and Park, Woong-Yang and Choi, Joon Young and Lee, Kyung-Han and Lee, Se-Hoon and Moon, Seung Hwan: Comparative analysis of batch correction methods for FDG PET/CT using metabolic radiogenomic data of lung cancer patients. Scientific Reports **13**(1), 18247 (2023)

[68] Mottola, Margherita and Ursprung, Stephan and Rundo, Leonardo and Sanchez, Lorena Escudero and Klatte, Tobias and Mendichovszky, Iosif and Stewart, Grant D and Sala, Evis and Bevilacqua, Alessandro: Reproducibility of CT-based radiomic features against image resampling and perturbations for tumour and healthy kidney in renal cancer patients. Scientific Reports **11**(1), 11542 (2021)

[69] Bernatz, Simon and Zhdanovich, Yauheniya and Ackermann, Jörg and Koch, Ina and Wild, Peter J and Dos Santos, Daniel Pinto and Vogl, Thomas J and Kaltenbach, Benjamin and Rosbach, Nicolas: Impact of rescanning and repositioning on radiomic features employing a multi-object phantom in magnetic resonance imaging. Scientific reports **11**(1), 14248 (2021)

[70] Leijenaar, Ralph TH and Nalbantov, Georgi and Carvalho, Sara and Van Elmpt, Wouter Jc and Troost, Esther GC and Boellaard, Ronald and Aerts, Hugo JWL and Gillies, Robert J and Lambin, Philippe: The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports **5**(1), 11075 (2015)

[71] Marfisi, Daniela and Tessa, Carlo and Marzi, Chiara and Del Meglio, Jacopo and Linsalata, Stefania and Borgheresi, Rita and Lilli, Alessio and Lazzarini, Riccardo and Salvatori, Luca and Vignali, Claudio and others: Image resampling and discretization effect on the estimate of myocardial radiomic features from T1 and T2 mapping in hypertrophic cardiomyopathy. Scientific Reports **12**(1), 10186 (2022)

[72] Wuschner, Antonia E and Wallat, Eric M and Flakus, Mattison J and Shanmuganayagam, Dhanansayan and Meudt, Jennifer and Christensen, Gary E and Reinhardt, Joseph M and Miller, Jessica R and Lawless, Michael J and Baschnagel, Andrew M and others: Radiation-induced Hounsfield unit change correlates with dynamic CT perfusion better than 4DCT-based ventilation

measures in a novel-swine model. Scientific reports **11**(1), 13156 (2021)

[73] Carré, A., Klausner, G., Edjlali, M., Lerousseau, M., Briend-Diop, J., Sun, R., Ammari, S., Reuzé, S., Alvarez Andres, E., Estienne, T., *et al.*: Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. Scientific reports **10**(1), 12340 (2020)

[74] Daisaki, Hiromitsu and Kitajima, Kazuhiro and Nakajo, Masatoyo and Watabe, Tadashi and Ito, Kimiteru and Sakamoto, Fumi and Nakahara, Tadaki and Ishibashi, Mana and Toriihara, Akira: Usefulness of semi-automatic harmonization strategy of standardized uptake values for multicenter PET studies. Scientific Reports **11**(1), 8517 (2021)

[75] Landau, Susan M and Harrison, Theresa M and Baker, Suzanne L and Boswell, Martin S and Lee, JiaQie and Taggett, Jacinda and Ward, Tyler J and Chadwick, Trevor and Murphy, Alice and DeCarli, Charles and others: Positron emission tomography harmonization in the Alzheimer's Disease Neuroimaging Initiative: A scalable and rigorous approach to multisite amyloid and tau quantification. Alzheimer's & Dementia (2024)

[76] Nadav, Yayon and Amir, Dudai and Nora, Vrieler and Oren, Amsalem and Michael, London and Hermona, Soreq: Intensify3D: Normalizing signal intensity in large heterogeneous image stacks. Scientific reports **8**(1), 4311 (2018) https://doi.org/10.1038/s41598-018-22489-1

[77] Liao, J., Chin, K.-V.: Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics **23**(15), 1945–1951 (2007)

[78] Anggraeny, F.T., Purbasari, I.Y., Munir, M.S., Muttaqin, F., Mandyarta, E.P., Akbar, F.A.: Analysis of simple data imputation in disease dataset. In: International Conference on Science and Technology (ICST 2018), pp. 471–475 (2018). Atlantis Press

[79] Ulriksborg, T.R.: Imputation of missing time series values using statistical and mathematical strategies. Department of Informatics (2022)

[80] Jantscher, Michael and Gunzer, Felix and Kern, Roman and Hassler, Eva and Tschauner, Sebastian and Reishofer, Gernot: Information extraction from German radiological reports for general clinical text and language understanding. Scientific Reports **13**(1), 2353 (2023)

[81] Ma, Ming-Wei and Gao, Xian-Shu and Zhang, Ze-Yu and Shang, Shi-Yu and Jin, Ling and Liu, Pei-Lin and Lv, Feng and Ni, Wei and Han, Yu-Chen and Zong, Hui: Extracting laboratory test information from paper-based reports. BMC Medical Informatics and Decision Making **23**(1), 251 (2023)

[82] Theodorou, Brandon and Xiao, Cao and Sun, Jimeng: Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. Nature communications **14**(1), 5305 (2023)

[83] Yang, Xi and Chen, Aokun and PourNejatian, Nima and Shin, Hoo Chang and Smith, Kaleb E and Parisien, Christopher and Compas, Colin and Martin, Cheryl and Costa, Anthony B and Flores, Mona G and others: A large language model for electronic health records. NPJ digital medicine **5**(1), 194 (2022)

[84] Hassan, Esraa and Abd El-Hafeez, Tarek and Shams, Mahmoud Y: Optimizing classification of diseases through language model analysis of symptoms. Scientific reports **14**(1), 1507 (2024)

[85] Alba, Charles and Xue, Bing and Abraham, Joanna and Kannampallil, Thomas and Lu, Chenyang: The foundational capabilities of large language models in predicting postoperative risks using clinical notes. npj Digital Medicine **8**(1), 95 (2025)

[86] Wang, Kanix and Stevens, Robert and Alachram, Halima and Li, Yu and Soldatova, Larisa and King, Ross and Ananiadou, Sophia and Schoene, Annika M and Li, Maolin and Christopoulou, Fenia and others: NERO: a biomedical named-entity (recognition) ontology with a large, annotated corpus reveals meaningful associations through text embedding. NPJ systems biology and applications **7**(1), 38 (2021)

[87] Jee, Justin and Fong, Christopher and Pichotta, Karl and Tran, Thinh Ngoc and Luthra, Anisha and Waters, Michele and Fu, Chenlian and Altoe, Mirella and Liu, Si-Yang and Maron, Steven B and others: Automated real-world data integration improves cancer outcome prediction. Nature, 1–9 (2024)

[88] Sun, Weiyi and Rumshisky, Anna and Uzuner, Ozlem: Evaluating temporal relations in clinical text: 2012 i2b2 challenge. Journal of the American Medical Informatics Association **20**(5), 806–813 (2013)

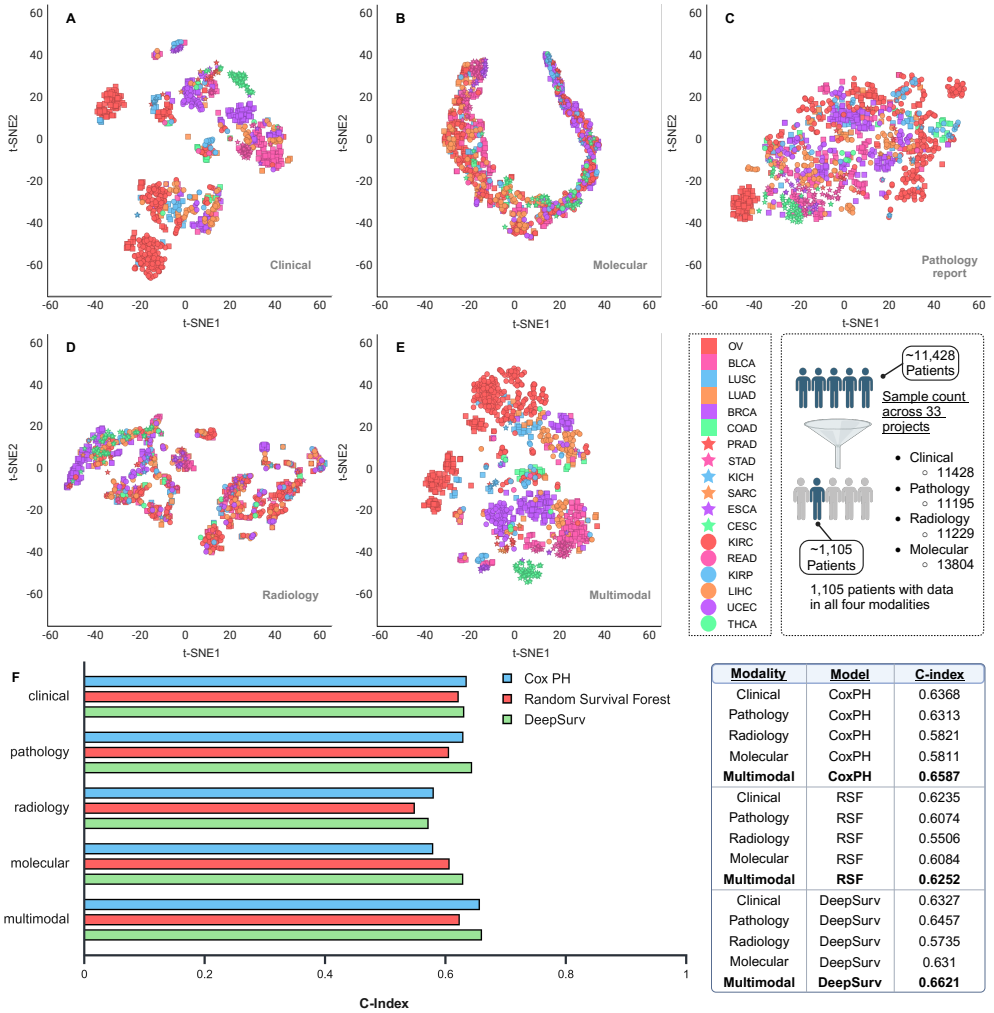
[89] Kumar, Vishesh and Stubbs, Amber and Shaw, Stanley and Uzuner, Özlem: Creation of a new longitudinal corpus of clinical narratives. Journal of biomedical informatics **58**, 6–10 (2015)

[90] Lhoest, Q., Moral, A., Jernite, Y., Thakur, A., Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., Wolf, T.: Datasets: A community library for natural language processing. In: Adel, H., Shi, S. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 175–184. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https:

//doi.org/10.18653/v1/2021.emnlp-demo.21

[91] Wu, Pengxiang and Wang, Siman and Dela Rosa, Kevin and Hu, Derek: FORB: A Flat Object Retrieval Benchmark for Universal Image Embedding. Advances in Neural Information Processing Systems **36** (2024)

[92] Yang, Kaiyu and Swope, Aidan and Gu, Alex and Chalamala, Rahul and Song, Peiyang and Yu, Shixing and Godil, Saad and Prenger, Ryan J and Anandkumar, Animashree: Leandojo: Theorem proving with retrieval-augmented language models. Advances in Neural Information Processing Systems **36** (2024)



Modality	Model	C-index
Clinical	CoxPH	0.6368
Pathology	CoxPH	0.6313
Radiology	CoxPH	0.5821
Molecular	CoxPH	0.5811
Multimodal	**CoxPH**	**0.6587**
Clinical	RSF	0.6235
Pathology	RSF	0.6074
Radiology	RSF	0.5506
Molecular	RSF	0.6084
Multimodal	**RSF**	**0.6252**
Clinical	DeepSurv	0.6327
Pathology	DeepSurv	0.6457
Radiology	DeepSurv	0.5735
Molecular	DeepSurv	0.631
Multimodal	**DeepSurv**	**0.6621**

Fig. 2 Multimodal embeddings enhance cancer subtype clustering and survival prediction. (A–E) t-SNE plots of patient embeddings (n = 1,105), colored by cancer types, for each individual modality and their integration: (A) clinical text, (B) molecular profiles, (C) pathology reports, (D) radiology images, and (E) multimodal embeddings. The multimodal embedding space demonstrates more distinct and compact clusters, indicating improved cancer-type separability compared to single-modality embeddings. **(F)** Concordance index (C-index) values for survival prediction across modalities and models, Cox proportional hazards (CoxPH), random survival forest (RSF), and DeepSurv. Multimodal embeddings yield the highest performance across models, illustrating their clinical utility for prognostic modeling.

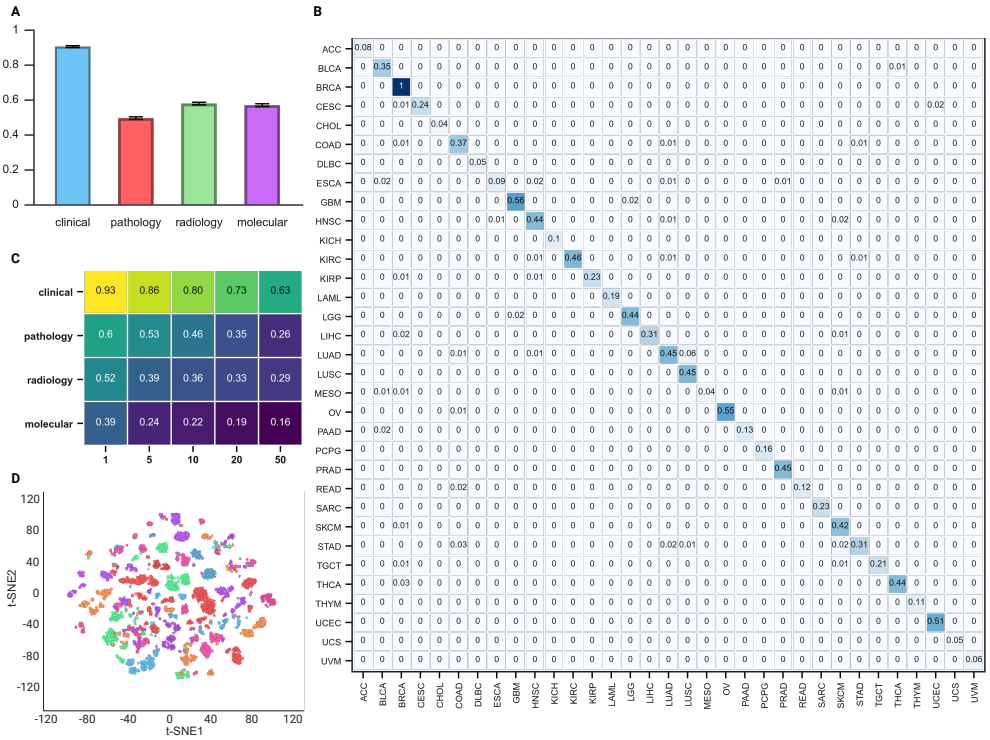


Fig. 3 HONeYBEE embeddings support robust cancer classification and retrieval tasks. (A) Classification accuracy of random forest classifiers trained on embeddings from different modalities. Clinical embeddings achieve the highest accuracy, highlighting their strong discriminative power for cancer-type prediction. **(B)** Confusion matrix for the clinical embedding classifier, showing cancer type-specific performance. **(C)** Precision@k retrieval curves for each modality, demonstrating that clinical embeddings retrieve more same-type cases consistently across varying k values. **(D)** t-SNE visualization of clinical embeddings, colored by cancer type, revealing well-separated clusters and structured latent space.

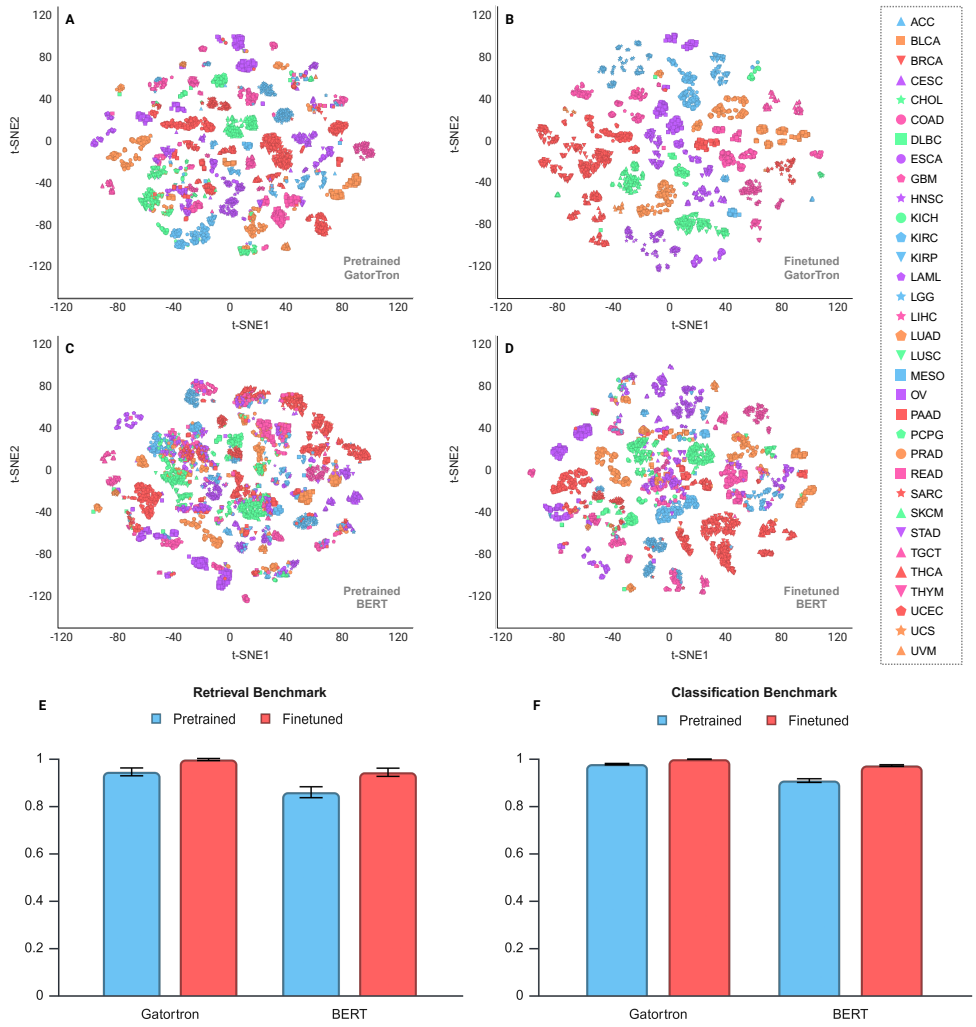


Fig. 4 Fine-tuning enhances clinical embeddings for retrieval and classification tasks. (A–D) t-SNE visualizations of clinical text embeddings extracted from TCGA data using (A) pretrained GatorTron, (B) fine-tuned GatorTron, (C) pretrained BERT, and (D) fine-tuned BERT models. Each point represents a clinical document, colored by TCGA cancer type. Fine-tuned models, particularly GatorTron, exhibit more distinct clustering by cancer type, suggesting improved embedding quality and class separation. **(E)** Retrieval performance comparison between pretrained and fine-tuned models, measured by the proportion of correctly retrieved same-cancer-type neighbors. Fine-tuning improves retrieval precision, with fine-tuned GatorTron achieving the highest scores. **(F)** Classification accuracy of random forest models trained on embeddings from each model. Fine-tuned embeddings significantly outperform pretrained counterparts, with GatorTron again demonstrating the strongest performance.

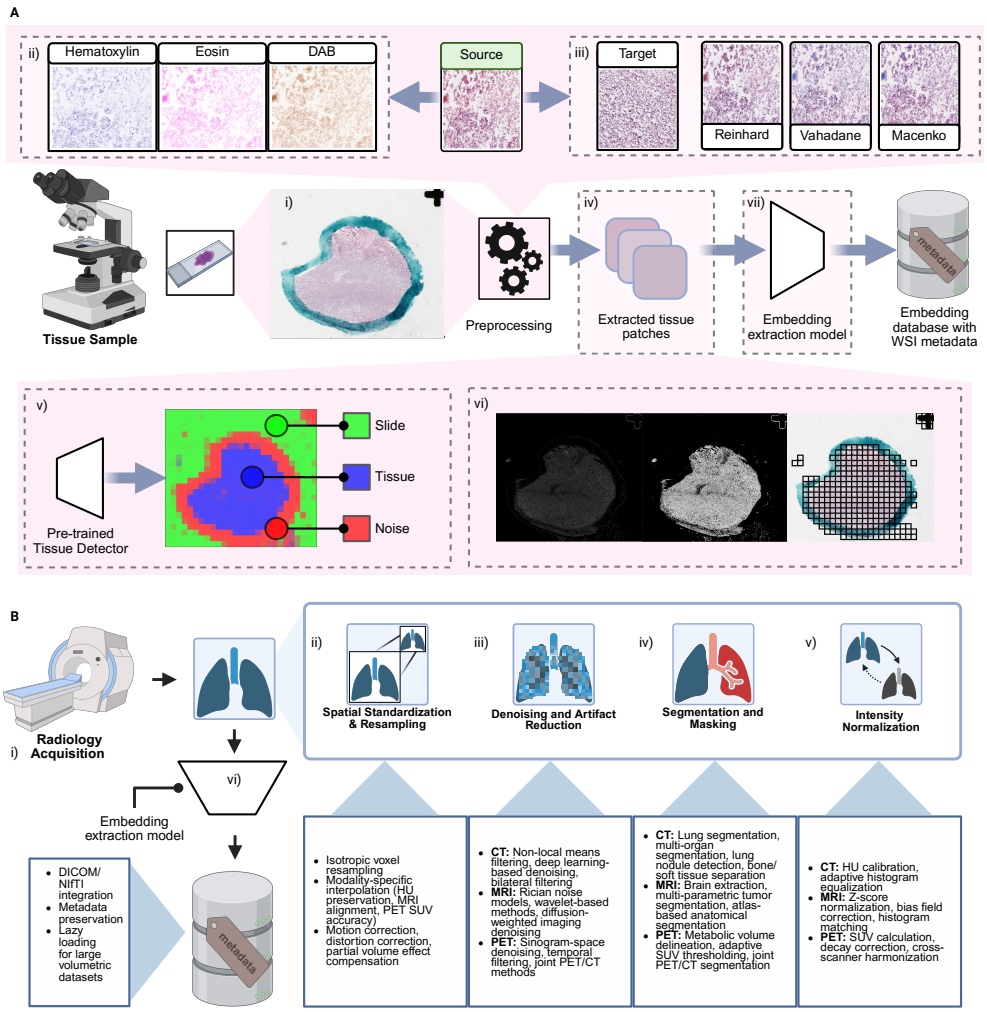


Fig. 5 Pathology and radiology data processing pipelines implemented in HONeYBEE. (A) Whole Slide Image Processing. (i) Tissue samples are digitized into gigapixel WSIs using slide scanners. (ii) Stain normalization is applied to reduce inter-slide variability caused by differences in hematoxylin, eosin, or DAB staining. Methods include Reinhard, Vahadane, and Macenko normalization techniques. (iii) The source WSI undergoes preprocessing, including tissue segmentation and filtering. (iv) Tissue regions are divided into smaller, information-rich patches for analysis. (v) A pretrained tissue detector model classifies regions as slide background, tissue, or noise to identify high-quality tissue areas. (vi) Grid-based patch extraction is performed over valid tissue regions. (vii) A pretrained embedding extraction model processes the tissue patches to generate fixed-length vectors. These embeddings, along with metadata, are stored in a structured database for downstream AI applications. **(B) Radiological Image Processing.** (i) Radiology images are acquired and ingested in DICOM or NIfTI format. (ii) Spatial standardization and resampling are applied to harmonize voxel spacing and orientation. (iii) Denoising and artifact reduction methods, such as non-local means or deep learning-based techniques, improve signal quality. (iv) Segmentation models isolate anatomical structures or lesions. (v) Intensity normalization ensures consistency across patients and scanners. (vi) Preprocessed images are passed through an embedding extraction model, and the resulting feature vectors are stored with metadata. This standardized pipeline enables downstream analysis such as classification, retrieval, and prognosis modeling across both pathology and radiology modalities.

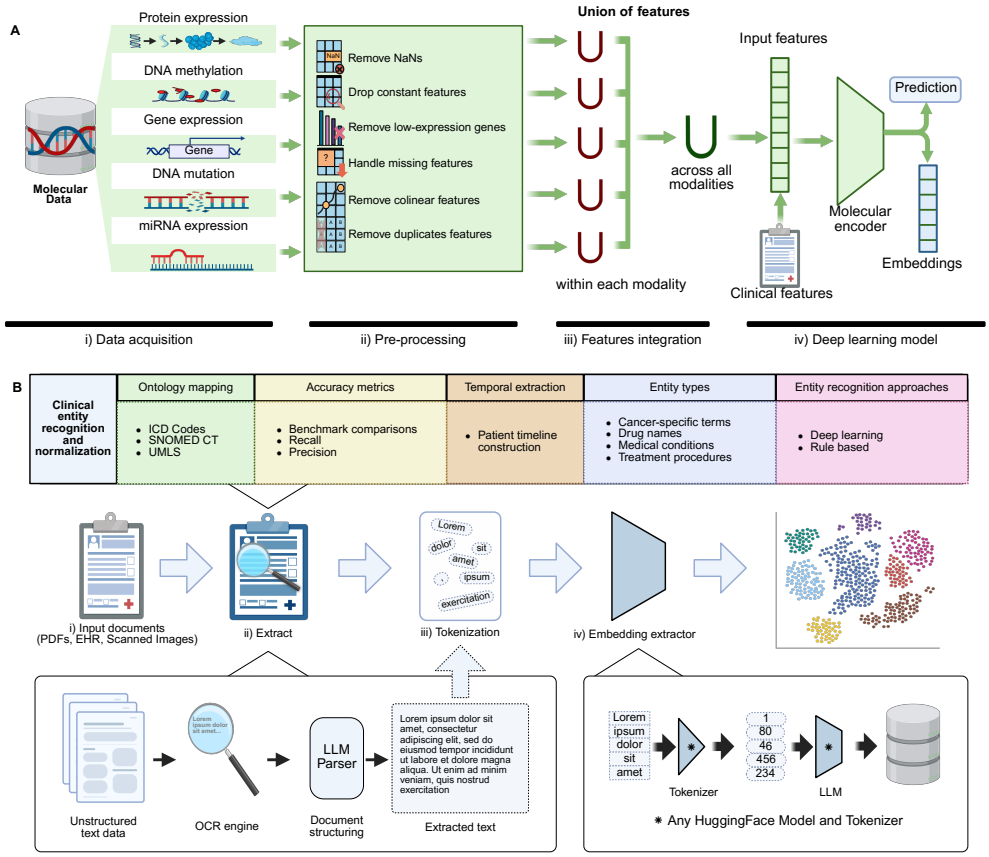


Fig. 6 Molecular and clinical data processing pipelines implemented in HONeYBEE. (A) Molecular Processing: Multimodal molecular data—such as protein expression, DNA methylation, gene expression, DNA mutation, and miRNA expression—is acquired from public repositories. Preprocessing includes the removal of missing values, constant or duplicate features, low-expression genes, and collinear features. Features are first unified within each modality, then across modalities, and subsequently integrated with clinical data. The combined feature set is passed through a molecular encoder [24] to generate embeddings or predictions for downstream tasks. **(B) Clinical Processing:** Structured and unstructured clinical data, including EHRs, PDFs, and scanned reports, undergo entity recognition, normalization, and embedding. Key stages include: (i) document input, (ii) entity extraction and OCR for scanned text, (iii) tokenization using domain-specific tokenizers, and (iv) embedding generation using Hugging Face-compatible language models. The pipeline supports concept mapping (e.g., ICD, SNOMED CT), accuracy benchmarking, timeline construction, and identification of cancer-specific terms, enabling structured integration of clinical narratives for downstream AI applications.