Multi-Omic Transfer Learning for the Diagnosis & Prognosis of Blood Cancers

Leonardo P. A. Biral
Department of Medicine
Computational Biology & Bioinformatics
Duke University, Durham, NC
leonardo.biral@duke.edu

Sandeep S. Dave
Department of Medicine
Duke Cancer Institute
Duke University, Durham, NC
sandeep.dave@duke.edu

Abstract

Blood cancers are common, affecting over 10 million individuals worldwide. Genomic approaches have shown promise in better diagnosing and risk-stratifying these cancers, but scaling up those approaches with machine learning has been challenging due to a lack of sufficient training data and robust classification approaches. Current models ignore disease taxonomies, underrepresent the full diversity of diagnoses, and fail to handle out-of-distribution (OOD) subtypes. We sought to leverage the largest ever multi-omic blood cancer dataset of over 5,000 genomically-profiled tumors that span more than 160 distinct subtypes to develop BLOOM, a deep learning pipeline that learns stable embeddings across transcriptomic, mutational, and fusion inputs. These embeddings enable taxonomy-aware, OOD-robust diagnoses with high validation performance (97.9% diagnostic precision) and strong cross-institution generalization (95.1% diagnostic precision). BLOOM's embeddings also transfer to survival prediction, attaining validation c-indices of 0.739 and 0.690 for overall and progression-free survival respectively. Through these information-rich representations of diverse genomic inputs, BLOOM highlights the clinical utility of transfer learning for blood cancer diagnosis and prognosis.

1 Introduction

Blood cancers account for roughly 10% of all cancer cases and affect over 10 million patients globally, presenting a major clinical challenge due to their high mortality and extreme biological heterogeneity [1, 2]. There are more than 160 distinct subtypes of blood cancer, organized into a four-level taxonomy by the World Health Organization (WHO) Classification System for Hematolymphoid Tumors [3]. Yet, accurate diagnosis remains difficult as traditional workflows are time-consuming, resource-intensive, and often limited by incomplete profiling or subjective interpretation.

Most prior machine learning approaches for blood cancer diagnosis rely on images or single genomic data-types and span a narrow range of diagnoses, limiting their performance and clinical utility [4–6]. Moreover, nearly all published classifiers directly predict across diagnotic classes, ignoring the inherent hierarchical relatedness among subtypes [7–9]. Such flat frameworks cannot leverage shared lineage features linking related diseases and lack mechanisms for handling cases with OOD subtypes, making them unreliable in real-world clinical settings. In contrast, multi-omic embeddings that encode diverse genomic modalities better reflect the breadth of disease biology. Passing such embeddings into a hierarchy-aware model allows for flexible, OOD-robust predictions that account for the relationships between genomic features and disease lineages.

Another critical clinical need is predicting prognosis. Current risk models rely on broad clinical features or limited genomic data-types and cover small subsets of diagnostic categories. Such models

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multimodal Representation Learning for Healthcare.

often fail to capture the complexity of factors underlying patient outcomes, which limits their clinical applicability [10–13]. Generalizable multi-omic embeddings can allow models to connect genomic features with outcome-relevant biology, improving prognostic predictions. Such approaches can stratify patients into clinically meaningful risk groups, informing personalized treatment regimens.

We present Blood cancer Learning via Omics and Ontology-aware Modeling (BLOOM), a transfer learning framework that learns stable multi-omic embeddings with effective application to diagnostic and prognostic tasks. Our study draws on the Atlas of Blood Cancer Genomes (ABCG), a dataset generated from over 5,000 patients that is the largest, most comprehensive multi-omic blood cancer cohort to date [14]. It spans 163 blood cancer subtypes we profiled with matched gene expression, mutations, and fusion calling. To our knowledge, no prior study has integrated such diverse omics data-types across this broad a spectrum of blood cancers. In BLOOM, we use a source model to encode these multi-omic inputs into a low-dimensional embedding. We show this embedding leads to hierarchical, OOD-robust diagnoses and accurate prognoses across institutions when passed into specialized target models, establishing a foundation for multi-omic transfer learning in precision hematologic oncology.

2 Methods

2.1 BLOOM Pipeline

The ABCG cohort (n=5,476; 163 diagnoses) was processed through bioinformatic pipelines for gene expression, variant calling, and fusion detection (Figures 1a,S2). The cohort was then divided into an in-distribution (ID) set (n=4,081) from 34 well-represented diagnoses and an OOD set (n=1,395) of 129 rare entities (Figure S1). We randomly split the ID set into train, validation, and test sets with the test and OOD sets used solely for post-training evaluation. To determine model robustness to batch effects, we also used leave-one-hospital-out cross-validation (LOHOCV) on models trained both with and without clinical data evaluating performance on the hold-out set (Figure S6).

2.2 Source Model: Diagnosis-Aligned Multi-Omic Embedding

Our source model converts high-dimensional transcriptomic, mutation, and fusion features into low-dimensional, information-rich representations using an expression encoder and integration head (Figure 1b, Table S1). The expression encoder manages the high dimensionality of the transcriptomic input (12,438 genes), learning a 256-dimensional embedding that preserves relevant expression patterns. This embedding is concatenated with mutation and fusion features which is then processed by the integration head, a multilayer perceptron (MLP) that classifies among the 34 ID blood cancer subtypes, outputting a 34-dimensional vector of class scores. These class scores are temperature-scaled and concatenated to a 256-dimensional learned feature vector from the integration head's penultimate layer to produce a diagnosis-aligned multi-omic embedding (DAME) that captures the holistic genomic landscape of each sample. DAME is the BLOOM pipeline's source embedding that is transferred to the hierarchical diagnosis and survival prediction target tasks. Importantly, both components of the source model are jointly learned during training, allowing the expression encoder to generate embeddings directly informed by the diagnostic objective.

2.3 Hierarchical Softmax: Hierarchical Diagnosis Classification

DAMEs are inputted to a specialized Hierarchical Softmax model (HSM) that outputs a structured diagnosis aligned to the WHO-defined taxonomy. DAMEs first pass through the HSM's shared encoder, an MLP projecting them into a shared embedding space. This embedding enters a Softmax tree in which each internal node of the WHO taxonomy is represented by a dedicated linear layer whose Softmax output defines a conditional probability distribution over its children (Figure 1c).

Post-training, we define minimum confidence thresholds at each internal node. During inference, if the model's confidence for all child nodes under a parent falls below its threshold, prediction halts at that parent node. This mechanism allows classifications to degrade gracefully on uncertain or OOD samples, outputting higher-level predictions for safe and robust clinical deployment. We select a conservative threshold as misclassifications in a clinical setting can have severe treatment consequences. In other words, we allow frequent higher-level predictions to ensure deeper predictions are of the highest confidence and precision (Table S3, Figure S4).

2.4 DeepHit: Survival Prediction

DeepHit, a neural network for survival analysis, leverages DAMEs to model time-to-event risk for patient outcome prediction [15]. DAMEs are passed into two parallel DeepHit networks, one modeling overall survival (OS) and the other progression-free survival (PFS), that each output a T-dimensional vector containing risk scores for each time interval (Figure 1d, Table S4).

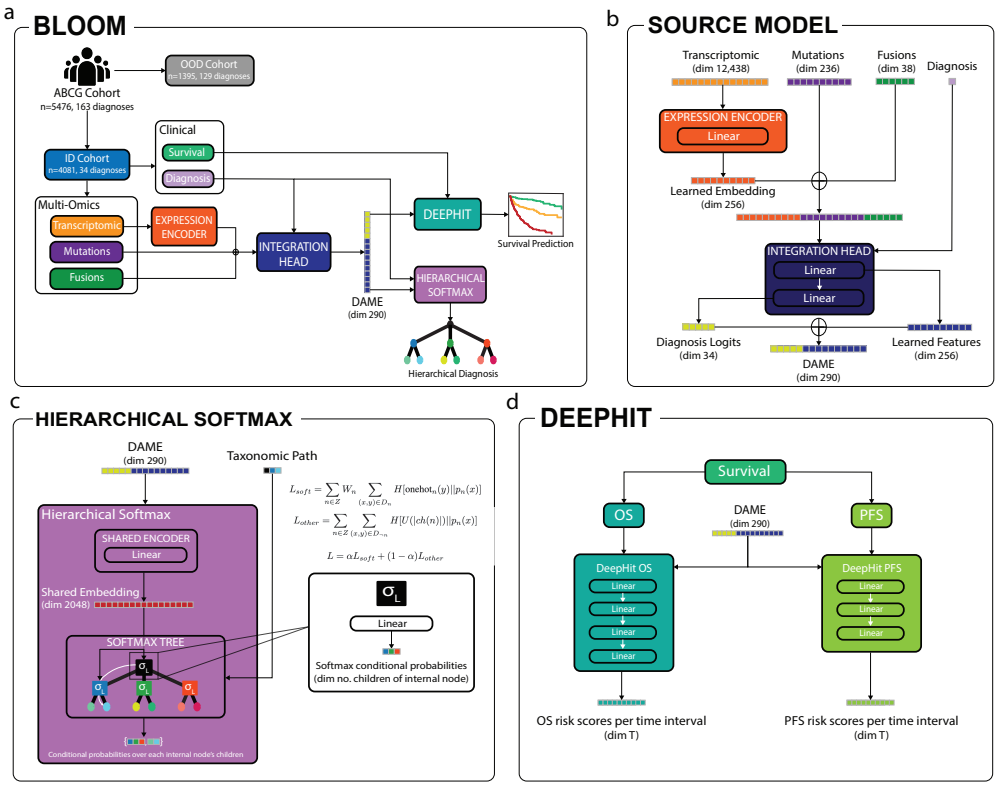

Figure 1: (a) BLOOM pipeline. (b) Source model. (c) HSM. (d) DeepHit for OS and PFS.

3 Results

3.1 Source Task: Creating DAMEs

We first evaluated the discriminative quality of DAMEs before transferring them to the target tasks. The source model effectively classified over the 34 ID diagnoses with validation, test, and LOHOCV hold-out macro-averaged AUROCs of 0.975±0.1%, 0.981±0.1%, and 0.967±1.2% which corresponded to accuracies of 74.6±1.0%, 75.1±0.8%, and 70.2±4.4% respectively (Figures S3,S6). t-SNEs of the source model's learned features reveal clear separation in the ID set among cells-of-origin (COOs), diagnoses, and DeepHit 3-year OS risk tertiles demonstrating the source model's greater ability to capture clinically-relevant genomic structure compared to raw gene expression (Figures 2a-b,S2-3) [16]. Our source model achieved significantly higher diagnostic accuracy than one using a Graph Neural Network (GNN) integration head ($p<0.01$) and outperformed Decision Tree, Random Forest, and XGBoost classifiers in COO and diagnostic macro-averaged AUROC and accuracy ($p<1e-4$) (Figure S3) [17].

3.2 Target Task 1: Hierarchical Blood Cancer Diagnosis

The HSM achieved validation precisions of 98.2±0.1% and 97.9±0.6% in classifying COO (level 1) and specific diagnoses (level 4) respectively as well as high ROC separation across all internal nodes of the hierarchy (Figures 2c,S4). This strong ID performance extended to the OOD set, classifying COO with a precision of 91.6±0.4% and achieving a hierarchically-weighted precision of 89.3±1.3%. Robust performance on OOD cases is due to effective threshold calibration causing the HSM to halt predictions at higher taxonomic levels, only predicting to the diagnosis level in 2.29±1.1%

of OOD cases. Conversely, the HSM predicted to the diagnosis level in 11.5±1.6% of validation cases, revealing significantly different diagnostic behaviors for ID and OOD samples that match our desiderata (p=1.06e-6). LOHOCV hold-out performance was similar with COO, diagnostic, and weighted precisions of 98.1±1.0%, 95.1±4.2%, and 97.9±1.0% respectively (Figures 2c,S6).

3.3 Target Tasks 2 & 3: Overall and Progression-Free Survival Analysis

Kaplan-Meier (KM) curves of DeepHit-predicted PFS and OS 3-year risk tertiles show significant stratification (p<1e-6) (Figures 2d–e). The validation concordance indices (c-indices) of the DeepHit OS and PFS models were 0.739±3.0e-3 and 0.690±6.1e-3 respectively, significantly outperforming age at diagnosis Cox models as well as DeepHit models trained on clinical diagnostic labels, raw genomic inputs, and the source model's learned features (p<0.05) (Figure S5). High OOD and LOHOCV hold-out performance were also achieved with OS c-indices of 0.697±1.1e-2 and 0.742±2.5e-2 and PFS c-indices of 0.634±1.6e-2 and 0.665±2.4e-2 respectively (Figures S5-6).

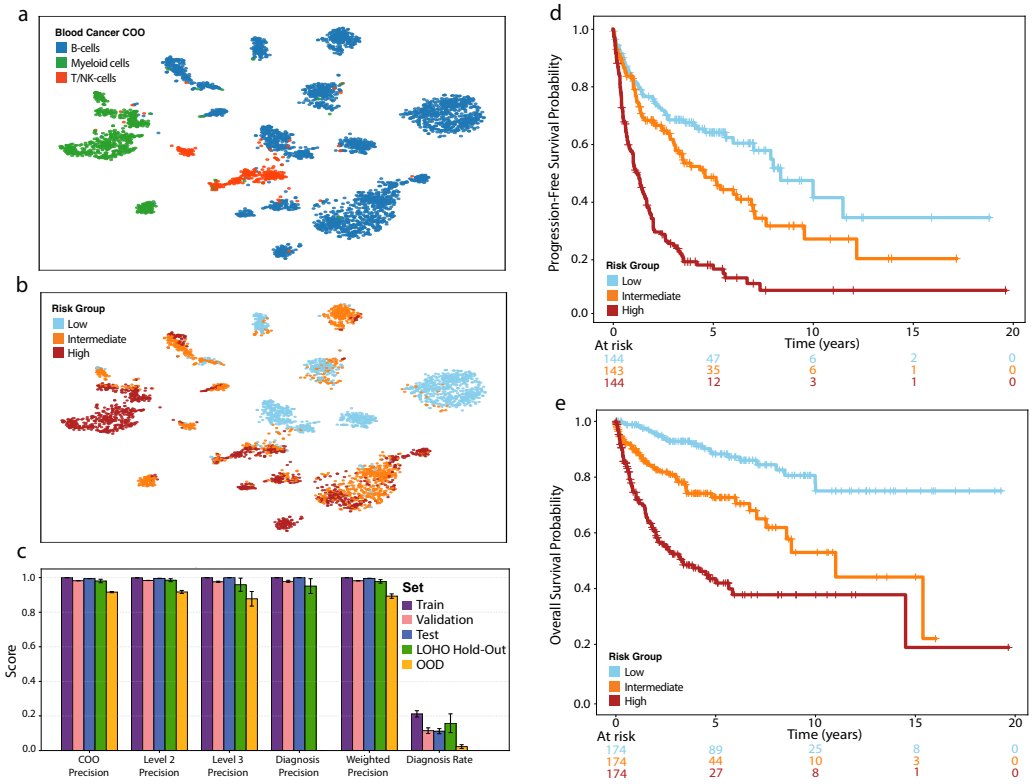

Figure 2: Learned feature t-SNEs colored by (a) COO and (b) DeepHit 3-year OS risk tertiles. (c) HSM performance. Validation set KM curves of DeepHit (d) PFS and (e) OS 3-year risk tertiles.

4 Discussion

BLOOM establishes a unified transfer learning framework that generates embeddings from transcriptomic, mutation, and fusion data for two clinically impactful tasks: diagnosis and risk stratification of blood cancers. The high ID precision and robust OOD-handling of our HSM and the DeepHit models' significant ability to stratify prognostic subgroups highlight the predictive value of DAMEs. The target models' consistent LOHOCV hold-out performance when trained both with and without clinical data strongly suggests DAMEs capture true predictive signal rather than institution-dependent batch effects. Future extensions could incorporate additional data modalities such as copy-number alterations, viral detection, and biopsy site images to maximize the breadth of information encoded by DAMEs. Additionally, fine-tuning the source model while training the target models could enhance the specialization of the embeddings for the specific task. Finally, alternative source embeddings such as a contrastive model that maps multi-omic data to textual descriptions of each malignancy could be explored. Overall, the BLOOM pipeline highlights the value of transfer learning: rich, multi-omic source embeddings can be the foundation for clinically impactful precision oncology.

5 Supplementary Methods

5.1 Blood Cancer Taxonomy

Tumor samples in the ABCG cohort were mapped to the taxonomy defined by the WHO Classification System for Hematolymphoid Tumors (5th ed.). Adding a root (blood cancer), this taxonomy forms a rooted directed acyclic graph (DAG) spanning five levels, from the root through level 1 (COO) all the way down to the leaf nodes representing specific subtypes (Figure S1a). Consistent with population frequencies, mature B-cell neoplasms, acute myeloid leukemia (AML), and mature T/NK-cell neoplasms are the most common subcategories for B, myeloid, and T/NK-cell malignancies respectively, underscoring both the depth and breadth of the ABCG dataset (Figures S1b–c) [1, 2]. Sample counts for level 3 and ID diagnosis categories are also provided (Tables S5–6).

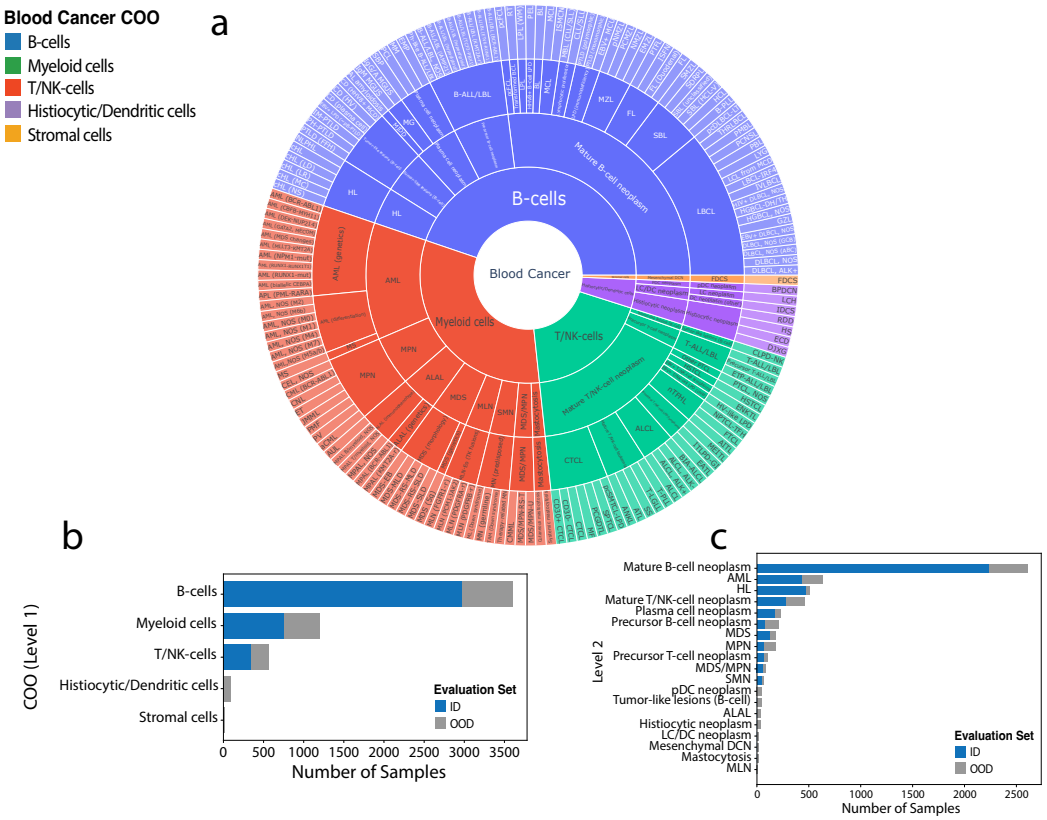

Figure S1: (a) WHO blood cancer taxonomy. (b) COO and (c) hierarchy level 2 counts colored by evaluation set. Abbreviations: HL (Hodgkin Lymphoma), MDS (Myelodysplastic Syndrome), MPN (Myeloproliferative Neoplasms), SMN (Secondary Myeloid Neoplasms), pDC (Plasmacytoid Dendritic Cell), ALAL (Acute Leukemias of mixed or ambiguous lineage), LC/DC (Langerhans Cell or other Dendritic Cell), DCN (Dendritic Cell Neoplasms), Myeloid/Lymphoid Neoplasms (MLN).

The ABCG cohort's genomic data is highly diverse and concordant with the published literature. Mutation profiles show recurrent alterations in known oncogenic genes such as *KMT2D*, *TP53*, *BCL2*, and *NPM1* and a distribution of variants consistent with the literature such as the overrepresentation of *TET2* alterations in T/NK-cell cancers (p=1.01e-7) (Figure S2a) [18]. t-SNE visualization of transcriptomic profiles confirms this heterogeneity, revealing clear separation between myeloid and lymphoid malignancies as well as isolated clustering within lineages (Figure S2b). This indicates the hierarchical structure of subtype relatedness is represented within the genomic data. Fusion analysis reveals extensive heterogeneity across and within these lineages while confirming known tumor biology (Figure S2c). For instance, we observe that 32 of 34 (94.1%) of cases with *PML-RARA* occur in Acute Promyelocytic Leukemia (APL), a myeloid malignancy characterized by that fusion [19].

Clinical outcomes were measured by PFS and OS. Across the full cohort, the median survival time was 4.2 years (95% CI 3.7-4.7) with 1362 (52.7%) events and 16 years (95% CI 15-21) with 830 (26.7%) events for PFS and OS respectively. KM curves reveal notable stratification across COOs, with B-cell malignancies exhibiting broadly better PFS (p=2.16e-42) and OS (p=1.11e-70) than myeloid and T/NK-cell malignancies (Figures S2d-e).

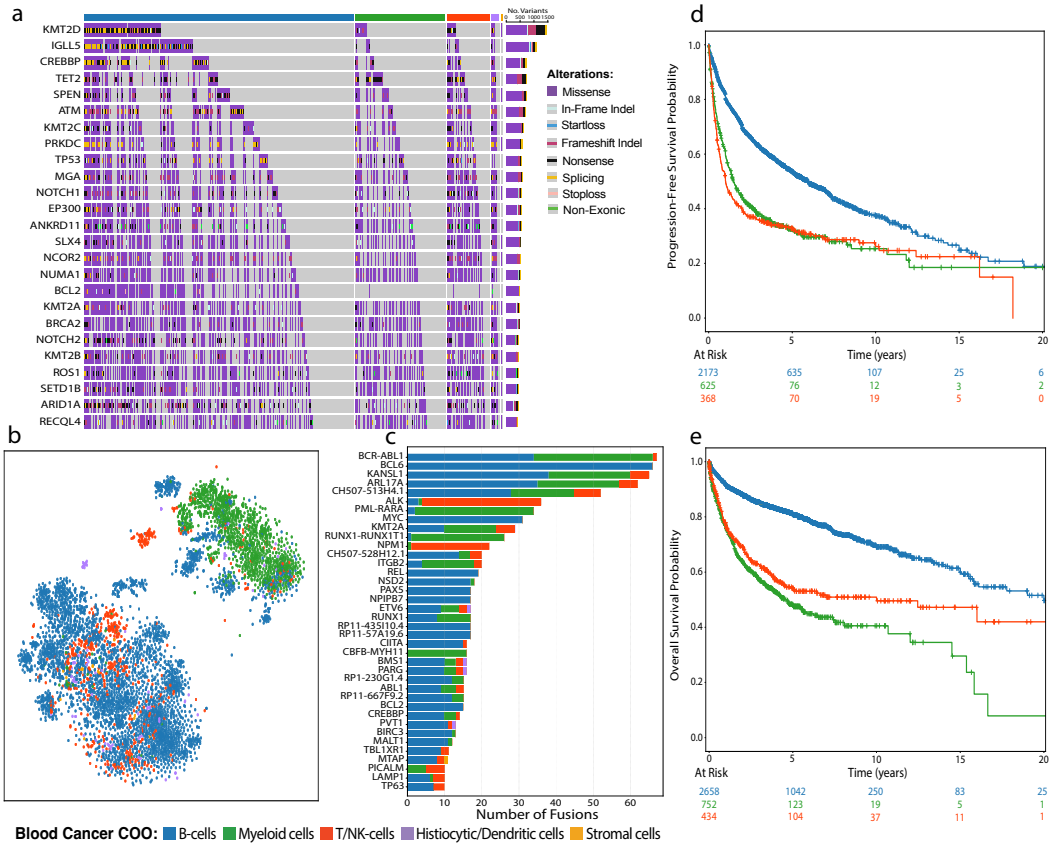

Figure S2: (a) Oncoprint of the 25 most recurrently mutated genes in the cohort column-split by COO with mutations colored by variant type. (b) Gene expression t-SNE colored by COO. (c) Number of fusions in the cohort colored by COO. (d) PFS and (e) OS KM curves stratified by COO. Note histiocytic/dendritic and stromal cell malignancies were excluded in d-e as few such cases had survival data.

5.2 Sequencing

Our study uses cases from the ABCG dataset, an international collaborative effort from over 25 clinical sites in which we aimed to collect and sequence cases that represent the full blood cancer family. The study cohort spans 163 hematologic malignancies and includes 5,476 formalin-fixed, paraffin-embedded tumor biopsies collected in accordance with an IRB-approved protocol. All tissue biopsies were deidentified and diagnoses were confirmed by expert panel-based pathology review using purpose-built tools. Each tumor biopsy has gene expression and fusion calls from whole transcriptome sequencing as well as matched gene mutation data from whole exome and targeted panel sequencing. All nucleic acid extractions and sequencing library preparations were performed with Duoseq (Research Kit EPXv3, Data Driven Bioscience, Durham, NC) following the manufacturer's protocol. All libraries were sequenced on the Illumina platform according to manufacturer recommendations.

5.2.1 DNA Sequencing Analysis

FASTQ files containing DNA sequencing reads were trimmed using Trimmomatic (v0.39) in paired end mode to remove adapter sequences and low-quality reads [20, 21]. Using Sentieon BWAmem (v201911) with the default settings, DNA reads were aligned to the human genome (GRCh38.p12, with a PAR mask on chrY) [22]. PCR duplicate reads were marked using Picard (v2.8.1) [23]. Picard, FASTQC (v0.11.8), and samtools (v1.13) were used to extract quality control metrics [24].

5.2.2 Variant Calling Analysis

Variant calling analysis was performed using DNA reads from exome samples and targeted panel sequencing where available. Variants were called using a pipeline that uses the union of variants called by each of Strelka2 (v2.9.10), DeepVariant (v1.1.0), and Sentieon Haplotyper (v201911) [25, 26, 22]. Synonymous single nucleotide variants (SNVs), variants found in the separate panel of 263 normal control samples, and variants with a population frequency of greater than 0.01 reported in the gnomAD [27] databases were excluded. We only included variants included in the union of genes previously described by three major blood cancer genomics profiling studies [28–30].

Variant calling results were split into two components: insertions and deletions (indels) and pathogenic SNVs. For each patient, indels were summarized as a row vector of indel counts across 188 genes, each of which contained at least 1 indel in a minimum of 10 cases. These counts were standardized in the ID (train, validation, test) and OOD sets to the train set distribution upstream of training. We identified pathogenic SNVs among the called variants using a curated list of genomic loci with known relevance in hematological malignancies. We excluded any SNVs in this set found in fewer than 10 cases across the cohort leaving a set of 48 recurrent pathogenic SNVs. We represent this data for each patient as a binary row vector indicating presence or absence of each SNV.

5.2.3 RNA Sequencing & Gene Expression Analysis

FASTQ files containing RNA-Seq reads were trimmed using Trimmomatic (v0.39) in paired end mode to remove Illumina specific adapter sequences and low-quality reads [21]. STAR aligner (v2.7.1) was utilized for mapping RNA reads to the human genome and transcriptome (GRCh38.p12, with a PAR mask on chrY) [31]. Reads that had at least one primary alignment to the human genome were extracted using samtools (v1.13) and additionally filtered based on manufacturer recommendations using GATK (v4.1.2.0) [24, 32]. Transcript quantification was performed using salmon (v1.2.1), and the transcript-level data were summarized to the gene level using the tximport (v1.14.2) library in R [33, 34]. The result was a matrix representing raw read counts per gene per patient sample. This matrix was restricted to include only the expression of the protein-coding genes captured by exome sequencing. Genes with missing or uniformly low expression were excluded from further analysis and the remaining expressed genes were log2-normalized in R using the DESeq2 (v1.40.2) tool [35]. In all, we had expression values for 12,438 genes for each patient. These expression values were standardized in the ID and OOD sets to the train set distribution upstream of training.

5.2.4 Fusion Analysis

Fusions were identified using Arriba (v2.1.0) [36]. To retain high-confidence fusion events for downstream analysis, we applied a multi-step filtering strategy. First, we selected fusions annotated in Mitelman, a curated database of recurrent, cancer-associated fusions, or with an automated confidence label of high [37]. We then excluded fusion events lacking annotated transcript IDs for either fusion partner. To ensure we selected only biologically and technically supported events, we calculated a total support metric for each fusion as the sum of split reads from both partners and discordant mate-pair support. Fusions with fewer than 5 supporting reads were excluded from further analysis. We identified 10 unique fusion pairs (e.g. *BCR-ABL1*) and 28 single genes (e.g *ALK*) involved in fusions with different partners in at least 10 cases in the cohort. We represented these 38 fusion features per patient as a binary row vector indicating presence or absence.

5.3 Model Development

After samples were processed through bioinformatic pipelines for gene expression, variant calling, and fusion detection, the cohort was divided into an ID set (n=4,081) of 34 diagnoses that had

at least 50 representative cases and an OOD set (n=1,395) of 129 rarer entities. The ID set was randomly divided with a 90/10 train/test split and the resulting train set underwent a random 80/20 train/validation split with ID diagnostic categories stratified across all splits. We ensured the same samples were in each of the train, validation, and test sets for both source and targets models to prevent data leakage. The train and validation sets were used for all model training and development while the test and OOD sets were used only for post-training evaluation.

5.3.1 Source Model

The source model is composed of an expression encoder and integration head implemented and trained end-to-end in `PyTorch` [38]. Both of these components were jointly learned during training. We performed 1000 trials of Bayesian Optimization (BO) with `Optuna` to identify optimal architectures and hyperparameters for both source model elements [39] (Table S1). For consistent DAME generation for the target tasks, we trained the source model once with the optimized architectures and hyperparameters and froze the weights.

Table S1: Hyperparameters for source model

Expression Encoder	
Component	**Value**
Architecture	
Input feature dimension	12,438
Encoder layers	1 (Linear)
Output embedding dim	256
Activation function	ReLU
Regularization	
Dropout rate	0.5
Integration Head	
Component	**Value**
Architecture	
Input feature dimension	256
Hidden layers	1 (Linear)
Hidden layer embedding dim	256
Output dim	34
Activation function	ReLU
Regularization	
Dropout rate	0.5
Training	
Optimizer	Adam
Learning rate	2.24×10^{-5}
Weight decay (L2)	1.68×10^{-3}
Batch size	64
BO rounds	1000
Max epochs per trial	250
Early stopping patience	5
Loss function	Cross Entropy Loss

Performance The source model has strong predictive performance especially considering it classifies across 34 diagnoses. The validation confusion matrix reveals most of the misclassifications are reasonable mistakes (Figure S3a). For instance, a plurality of misclassifications occur across 3 highly related Diffuse Large B-cell Lymphoma, not otherwise specified (DLBCL, NOS) subtypes: Activated B-cell (ABC), Germinal Center B-cell (GCB), and unclassified [40]. When one considers the unclassified category can actually include ABC or GCB DLBCL, NOS cases simply because ABC and GCB status were not tested for that patient, the misclassifications are even more reasonable.

t-SNE vizualization of the learned features from the frozen source model's penultimate layer colored by diagnosis further confirms its discriminative ability across the 34 ID subtypes, revealing strong separation even among granular subtypes not just higher-level COO categories (Figure S3b). Finally, ROC curves of the 10 diagnoses with the most representative cases and the micro-averaged ROC curve further highlight the frozen source model's high diagnostic performance (Figure S3c).

Comparison to alternative models We compared the mean predictive performance of our source model trained with 5 different random states in classifying diagnosis and COO to the following classifiers (Figure S3d):

1. Decision Tree (DT)
2. Random Forest (RF)
3. XGBoost (XGB)
4. BLOOM source model with a GNN integration head

We implemented the DT and RF baselines using `scikit-learn`'s `DecisionTreeClassifier` and `RandomForestClassifier`, and the XGB baseline using the `XGBoost` Python package [41, 42]. We train these classifiers on the same mutation and fusion features passed into our source model, but replace the raw expression features with tumor microenvironment (TME) proportions deconvoluted from the full transcriptomic data using FARDEEP [43]. These TME proportions correspond to the relative fractions of immune cells in a given patient, which are highly relevant in blood cancers diagnosis, especially in determining COO. This controlled dimensionality vastly reduces the size of the feature space, improving these classifiers' computational efficiency. We train separate DT, RF, and XGB models for diagnosis and COO classification with the same train/validation/test split used for source model development. Using 10-fold grid search CV with cross-entropy loss as the evaluation metric, we optimize the hyperparameters of the DT (cost-complexity parameter), RF (number of estimators, maximum depth, and minimum samples per leaf), and XGB (number of estimators, learning rate, maximum depth, minimum child weight, and the L2-regularization).

An alternative to using an MLP as the integration head of the BLOOM source model is a GNN, which integrates multi-omic features while explicitly modeling the hierarchical relationships among diagnostic categories [17]. Each node represents a disease entity, and edges encode hierarchical dependencies between parent and child classes. The GNN receives the multi-omic feature vector and propagates it through this hierarchical graph, producing hierarchy-aware multi-omic embeddings (HAMEs) informed by both molecular similarity and taxonomic proximity. We identify optimal expression encoder and GNN architectures and hyperparameters using the same BO protocol as for the original source model.

While the DT, RF, and XGB classifiers significantly outperformed random-chance performance ($p<1e-4$), the BLOOM source model achieved significantly higher validation and test classification accuracy and macro-averaged AUROC for both diagnosis and COO ($p<1e-4$). The BLOOM source model also significantly outperformed the GNN-based source model in both validation and test diagnostic accuracy ($p<0.01$), but had no significant difference in COO classification or diagnostic macro-averaged AUROC (Figure S3d). Note that COO predictions for the BLOOM and GNN source models were obtained by grouping diagnoses in these models' respective prediction spaces by their mutually exclusive COO categories and summing the classification probabilities of these groups rather than retraining them to directly classify COO as done for the DT, RF, and XGB classifiers.

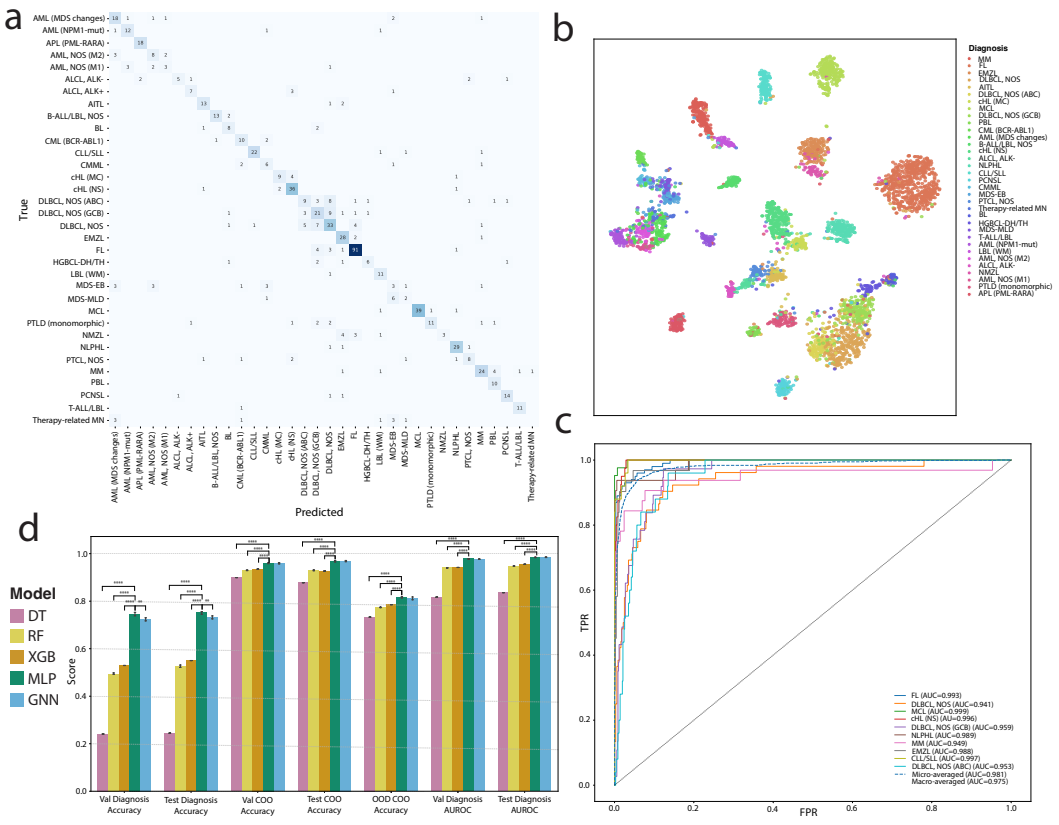

Figure S3: (a) Confusion matrix of the frozen source model's validation set predictions. (b) t-SNE of the frozen source model's learned features colored by ID diagnostic class. (c) Frozen source model ROC curves of the 10 diagnoses with the most representative cases and the micro-averaged ROC curve. (d) BLOOM source model, DT, RF, XGB, and source model with GNN integration head comparative diagnostic and COO classification macro-averaged AUROC and accuracy. ** indicates p<0.01, **** indicates p<1e-4.

Although the GNN-based source model exhibited lower classification accuracy than the BLOOM source model, the resulting HAMEs offer several potential advantages over DAMEs. HAMEs are 70-dimensional, corresponding to the GNN's logits for each of the 36 internal nodes in the WHO-defined taxonomy for the 34 ID diagnostic leaves. This means HAMEs explicitly encode hierarchical dependencies and intermediate category representations captured by the GNN. By embedding information from both fine-grained diagnoses and higher-order taxonomic levels, HAMEs may enhance downstream target model performance. This makes using a GNN integration head in the source model a promising avenue for future exploration despite the BLOOM source model performing best on the source task.

5.3.2 Hierarchical Softmax

Loss function Because the WHO-defined taxonomy is a rooted DAG, there exists a unique path from the root to each leaf. At each internal node of this hierarchy, the HSM has a linear layer with a Softmax activation function that outputs the conditional probability distribution across its child nodes. During inference, the model computes the joint probability of each diagnosis as the product of conditional probabilities along the corresponding taxonomic path, enabling consistent predictions that respect the blood cancer ontology.

Given this nested prediction task reliant on conditional probabilities, we implemented a loss function that balances confidence along the true diagnostic path with calibrated behavior at off-path nodes (Equation 1) [44, 45]. The loss includes a path-specific cross-entropy term that encourages high confidence along the ground truth hierarchical path (Equation 2). It also weighs prediction errors at internal nodes with many descendant nodes more than errors at internal nodes with fewer descendants

as higher-level misclassifications lead to a greater number of downstream predictions being incorrect (Equation 3). In addition, for all internal nodes not on the sample's path, we apply a term that penalizes deviations from a uniform distribution over those nodes' children (Equation 4). This prevents the HSM from becoming overconfident in branches outside the true prediction path, improving calibration and robustness, which is particularly important when handling samples from OOD subtypes.

$$L = \alpha L_{soft} + (1 - \alpha) L_{other} \tag{1}$$

$$L_{soft} = \sum_{n \in Z} W_n \sum_{(x,y) \in D_n} H[\text{onehot}_n(y) || p_n(x)] \tag{2}$$

$$W_n = \frac{|\{j \in \{1, ..., N\} : n \in anc(j)\}|}{|Z|} \tag{3}$$

$$L_{other} = \sum_{n \in Z} \sum_{(x,y) \in D_{\neg n}} H[U(|ch(n)|) || p_n(x)] \tag{4}$$

where n is a node in the set Z of all internal nodes with ancestors $anc(n)$ and children $ch(n)$. D_n is the set of tuples (x_i, y_i) of samples whose ancestors contain n in which x_i represents sample i's input features and y_i is its ground-truth leaf node label. $H[s||t]$ is the cross-entropy from s to t. $p_n(x_i)$ is the model's predicted probability distribution over $ch(n)$ for sample i. α is a hyperparameter dictating the relative weight of L_{soft} to L_{other} in the cumulative loss function L.

Per α model optimization Since α has a large impact on what the model prioritizes in the prediction task, we performed BO to identify optimal HSM architectures and hyperparameters for each $\alpha \in [0.05, 1]$ with a step size of 0.05. For each BO study, we ran 100 trials with the maximum epoch count per trial set to 250 with a patience of 5 epochs. We optimized the following parameters in the HSM's shared encoder: number of hidden layers, number of nodes per layer, and the regularization of these layers. We also optimized five key hyperparameters: learning rate, weight decay, batch size, and the temperature-scaling factor applied to the 34-dimensional vector of class scores outputted by the source model before concatenation with the integration head's penultimate layer features. This process returned an optimized HSM for each of the 20 α parameters explored. All HSMs were implemented and trained end-to-end in `PyTorch` [38].

Adaptive confidence thresholding To mitigate overconfident misclassifications, we evaluated a node-specific confidence threshold parameter τ. We first computed node-level confidence scores by evaluating the HSM's predicted path probabilities along all possible root-to-leaf paths and recording, for each node, the probability of selecting its correct child (Equation 5). For each n, this produced a set of scores $\{s_i(n, c), y_i\}_{i=1}^{N_n} \forall c \in ch(n)$ where s_i is sample i's cumulative predicted path probability from the root to one of n's children c and $y_i \in \{0, 1\}$ indicates whether the child lies along the true path for the set of samples N_n at node n.

$$s_i(n, c) = (\prod_{j=1}^{l-1} P(v_{j+1} | v_j, x_i)) P(c | n, x_i) \tag{5}$$

where the path to n is $\gamma(n) = (v_{1=root}, v_2, ..., v_n)$. We then used these node-wise data to estimate the threshold τ_n that ensures minimal desired precision $\tau \in [0, 1]$ at n (Equations 6-7).

$$Precision_n(t) = \sum_{c \in ch(n)} \sum_{i \in N_n} \frac{\mathbf{1}(s_i(n, c) \geq t) y_i}{\mathbf{1}(s_i(n, c) \geq t)} \tag{6}$$

$$\tau_n = min\{t | Precision_n(t) \geq \tau\} \tag{7}$$

During inference, the classifier traverses the taxonomy top-down. Starting at the root, it selects the child (COO) with the highest conditional probability and continues along that path greedily

selecting the most probable child at each internal node, updating the cumulative path probability. If the cumulative probability at a given n's prediction over $ch(n)$ along this path falls below τ_n, the model halts prediction and returns n, the last confidently assigned taxonomic class. This adaptive thresholding causes the HSM to produce shallower, higher-level classifications for ambiguous or OOD samples while maintaining high precision in confident cases.

Selecting α and τ The goal is to find a (α, τ) pair that has high validation and OOD performance in the following metrics: per-hierarchy level precision (Equation 8), hierarchically-weighted precision (Equation 9), prediction depth (Equation 10), and λ-weighted hierarchical distance (Equation 11). We performed a grid search over (α, τ) pairs and selected candidate configurations with high validation and OOD performance in these metrics from which we selected our final model (Table S2).

$$Prec_l = \frac{1}{N_l} \sum_{i:|\hat{p}_i| \geq l} \mathbf{1}\{\hat{y}_{i,l} = y_{i,l}\} \tag{8}$$

$$W_{prec} = \frac{\sum_{l=1}^{|\gamma|} \pi_l Prec_l}{\sum_{l=1}^{|\gamma|} \pi_l}, \pi_l = \frac{N_l}{N} \tag{9}$$

$$D = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{\gamma}_i|}{|\gamma_i|} \tag{10}$$

$$d_\lambda(\gamma, \hat{\gamma}) = \frac{1}{N} \sum_{i=1}^{N} [(|\gamma_i| - DCA(\gamma_i, \hat{\gamma_i})) + \lambda(|\hat{\gamma_i}| - DCA(\gamma_i, \hat{\gamma_i}))] \tag{11}$$

where N_l is the number of samples that are predicted at level l of a hierarchy and λ is the penalty assigned to misclassifications due to overconfidence relative to early stopping. $DCA(\gamma_i, \hat{\gamma_i})$ is the deepest common ancestor between the ground truth root to leaf path γ_i and the predicted path $\hat{\gamma_i}$ for sample i: $DCA(\gamma_i, \hat{\gamma_i}) = max\{k \in \{0, 1, ...|\hat{\gamma_i}|\} : \gamma_{i,j} = \hat{\gamma}_{i,j} \forall 1 \leq j \leq k\}$. Note that for this hierarchy, $|\gamma_i|$ is fixed at 5 as that is the number of levels from root to leaf in the WHO taxonomy.

Table S2: HSM candidate configurations

α	τ	λ	Highly performing metrics
0.15	0.982		OOD $Minimax_{d_\lambda} \forall \lambda$
0.2	0.991	20	Val & OOD d_λ
0.2	0.995	50	Val & OOD d_λ
0.25	0.593		OOD W_{prec} & D F1-score
0.3	0.924	2	Val & OOD d_λ
0.5	0.654		Val W_{prec} & D F1-score
0.5	0.907	1.25	Val & OOD d_λ
0.5	0.908	1.5	Val & OOD d_λ
0.7	0.905	1	Val & OOD d_λ
0.75	0.915	1.75	Val & OOD d_λ
0.8	0.99	10	Val & OOD d_λ
0.9	0.996	100	Val & OOD d_λ, Val & OOD W_{prec} F1-score
0.95	0.987	5	Val & OOD d_λ
1	0.979	3	Val & OOD d_λ
1	0.992		Val $Minimax_{d_\lambda} \forall \lambda$

For our final model selection, we first filtered out any candidate configurations with OOD weighted precisions of less than 80% to ensure our model effectively handles OOD samples. We then prioritized precision because misclassifications at any taxonomic level risk severe patient consequences in a clinical setting. However, once precision was high, we used prediction depth to assess the model's ability to return granular classifications. Given these criteria, we selected the HSM with a (α, τ) configuration of (0.95, 0.987) (Table S3).

Table S3: Hyperparameters for the selected HSM

Component	Value
Architecture	
Input feature dimension	290
Shared encoder layers	1 (Linear)
Shared encoder embedding dim	2048
Activation function	ReLU
Regularization & Threshold	
Dropout rate	0.15
Confidence threshold (τ)	0.987
Training	
Class score temperature-scaling	2
Optimizer	Adam
Learning rate	1.00×10^{-4}
Weight decay (L2)	7.71×10^{-3}
Batch size	128
BO rounds	100
Max epochs per trial	250
Early stopping patience	5
Loss function	Equation 1 ($\alpha = 0.95$)

HSMs with comparable validation precisions either underperformed on OOD cases or produced vastly shallower predictions too generic to be clinically useful (Figures S4a–b). Micro-averaged ROC curves highlight the selected HSM's high predictive discrimination at each internal node (Figure S4c). Overall, our selected HSM configuration balances high precision with meaningful depth on both validation and OOD sets.

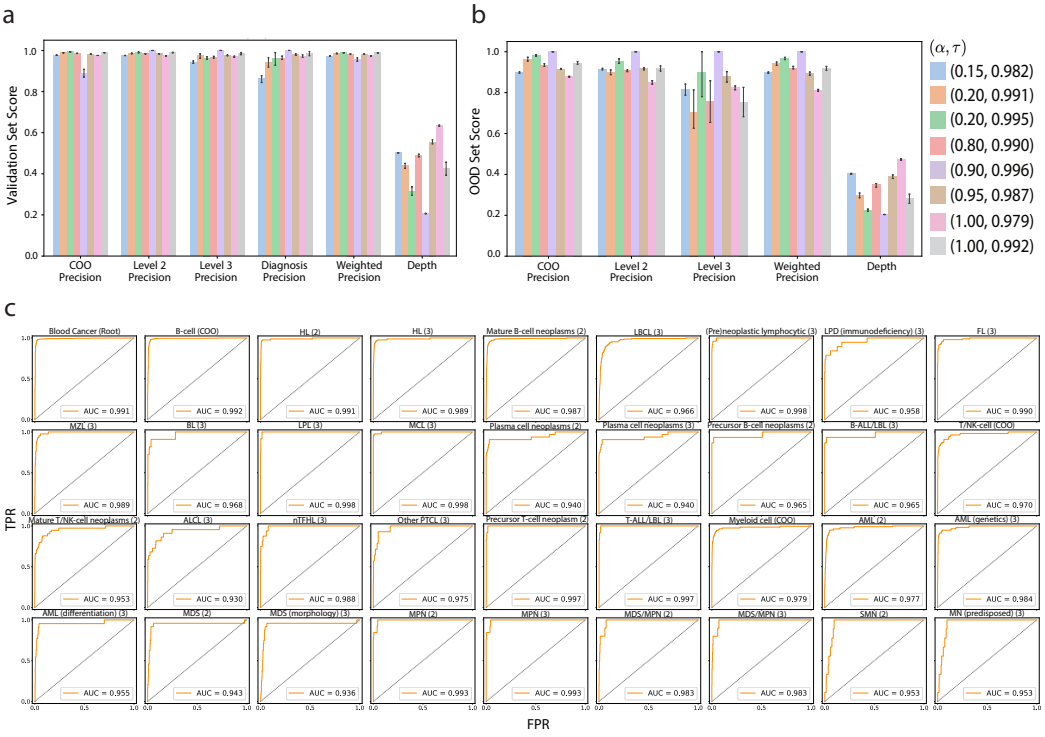

Figure S4: (a) Candidate models' (a) validation and (b) OOD performance metrics. (c) Selected HSM's micro-averaged validation ROC curves at each internal node of the taxonomy.

A limitation of the selected HSM is the relatively low rate of diagnosis-level predictions, driven by a high fraction of higher-level outputs (Figures 2c, S4a-b). While this was expected given our

conservative thresholding, future work would focus on improving prediction depth. First, we will enhance DAMEs by incorporating additional data-types and test alternative embedding strategies such as using a GNN integration head in our source model as previously discussed. Another idea is to train a contrastive model that creates a shared embedding between the multi-omic data and subtype-specific textual descriptions potentially yielding richer, more discriminative embeddings. Second, we will explore other hierarchical classification methods. Because hierarchical classifications form a taxonomic path, we can cast prediction as a sequence generation problem and evaluate transformers or other attention-based models that condition each child decision on prior taxonomic choices [46].

5.3.3 DeepHit

DeepHit is a neural-network–based survival model that directly learns the joint distribution of event occurrence and time through discrete-time likelihood estimation. In our implementation, we used `DeepHitSingle` from the `pycox` library to model PFS and OS [15, 47]. Our DeepHit models output a discrete probability mass function over T pre-specified time intervals (0 to 20 years with a step size of 1 month) corresponding to the conditional event-time distribution $P(T = t|X)$.

We performed 100 trials of BO to identify optimal architectures and hyperparameters for both DeepHit models' MLP backbones. We optimized the following parameters: number of hidden layers, number of nodes per layer, regularization, learning rate, weight decay, batch size, and the temperature-scaling factor applied to the source model's 34-dimensional vector of class scores (Table S4).

Table S4: Hyperparameters for DeepHit models

DeepHit OS

Component	Value
Architecture	
Input feature dimension	290
Hidden layers	4 (Linear)
Output embedding dim	32
Activation function	ReLU
Regularization	
Dropout rate	0.35
Training	
Logits temperature scaling	0.25
Optimizer	Adam
Learning rate	1.32×10^{-3}
Weight decay (L2)	6.75×10^{-5}
Batch size	128
BO rounds	100
Max epochs per trial	1000
Early stopping patience	5
DeepHit loss function α	0.2

DeepHit PFS

Component	Value
Architecture	
Input feature dimension	290
Hidden layers	3 (Linear)
Output dim	64
Activation function	ReLU
Regularization	
Dropout rate	0.5
Training	
Logits temperature scaling	2
Optimizer	Adam
Learning rate	3.86×10^{-4}
Weight decay (L2)	5.80×10^{-5}
Batch size	128
BO rounds	100
Max epochs per trial	1000
Early stopping patience	5
DeepHit loss function α	0.2

Comparison to alternative models We compared DAME-trained DeepHit models to alternative survival models and different feature sets to test if DAMEs optimally capture prognostic signal.

1. Cox proportional hazards: age at diagnosis
2. DeepHit: 1-hot encoded ID diagnosis labels (Figure S5a)
3. DeepHit: multi-omic data passed into the source model (Figure S5b)
4. DeepHit: learned features from the integration head's penultimate layer (Figure S5c)

To ensure fair comparison, all above DeepHit models were optimized with the same BO protocol used for the DAME-trained DeepHit models. The Cox model was trained using the `CoxPHFitter` method from the `lifelines` library [48]. All comparison models were trained and evaluated using the same train/validation/test split used for BLOOM source and target model development.

For both PFS and OS, the DAME-trained DeepHit models significantly outperformed the age at diagnosis Cox, the DeepHit diagnosis, and the genomics DeepHit models on the validation and test sets (p<1e-3) (Figures S5c–d). The DAME-trained DeepHit models also significantly outperformed the age at diagnosis Cox and the genomics DeepHit models on the OOD set(p<0.05). We did not evaluate OOD for the diagnosis-label models because its ID only inputs are undefined for OOD samples. The DAME-trained DeepHit model achieved significantly higher validation c-indices than the DeepHit model trained on the integration head's learned features for both PFS and OS (p<0.05).

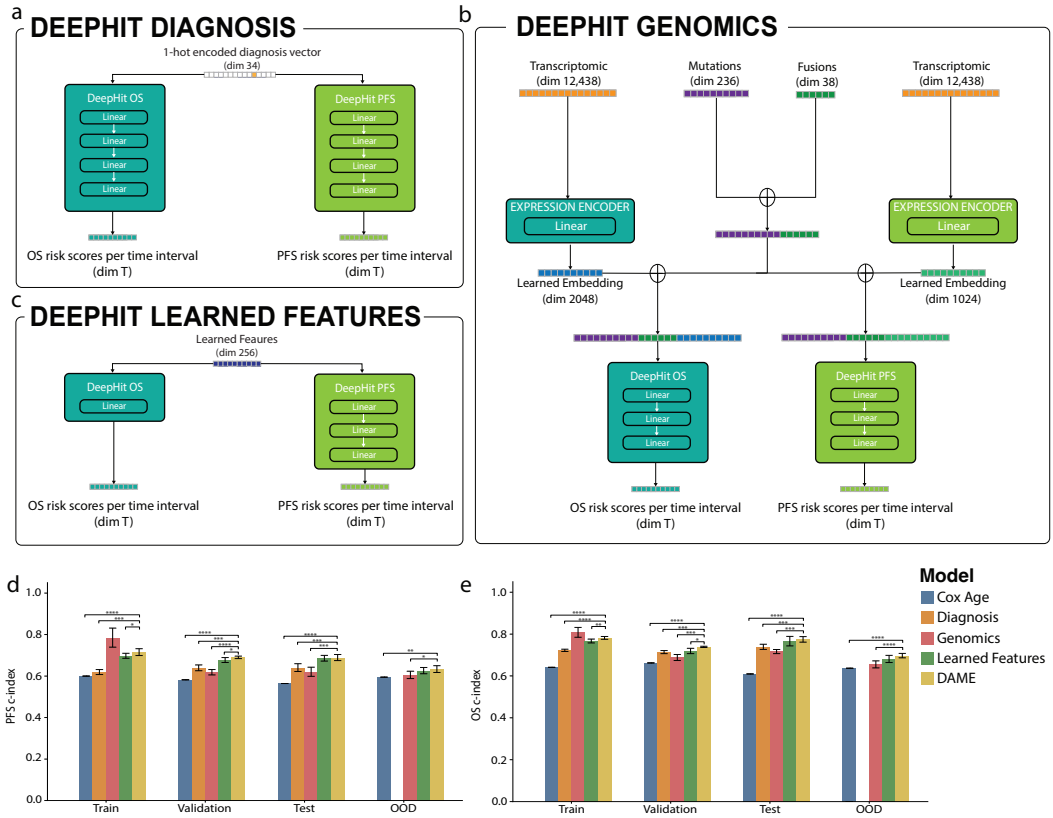

Figure S5: (a) Diagnosis-label DeepHit models. (b) Genomic data DeepHit models. Note that the expression encoder and DeepHit model are jointly optimized during training. (c) Learned feature DeepHit models. (d) PFS and (e) OS DeepHit model performance comparisons. * indicates p<0.05, ** indicates p<0.01, *** indicates p<1e-3, and **** indicates p<1e-4.

DAMEs captured prognostic-relevant information more effectively than age at diagnosis, a key clinical baseline. DAME-trained DeepHit models' outperformance of diagnosis-labeled DeepHit models demonstrates that despite being diagnosis-mapped, DAMEs effectively embed multi-omic signals associated with patient risk even within diagnoses. Additionally, their improvement over models trained only on the integration head's learned features validates that the class scores carry prognostic signal. DAME-trained DeepHit models also surpassed models trained on the complete genomic feature space likely because DAMEs effectively condense the high-dimensional multi-omic features. This is evidenced by the notably higher train set performance of DeepHit models trained on all multi-omic features relative to their poorer validation, test, and OOD results, suggesting the DAME-trained models are less overfit as a result of their focus on the strongest, most biologically relevant signals, which improves their generalizability.

5.4 Leave-One-Hospital-Out Cross-Validation

To assess whether source, HSM, and DeepHit model performance is robust to hospital-specific technical variation, we employed a LOHOCV strategy across the 28 institutions represented by our ID cohort (Figure S6a). In each iteration, all ID samples from one institution formed the hold-out test set while samples from the remaining institutions were randomly split into 80/20 train/validation

sets and trained using our previously described protocols. We computed performance on the internal validation and held-out test sets for every fold and report their averaged values (Figures S6b-c).

Clinical data We previously excluded clinical data because reporting can vary between institutions, potentially leading to institutional-dependent, non-random missingness that compounds batch effects. However, given their known importance, we repeated LOHOCV after appending a small set of low-missingness demographic features (sex, race, ethnicity, and age at diagnosis) to DAMEs. The source model received no demographic inputs because these features' relationship with diagnosis is vastly different from their association to prognosis opting us to directly expose the target models to the clinical data. Categorical clinical features were one-hot encoded, missing age values were imputed from the fold's train set mean, and age data was standardized to the train set distribution. Each feature had a dummy variable indicating missingness. All these preprocessing steps were performed on the training portion of each LOHOCV fold and applied to the validation and hold-out test sets to avoid leakage.

Computational budget and inclusion criteria To evaluate the generalizability of fixed architectures and minimize computational cost, we did not run BO on the source and target models within each fold instead retaining our previously described model architectures (Tables S1,S3-4). For consistent estimates, we excluded 12 institutions with fewer than 50 cases with PFS labels from our reported average hold-out test set performance metrics.

Results Across institutions, validation and hold-out test performance were closely matched for the source and target models trained both with and without clinical data, indicating good generalizability across sites. Target models trained with appended demographics did not materially outperform DAME-only models. This likely reflects the strong baseline established by DAMEs, our use of a narrow demographic feature set prioritized for low missingness rather than the full clinical space, and architectures and hyperparameters tuned for DAMEs alone, which may be suboptimal for the combined feature space.

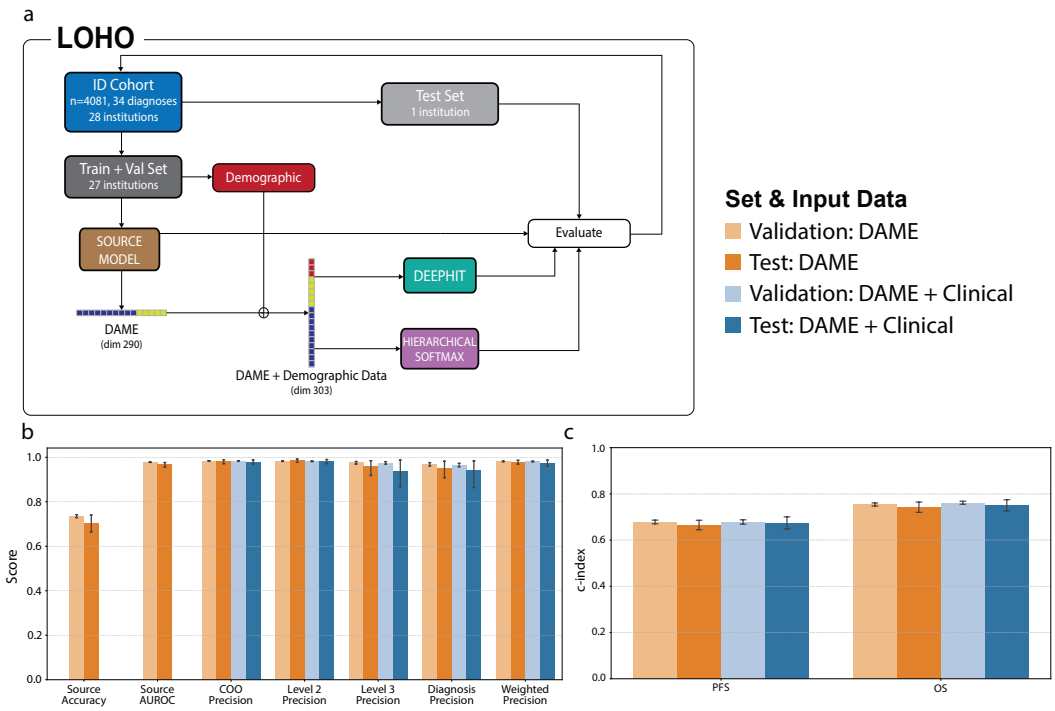

Figure S6: (a) LOHOCV process with clinical data. (b) Source model and HSM LOHOCV performance metrics. Note that AUROC refers to macro-averaged AUROC (c) PFS and OS DeepHit models' LOHOCV c-indices.

5.5 Compute Resources

For model development, BO studies were run on the Google Cloud Platform using a single `n1-highcpu-96` virtual machine (96 vCPUs, 86.4 GB RAM). This machine type was chosen to maximize parallel CPU throughput for `Optuna`'s evaluator. The main computational load was from performing HSM BO in which we performed a BO study for each of the 20 α hyperparameters evaluated. During this, we sharded studies by hyperparameter α launching one independent worker per α. With one study per each of the 20 α parameters evaluated, and 1 thread per trial, memory use still remained well below the 86.4 GB ceiling, as our per-trial footprint was modest with tabular/tensor batches kept in RAM.

5.6 Significance Tests

Categorical association test significance values were evaluated using 2-tailed Fisher's exact tests. Significance of KM curve stratification was evaluated using 2-tailed multivariate log-rank significance tests. 95% confidence intervals comparing model performance were calculated using a t-distribution.

5.7 Data Visualization

t-SNE plots were generated using the `TSNE` method from `sklearn` with perplexity parameters of 50 [41]. KM curves were generated using the `KaplanMeierFitter` and `add_at_risk_counts` methods from the `lifelines` library [48]. Plots were generated using the `matplotlib` and `seaborn` Python libraries except for the sunburst plot and oncoprint which were generated with the `plotly` Python library and the `ComplexHeatmap` R package's `oncoPrint` function respectively [49–52].

Table S5: Description of hierarchy level 3 categories

Category	N_{ID} (%)	N_{OOD} (%)
Large B-cell lymphomas (LBCL)	840 (80%)	216 (20%)
Follicular lymphomas (FL)	557 (97%)	16 (3%)
Hodgkin lymphomas (HL)	479 (94%)	29 (6%)
AML with defining genetic abnormalities	316 (76%)	101 (24%)
Marginal zone lymphomas (MZL)	237 (92%)	21 (8%)
Mantle cell lymphoma (MCL)	231 (99%)	2 (1%)
Plasma cell neoplasms	117 (80%)	45 (20%)
B-lymphoblastic leukemias/lymphomas (B-ALL/LBL)	83 (39%)	128 (61%)
AML defined by differentiation	123 (60%)	82 (40%)
Myeloproliferative neoplasms (MPN)	72 (40%)	109 (60%)
Myelodysplastic neoplasms (MDS) defined morphologically	133 (79%)	36 (21%)
(Pre)neoplastic small lymphocytic proliferations	138 (98%)	3 (2%)
Lymphoid proliferations (LPD) from immune deficiency	105 (79%)	28 (21%)
Anaplastic Large-cell Lymphomas (ALCL)	120 (96%)	5 (4%)
T-lymphoblastic leukemia/lymphomas (T-ALL/LBL)	64 (63%)	38 (37%)
Nodal T follicular helper cell lymphomas (nTFHL)	90 (96%)	4 (4%)
Myelodysplastic/myeloproliferative neoplasms (MDS/MPN)	58 (72%)	22 (28%)
Other peripheral T-cell lymphomas (Other PTCL)	77 (100%)	0 (0%)
Primary cutaneous T-cell LPD and lymphomas (CTCL)	0 (0%)	74 (100%)
Lymphoplasmacytic lymphomas (LPL)	67 (100%)	0 (0%)
Prediscposed myeloid neoplasms (MN(predisposed))	52 (83%)	11 (17%)
Burkitt lymphomas (BL)	62 (100%)	0 (0%)
Splenic B-cell lymphomas and leukemias (SBL)	0 (0%)	61 (100%)
Plasmacytoid Dendritic Neoplasms (pDC neoplasms)	0 (0%)	49 (100%)
Tumor-like lesions with B-cell predominance	0 (0%)	46 (100%)
EBV+ T/NK-cell lymphomas	0 (0%)	42 (100%)
Histiocytic neoplasms	0 (0%)	32 (100%)
Mature T/NK-cell leukemias	0 (0%)	30 (100%)
ALAL defined immunophenotypically	0 (0%)	26 (100%)
Cutaneous follicle center lymphomas (pcFCL)	0 (0%)	16 (100%)
Intestinal T/NK-cell LPD	0 (0%)	15 (100%)
Follicular dendritic cell neoplasms (FDCS)	0 (0%)	15 (100%)
Langerhans cell neoplasms (LC neoplasm)	0 (0%)	14 (100%)
MDS with genetic abnormalities	0 (0%)	13 (100%)
Mastocytosis	0 (0%)	13 (100%)
Myeloid sarcomas (MS)	0 (0%)	11 (100%)
Eosinophilic MLN (MLN-Eo) with tyrosine kinase (TK) fusions	0 (0%)	10 (100%)
ALAL with genetic abnormalities	0 (0%)	8 (100%)
Monoclonal gammopathies (MG)	0 (0%)	7 (100%)
HVV8+ B-cell LPD	0 (0%)	6 (100%)
Hepatosplenic T-cell lymphoma (HSTCL)	0 (0%)	4 (100%)
Transformations of indolent B-cell lymphoma	0 (0%)	3 (100%)
Monoclonal immunoglobulin deposition diseases (MIDD)	0 (0%)	2 (100%)
Other dendritic cell neoplasms	0 (0%)	1 (100%)
Childhood EBV+ T/NK-cell LPD	0 (0%)	1 (100%)

Table S6: Description of ID diagnoses

Category	N
Follicular lymphoma (FL)	557
Diffuse large B-cell lymphoma, NOS (DLBCL, NOS)	289
Mantle cell lymphoma (MCL)	231
Nodular sclerosis classical Hodgkin lymphoma (cHL (NS))	221
Germinal center B-cell type DLBCL, NOS (DLBCL, NOS (GCB))	207
Nodular lymphocyte predominant Hodgkin lymphoma (NLPHL)	179
Plasma cell myeloma (MM)	177
Extranodal marginal zone B-cell lymphoma (EMZL)	174
Chronic lymphocytic leukemia/small lymphoctyic lymphoma (CLL/SLL)	138
Activated B-cell type DLBCL, NOS (DLBCL, NOS (ABC))	136
AML with myelodysplasia-related changes (AML (MDS changes))	135
Monomorphic post-transplant LPD (PTLD (monomorphic))	105
Acute promyelocytic leukemia (APL (PML-RARA))	98
Primary CNS DLBCL (PCNSL)	93
Angioimmunoblastic T-cell lymphoma (AITL)	90
B-lymphoblastic leukemia/lymphoma, NOS (B-ALL/LBL, NOS)	83
AML with NPM1 mutations (AML (NPM1-mut))	83
MDS with excess blasts (MDS-EB)	83
Mixed cellularity classical Hodgkin lymphoma (cHL (MC))	79
Peripheral T-cell lymphoma, NOS (PTCL, NOS)	77
Chronic myelogeneous lymphoma with BCR-ABL1 (CML (BCR-ABL1))	72
M2 AML, NOS with maturation (AML, NOS (M2))	72
Lymphoplasmacytic lymphoma, Waldenstrom macroglobulinemia (LPL (WM))	67
T-lymphoblastic leukemia/lymphoma (T-ALL/LBL)	64
Nodal marginal zone lymphoma (NMZL)	63
Burkitt lymphoma (BL)	62
Anaplastic large cell lymphoma, ALK+ (ALCL, ALK+)	61
High grade B-cell lymphoma with MYC, BCL2, and/or BCL6 (HGBCL-DH/TH)	61
Anaplastic large cell lymphoma, ALK- (ALCL, ALK-)	59
Chronic myelomonocytic leukemia (CMML)	58
Plasmablastic lymphoma (PBL)	54
Therapy-related myeloid neoplasms (therapy-related MN)	52
M1 AML, NOS without maturation (AML, NOS (M1))	51
Myelodysplastic syndrome with multilineage dysplasia (MDS-MLD)	50

6 Acknowledgements

We thank the current and past members of the Sandeep Dave Lab, especially Ayush Batra, Fadzai Chinyengetere, Tushar Dave, Lanie Happ, Rachel Kositsky, Cassandra Love, Dennis Owusu, Razvan Panea, Jessi Rodgers, Veronica Russell, Jennifer Shingleton, Devang Thakkar, and Shari Tian for their work in sample processing and bioinformatic pipeline development. We also thank all collaborators in the ABCG consortium for providing and clinically reviewing samples; Katherine Dura of the Bill Majoros Lab for her contribution to the variant detection pipeline; and Nathaniel Blalock, Coban Brooks, Benjamin Perry, and Srinath Seshadri of the Philip Romero Lab for their advice in model development.

References

[1] Epidemiology National Cancer Institute, Surveillance and End Results (SEER) Program. Seer cancer stat facts. `https://seer.cancer.gov/statfacts/`, 2025.

[2] American Cancer Society. Cancer facts & statistics. `https://www.cancer.org/research/cancer-facts-statistics.html`, 2025.

[3] Weijie Li. The 5th edition of the world health organization classification of hematolymphoid tumors. *Leukemia [Internet]*, 2022. doi: doi:10.36255/exon-publications-leukemia-who-5th-edition-hematolymphoid-tumors.

[4] Ibraheem Hamdi, Hosam El-Gendy, Ahmed Sharshar, Mohamed Saeed, Muhammad Ridzuan, Shahrukh K. Hashmi, Naveed Syed, Imran Mirza, Shakir Hussain, Amira Mahmoud Abdalla, and Mohammad Yaqub. Breaking down the hierarchy: A new approach to leukemia classification. *arXiv*, page 104–113, 2023. doi: 10.1007/978-3-031-47076-9_11.

[5] Mahwish Ilyas, Muhammad Ramzan, Mohamed Deriche, Khalid Mahmood, and Anam Naz. An efficient leukemia prediction method using machine learning and deep learning with selected features. *PLOS ONE*, 20(5):e0320669, 2025. doi: 10.1371/journal.pone.0320669.

[6] Jan-Niklas Eckardt, Jan Moritz Middeke, Sebastian Riechert, Tim Schmittmann, Anas Shekh Sulaiman, Michael Kramer, Katja Sockel, Frank Kroschinsky, Ulrich Schuler, Johannes Schetelig, Christoph Röllig, Christian Thiede, Karsten Wendt, and Martin Bornhäuser. Deep learning detects acute myeloid leukemia and predicts npm1 mutation status from bone marrow smears. *Leukemia*, 36:111–118, 2022. doi: 10.1038/s41375-021-01408-w.

[7] Edian F Franco, Pratip Rana, Aline Cruz, Víctor V Calderón, Vasco Azevedo, Rommel T. J. Ramos, and Preetam Ghosh. Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. *Cancers*, 13(9):2013, 2021. doi: 10.3390/cancers13092013.

[8] Madison Darmofal, Shalabh Suman, Gurnit Atwal, Michael Toomey, Jie-Fu Chen, Jason C. Chang, Efsevia Vakiani, Anna M. Varghese, Anoop Balakrishnan Rema, Aijazuddin Syed, Nikolaus Schultz, Michael F. Berger, and Quaid Morris. Deep learning model for tumor type prediction from ngs panels. *Cancer Discovery*, 14(6):1064–, 2024. doi: 10.1158/2159-8290.CD-23-0996.

[9] L.O. Moraes, C.E. Pedreira, S. Barrena, A. Lopez, and A. Orfao. A decision-tree approach for the differential diagnosis of chronic lymphoid leukemias and peripheral b-cell lymphomas. *Computer Methods and Programs in Biomedicine*, 178:85–90, 2019. doi: 10.1016/j.cmpb.2019.06.014.

[10] Krzysztof Mrózek, Jessica Kohlschmidt, James S. Blachly, Deedra Nicolet, Andrew J. Carroll, Kellie J. Archer, Alice S. Mims, Karilyn T. Larkin, Shelley Orwick, Christopher C. Oakes, Jonathan E. Kolitz andBayard L. Powell, William G. Blum, Guido Marcucci, Maria R. Baer andGeoffrey L. Uy, Wendy Stock, John C. Byrd, and Ann-Kathrin Eisfeld. Outcome prediction by the 2022 european leukemianet genetic-risk classification for adults with acute myeloid leukemia: an alliance study. *Leukemia*, 37:788–798, 2023. doi: 10.1038/s41375-023-01846-8.

[11] Joaquim Carreras, Naoya Nakamura, and Rifat Hamoudi. Artificial intelligence analysis of gene expression predicted the overall survival of mantle cell lymphoma and a large pan-cancer series. *Healthcare*, 10(1):155, 2022. doi: 10.3390/healthcare10010155.

[12] Adrian Mosquera Orgueira, Marta Sonia González Pérez, Jose Diaz Arias, Laura Rosiñol, Albert Oriol, Ana Isabel Teruel, Joaquin Martinez Lopez, Luis Palomera, Miguel Granell, Maria Jesus Blanchard, Javier de la Rubia, Ana López de la Guia, Rafael Rios, Anna Sureda, Miguel Teodoro Hernandez, Enrique Bengoechea, María José Calasanz, Norma Gutierrez, Maria Luis Martin, Joan Blade, Juan-Jose Lahuerta, Jesús San Miguel, Maria Victoria Mateos, and the PETHEMA/GEM Cooperative Group. Unsupervised machine learning improves risk stratification in newly diagnosed multiple myeloma: an analysis of the spanish myeloma group. *Blood Cancer Journal*, 12:76, 2022. doi: 10.1038/s41408-022-00647-z.

[13] Devang Thakkar, Brian T. Hill, Rachel Kositsky, Shari Tian, Leonardo Biral, Veronica Russell, Tushar Dave, Cassandra L. Love, Caroline Roth, Matthew Stuart McKinney, Ahmed Galal, Jadee Neff, Agrima Mian, Ellen Kendall, Sarah L. Ondrejka, Matthew Chiaramonte, Govind Bhagat, Kenneth N. Ofori, Ran Reshef, Alexandra E. Kovach, Tarsheen Sethi, Emily F. Mason, Shakthi Bhaskar, Olalekan O. Oluwole, Chad McCall, Christopher Pallas, Nilanjan Ghosh, Robert Ferdman, George Chen, Francisco Hernandez-Ilizaliturri, Joanna Zurko, Ashley Cunningham, Nirav C. Shah, Boyu Hu, Deborah M. Stephens, Monalisa Ghosh, Neil Bailey, Krish Patel,

John M. Pagel, Kavya Kannamma Kannan, Eric D. Hsi, Rakhee Vaidya, Andrew Ip, Andre Goy, Swetha Kambhampati, Robert Ohgami, Charalambos Andreadis, Christian A. Gordillo, Brianna Just, Jonathon B. Cohen, Arielle Baim, Julie Barbello, Erika Cavallone, Veronika Bachanova, Maureen Laschen, Andinet Teferra, Frederique Lorcy, Sylvain Lamure, Guillaume Cartron, Valerie Dardalhon, Kikkeri N. Naresh, Magdalena Czader, Anupama Reddy, Elizabeth Thacker, Clayton Parker, Lanie Happ, and Sandeep Dave. Molecular and clinical determinants of car-t therapy response in dlbcl. *Blood*, 144:3409–3409, nov 2024.

[14] Cassandra Love, Raju Pillai, Sarah L. Ondrejka, Govind Bhagat, Amy Chadburn, Matthew McKinney, Jean L. Koff, Dina Sameh Soliman, Magdalena Czader, Jr. Abner Louissaint, Shaoying Li, Choon Kiat Ong, Amir Behdad, Andrew M. Evens, Yasodha Natkunam, Peter H. Norgaard, Sirpa Leppa, Eric Tse, Jennifer R. Chapman, Catalina Amador, Yuri Fedoriw, Agata M. Bogusz, Andrew G Evans, Rashmi S. Goswami, Ridas Juskevicius, Mina L. Xu, Kikkeri N. Naresh, Barbara Xiong, Adam Snowden, Anabel Thurman, Eileen Smith, Tushar Dave, Rachel Kositsky, Devang Thakkar, Veronica Russell, Caroline J. Roth, and Sandeep Dave. The atlas of blood cancer genomes (abcg) project: A comprehensive molecular characterization of leukemias and lymphomas. *Blood*, 138:2213, 2021. doi: https://doi.org/10.1182/blood-2021-151346.

[15] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[17] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

[18] Shannon A. Carty. Biological insights into the role of TET2 in T cell lymphomas. *Frontiers in Oncology*, 13:1199108, 2023. doi: 10.3389/fonc.2023.1199108.

[19] Meaghan M. Ryan. Acute promyelocytic leukemia: A summary. *Journal of the Advanced Practitioner in Oncology*, 9(2):178–187, mar 2018. doi: 10.6004/jadpro.2018.9.2.6.

[20] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010. doi: 10.1093/nar/gkp1137.

[21] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014. doi: 10.1093/bioinformatics/btu170.

[22] Donald Freed, Rafael Aldana, Jessica A. Weber, and Jeremy S. Edwards. The sentieon genomics tools - a fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv*, 2017. doi: 10.1101/115717.

[23] Picard. `https://github.com/havakv/pycox`.

[24] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009. doi: 10.1093/bioinformatics/btp352.

[25] Christopher T. Saunders, Wendy S. W. Wong, Sajani Swamy, Jennifer Becq, Lisa J. Murray, and R. Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012. doi: 10.1093/bioinformatics/bts271.

[26] Ryan Poplin, Pi-Chuan Chang, David Alexander, Stephen Schweighauser, Justin Duan, Cheng Lai, Alistair McIntyre, Jun Ding, Natalie Leibovich, Mark A. DePristo, Eric Banks, Anton Korobeynikov, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, 2018. doi: 10.1038/nbt.4235.

[27] Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alfoldi, Qianyi Wang, Andrea Ganna, Daniel P. Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020. doi: 10.1038/s41586-020-2308-7.

[28] Bjoern Chapuy, Chip Stewart, Andrew J. Dunford, Jaegil Kim, Angel Kamburov, Robert A. Redd, et al. Molecular subtypes of diffuse large b-cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature Medicine*, 24(5):679–690, 2018. doi: 10.1038/s41591-018-0016-8.

[29] Anupama Reddy, Jenny Zhang, Nicholas S. Davis, Andrea B. Moffitt, Cory L. Love, Andrew Waldrop, et al. Genetic and functional drivers of diffuse large b cell lymphoma. *Cell*, 171(2):481–494.e15, 2017. doi: 10.1016/j.cell.2017.09.027.

[30] George W. Wright, David W. Huang, John D. Phelan, Sarah E. Coupland, Hartmut Koeppen, Sunita Nayar, et al. A probabilistic classification tool for genetic subtypes of diffuse large b cell lymphoma with therapeutic implications. *Cancer Cell*, 37(4):551–568.e14, 2020. doi: 10.1016/j.ccell.2020.03.015.

[31] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jasmin Drenkow, Chris Zaleski, Sonali Jha, Jorg Battey, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. doi: 10.1093/bioinformatics/bts635.

[32] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010. doi: 10.1101/gr.107524.110.

[33] Rob Patro, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017. doi: 10.1038/nmeth.4197.

[34] Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4:1521, 2015. doi: 10.12688/f1000research.7563.2.

[35] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014. doi: 10.1186/s13059-014-0550-8.

[36] Sebastian Uhrig, Juliane Ellermann, Tobias Walther, Paul Burkhardt, Michael Fröhlich, Bernhard Hutter, and et al. Accurate and efficient detection of gene fusions from rna sequencing data. *Genome Research*, 31(3):448–460, March 2021. doi: 10.1101/gr.257246.119.

[37] Felix Mitelman, Bertil Johansson, and Fredrik Mertens. Mitelman database of chromosome aberrations and gene fusions in cancer. `https://mitelmandatabase.isb-cgc.org/`, 2025. Database last updated July 10, 2025.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. URL `http://arxiv.org/abs/1912.01703`.

[39] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3292500.3330701.

[40] Grzegorz S. Nowakowski and Myron S. Czuczman. ABC, GCB, and double-hit diffuse large B-cell lymphoma: Does subtype make a difference in therapy selection? *American Society of Clinical Oncology Educational Book*, 35:e449–e457, 2015. doi: 10.14694/EdBook_AM.2015.35.e449.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python, 2011.

[42] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. URL `http://arxiv.org/abs/1603.02754`.

[43] Yuning Hao, Ming Poon, Yan, Blake R. Heath, Yu L. Lei, and Yuying Xie. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS Comput Biol.*, 15(5), 2019. doi: 10.1371/journal.pcbi.1006976.

[44] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv*, 2021.

[45] Randolph Linderman, Jingyang Zhang, Nathan Inkawhich, Hai Li, and Yiran Chen. Fine-grain inference on out-of-distribution data with hierarchical classification. *arXiv*, 2022. doi: arXiv:2209.04493.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

[47] Havard Kvamme. pycox: Deep survival analysis with pytorch. `https://github.com/havakv/pycox`, 2019.

[48] Cameron Davidson-Pilon. lifelines: survival analysis in python, 2019. URL `https://doi.org/10.21105/joss.01317`.

[49] J. D. Hunter. Matplotlib: A 2d graphics environment, 2007.

[50] Michael L. Waskom. seaborn: statistical data visualization, 2021. URL `https://doi.org/10.21105/joss.03021`.

[51] Plotly Technologies Inc. Collaborative data science, 2015. URL `https://plot.ly`.

[52] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016. doi: 10.1093/bioinformatics/btw313.