Assessing under-specification in cancer predictive Multi-attention models trained on gene expression data..

Author_1 (Ahmed Ech-CHERIF) 1,2,*,¶ , Tarek BENAMEUR 2,¶ , Ossman El Wassila Idriss 2 Houria ECH-CHERIF 4

Abstract

Machine learning models, particularly transformer-based architectures, have shown strong performance in cancer classification tasks using gene expression data. However, their clinical deployment remains challenging due to concerns about reliability and generalization. In this study, we evaluate the robustness of transformer models trained on gene expression profiles from 33 cancer types

from a large-scale national oncology genomics program ". While these models achieve high accuracy on internal validation sets, we identify substantial performance variability across random seeds and a significant decline in accuracy on external, out-of-distribution (OOD) test cohorts drawn from GEO datasets. This phenomenon—where models trained under identical conditions yield divergent outcomes—indicates a critical limitation known as underspecification. We systematically assess this issue using seed variance analysis, sensitivity to label uncertainty, and performance under dataset shift. Our findings reveal that standard training pipelines may produce models that underperform outside controlled test environments. To explore possible remedies, we evaluate strategies such as ensemble learning, uncertainty quantification, and robust training objectives that may help mitigate variability. These results have important implications

for the development of trustworthy machine learning tools in clinical oncology and emphasize the need for rigorous evaluation beyond internal accuracy metrics

Keywords: Transformer models, Cancer classification, Gene expression, Machine learning robustness, Out-of-distribution generalization, Clinical decision support, Biomedical AI

Introduction

Background on Gene Expression in Oncology

Simultaneously, measuring the expression levels of many genes such as oncogenes, is known to capture the functions of the cells being examined [1,2], and any dysregulation of the latter is reflected in the former, which may indicate the presence of some specific diseases such as certain types of cancer. In clinical decision support systems [3,4,5,6], observed gene expression levels provide important clues to the pathologist into patient prognosis and eventually, her response to therapy [7,4,5,6]. Owing to the widespread use of Next Generation Sequencing (NGS) platforms and the relatively low cost of performing transcriptomic analysis [66], numerous experimental studies have been carried out in various biological and medical domains targeting many species [8,9,10], which resulted in considerable volumes of data relating various types of phenotypes such as cancer, to some large subsets of genes expression levels [11,12,13]. Some of these data sets have been further curated and made readily available in dedicated web portals [14,15,16,17].

Rise of Deep Learning and Transformer Models in Biomedicine

Following the recent remarkable success of Deep Learning (DL) models in diverse tasks ranging from visual perception to acoustic modeling and linguistic inference, to name a few, a plethora of DL models, trained on both whole slide images (WSI) [18] and micro-array gene expression data have been proposed to predict various phenotypes, particularly cancer types, and very encouraging results on independent test sets have been reported [19,20,21]. During training of such models, the model with the lowest leave-out error, akin validation error, is selected to avoid overfitting i.e., poor generalization on novel examples [22]. Due to the sheer number of parameters, particularly in deep learning pipelines, which is in the order of millions, and the hyper-parameters such as the network architecture, learning rate, random seed, etc., which must be determined during training, models with the same validation error have been found to behave differently when used to predict a phenotype [23]. This is due to the violation of the independently and identically distributed (i.i.d) assumption, which does not hold in practice [24,25,26,27,28]. This phenomenon is known as under-specification in the computer science parlance and has been extensively investigated in [25,29]. Moreover, optimized, and fine-tuned models particularly, the most prominent convolutional neural networks (CNN) in computer vision applications, have been found to be notoriously unstable in the sense that a small perturbation in the input image, unnoticeable to the human eye, results in erroneous prediction[50]. Numerous research works have focused on devising countermeasures to enhance robustness of deep learning models [24,25,26,27,28,30,33,32]. However, notwithstanding these efforts, there is strong mathematical evidence, which shows that deep learning models with high accuracy are

generally unstable, although accurate and stable models exist, but modern deep learning pipelines fail to select such models [3].

Due to its inherent reliability, wide acceptance and trustworthiness in the medical domain, microarray gene expression data has been extensively used in oncology studies to classify tumor types [35], and in precision medicine to provide personalized treatments of cancer [36]. Moreover, the latter data present four main important challenges to the above-mentioned ML methods: the first challenge pertains to its nonlinear separability due to the numerous interactions among the genes involved in the model, which rules out linear learning machines such as multi-class linear support vector machines [37]. The second challenge is the relatively low volume of the training dataset compared to the very large number of the genes involved, making it very hard for deep learning methods to learn a useful model, as these models require very large training data sets due to the sheer number of parameters namely, the network weights [38] . The third challenge concerns the difficulty of generating labeled artificial examples, which is a common practice in deep learning to avoid over-fitting[39].

ML models, which deal better with the above-mentioned challenges of micro-array gene expression data, have been the subject of intensive investigations. In this respect, Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), etc. have been trained on micro-array gene expression data and were found to yield comparable accuracy on independent test data[40,41]. Motivated by the success of CNN on image data, deep learning models and particularly, 1d-CNN and 2-d CNN [41,42] were also shown to produce good models that predict cancer type from micro array gene expression data.

However, the latter models were not analyzed in terms of robustness nor how their test accuracy drops when deployed on test data, which is not i.i.d.

To the best of our knowledge, and despite the plethora of machine learning models proposed in the medical domain along with their impressive accuracy rates, none of them has been approved by the FDA as a triaging tool [67].

Very recently, self-attention-based models, also known as transformers , which model better the interactions among the input features, as compared to standard deep learning models [55], have been also proposed and tested for the prediction of cancer types from micro array gene expression data [56]. They also address the transparency issues inherent to deep learning systems namely, – black box models - to some extent and may be a first step towards explainable AI systems as they harness the attention mechanism in their prediction phase.

Underspecification in Machine Learning Pipelines

Under specification in Machine Learning Pipelines

“An underspecified model pipeline is one where many classifiers can achieve equivalent performance on the training and validation data yet differ in terms of generalization.” [39,43]. under specification seriously hampers the introduction of models selected by standard pipelines into practice. Assessing the under specification negative effects of attention-based transformers models and attempting to mitigate its negative effects namely, the drop in test accuracy caused by distribution shifts, especially, in cancer predictive models is an important first step towards the adoption of robust predictive

models in daily medical practice. Despite their great promise to accurately model micro-array gene expression data, owing to their ability to efficiently model gene interactions, we provide, in this paper, experimental evidence to show that transformer predictive models, obtained from standard learning pipelines, suffer indeed from the negative effects of under specification as their deep learning counterpart models. We hypothesize that attention-based models trained on i.i.d cancer gene expression data exhibit performance degradation on clinically realistic OOD datasets due to under specification. Results of experiments will be reported and discussed, and some potential strategies for avoiding under specification effects are further discussed n the following sections

Figure 1: Assessment Pipeline

Objectives of This Study.

This study aims to investigate the extent to which underspecification affects predictive modeling in cancer classification using gene expression data. Specifically, we aim to:
(i) evaluate the performance' variability of transformer-based models across multiple random seeds and platforms;
(ii) assess the robustness of predictions under out-of-distribution (OOD) conditions; and
(iii) examine the sensitivity of model outcomes to ground truth label uncertainty.

By systematically analyzing these dimensions, we seek to identify critical limitations in current modeling practices and propose methodological strategies to mitigate under specification in clinical oncology applications.

2. Related Works:

Early Statistical Models

Machine Learning Cancer Prediction has been the subject of active research for the past two decades [44,45,46]. Early cancer prediction efforts during the 1990s, relied primarily on statistical models, using handcrafted features [64], and expert-defined biomarkers, derived from clinical, imaging, and molecular data [47,48]. Models based on logistic regression, such as Cox proportional hazards models [49], and decision trees [44] were commonly used to predict cancer risk and outcomes, especially for breast and prostate cancer. These models required manual selection of features such as age, tumor size, hormone receptor status, and genetic mutations (e.g., BRCA1/2) and often lacked the ability to capture nonlinear or high-dimensional interactions [50].

Classical Machine Learning Approaches (SVMs, k-NN, Decision Trees)

---------- Start----------

The advent of high-throughput technologies especially, microarrays in the early 2000s enabled gene expression-based prediction models [51,52]. Notably, the work by Golub et al. (1999) on leukemia classification using gene expression profiling [51] marked a shift toward data-driven approaches in oncology. This era also saw the rise of support vector machines (SVMs) and k-nearest neighbors (k-NN) classifiers applied to molecular datasets, paving the way for personalized medicine [53,54]. Classification from gene expression data focused on classical machine learning algorithms. Support Vector Machines (SVMs), Decision Trees, and k-Nearest Neighbors (k-NN) were widely used

due to their robustness on small datasets and high-dimensional inputs. SVMs in particular showed strong performance with high-dimensional transcriptomic data, especially when combined with feature selection methods to reduce gene set dimensionality [55]. However, these models typically relied on linear or kernel-based assumptions and were limited in modeling complex gene-gene interactions. Moreover, they were often sensitive to noise and lacked the capacity to scale with larger, modern datasets [56].

Deep Learning and Neural Networks Models

"The application of CNNs to 1D gene expression data marked a turning point in cancer prediction tasks [57]. Indeed, the application of 1D-CNNs to extract features from microarray data, demonstrated performance improvements over SVMs and traditional MLPs [57,58]. The CNN's ability to learn local patterns and nonlinear transformations allowed for better class separation and generalization on test sets. Despite promising results, However, CNN-based models were often treated as black boxes, raising concerns about interpretability and reliability in clinical settings [33,34]. Additionally, these models showed vulnerability to data distribution shifts and adversarial perturbations — a critical issue when deploying models in heterogeneous clinical environments [59,32]. In this study, under specification refers to the phenomenon where multiple models with similarly strong in-distribution (ID) performance make divergent predictions on out-of-distribution (OOD) samples, despite being trained under the same objective. Unlike overfitting, which involves memorizing training data at the expense of generalization, or poor generalization, which reflects weak performance on unseen data, under specification

indicates the presence of multiple plausible model solutions that behave inconsistently on OOD cases [60] .

Transformer Models for Cancer Classification

Motivation for Transformer Models

The success of transformer-based architectures in natural language processing has motivated their extension to high-dimensional biomedical data, including gene expression. Unlike classical ML models that rely on feature selection and predefined interactions, transformers offer the ability to learn complex hierarchical patterns and global dependencies directly from raw data [61,62].

Gene expression profiles present a highly multivariate, non-sequential input space that benefits from the self-attention mechanism's ability to model long-range feature interactions. In cancer classification, such capacity is crucial to capture both local gene-gene co-regulation and higher-level pathway activation signals [63].

2.4.2 Applications of Transformers to Gene Expression Data

Several recent works have explored transformer-based models for cancer classification[61,62,63]:

T-GEM: Transformer for Gene Expression Modeling

T-GEM, one of the first transformer-based models applied directly to gene expression matrices for phenotype prediction across multiple cancer types was introduced in [61]. The model effectively captures nonlinear gene-gene interactions and pathway-level dependencies without requiring extensive feature engineering.

MOT: a Multi-Omics Transformer for multiclass classification tumour types predictions

transformers were used in [64] under the name “MOT” for integrating gene expression, methylation, and copy number data for pan-cancer classification. MOT demonstrated superior performance to CNNs and random forests across TCGA cohorts.

2.5.3 Survival Transformers

Some works have applied transformers for survival prediction tasks, leveraging censored survival data together with gene expression, achieving better hazard prediction than traditional Cox models [65,66].

2.5.4 Advantages of End-to-end learning without handcrafted feature extraction Over Classical Models

End-to-end learning without handcrafted feature extraction has the ability to handle variable gene sets (via masking or embedding), Incorporation of biological prior knowledge (e.g., pathway attention) and Improved OOD robustness (when properly regularized) and thus capturing nonlinear gene-gene interactions at scale [67].

End-to-end transformer-based learning frameworks eliminate the need for handcrafted feature extraction and can flexibly handle variable gene sets through masking or embedding strategies. By incorporating biological prior knowledge (e.g., pathway-guided

attention) and employing robust regularization, these models achieve improved out-of-distribution (OOD) generalization while capturing nonlinear gene–gene interactions and hierarchical biological relationships at scale [67].

2.5.5 Challenges and Open Questions

Despite their promise, transformers for cancer classification face important challenges namely,:

2.5.5.1 Sample size requirements for stable training

Label imbalance in rare cancers, generalization across platforms, batch effects and interpretability of learned attention weights motivate continued research into both model design and data curation for transformer-based oncology models [68]. This work focuses specifically on underspecification in Transformer-based cancer gene classifiers, where the learned representations may vary significantly depending on random seeds or OOD data type (e.g., tissue, platform, or time period), even when ID performance remains stable. Despite some prominent promising results, early models often suffered from limited generalization due to small sample sizes, high dimensionality, and batch effects [69]. These challenges highlighted the need for more robust algorithms and larger, standardized datasets—laying the groundwork for the deep learning and transformer-based approaches used nowadays [71]. Neural Network models have been proposed to

predict the cancer type from gene expression profiles and some quite encouraging results have been obtained [73]. Very recent important advances in ML models have achieved notable progress. However, despite this progression, very few studies have rigorously analyzed how models — especially attention-based ones — perform under distributional shifts common in real-world clinical settings [72]. Our work addresses this gap by evaluating how Transformer classifiers trained on TCGA data generalize under out-of-distribution conditions and proposes a systematic framework for diagnosing and reducing the effects of underspecification.

2.5 Gaps in Current Literature

Despite significant advances in the use of machine learning and deep learning for cancer prediction, several important persistent gaps remain in the current literature which need to be addressed for developing robust, interpretable, and equitable models capable of driving precision oncology forward [74].

Limited Integration of Multimodal Data

While gene expression, imaging, and clinical records have independently shown predictive utility, few studies effectively integrate all three at scale. Most models focus on single-modality data, limiting the ability to capture complex, biologically meaningful patterns across data types [75].

Overreliance on Retrospective Datasets

A large portion of published work relies on retrospective, pre-curated datasets such as TCGA or GEO. These datasets may not reflect real-world clinical variability, introducing biases and reducing the generalizability of the models to prospective cohorts.[76,77,78]

Lack of Transparency in Model Design

Many DL-based models—especially those leveraging transformers or other attention-based architectures—remain black-boxes. The literature often lacks comprehensive interpretability studies, making clinical adoption difficult and hindering biological insight [79].

Underspecification and Model Robustness

Recent research has highlighted the issue of underspecification [80,81]—where multiple models achieve similar validation performance but differ substantially in generalization behavior-. However, few cancer prediction studies systematically address this issue across model classes.

Scarce Benchmarking Across Cancer Types

There is a lack of rigorous, cross-cancer-type benchmarking. Many studies focus on a single cancer type or a few well-studied ones (e.g., breast, lung), leading to unclear conclusions about the generalizability and limitations of the proposed methods.[82]

Rare and Underrepresented Cancers

Most large-scale studies prioritize high-prevalence cancers. Consequently, rare cancers—despite having distinct molecular signatures—remain underexplored, perpetuating health disparities in algorithmic oncology [83].

Insufficient Reporting Standards

The lack of standardized reporting for dataset splits, preprocessing steps, hyperparameters, and evaluation metrics complicates replication efforts and hinders progress in comparative assessments [84]

Our analysis leverages gene expression and clinical metadata from The Cancer Genome Atlas (TCGA), encompassing 32 cancer types. The dataset was obtained from the T-GEM GitHub repository [85] , which provides a preprocessed version of TCGA expression as shown in table 1.

4. Materials and Methods

4.1 Dataset Sources and Preprocessing

We utilized the publicly available T-GEM dataset [85], which aggregates gene expression profiles from 32 cancer types curated from The Cancer Genome Atlas (TCGA). The dataset, originally compiled and benchmarked by Wang et al., was retrieved from the T-GEM GitHub repository [85] . Raw RNA-Seq expression data were log-transformed and z-score normalized within each cancer type to reduce platform-related biases.

4.1.1 Feature Selection: Top Variable Genes

To reduce noise and dimensionality in the gene expression matrix, we applied a feature selection step focused on the most informative genes. Specifically, we selected the top 1,500 most variable genes across all samples in the training set. Variability was quantified using the variance of log-transformed TPM values for each gene.

This unsupervised approach ensures that only genes with the greatest expression heterogeneity—often associated with regulatory activity or disease relevance—are retained. Selecting high-variance genes has been shown to improve classification performance by focusing model attention on biologically meaningful features, while reducing overfitting and computational burden.[86]

This feature selection step was applied before model training and consistently used across all experimental settings, ensuring comparability in both in-distribution and out-of-distribution evaluations

4.2 Transformer Model Architecture

We implemented a Transformer-based model for cancer type classification using gene expression data, inspired by adaptations of attention mechanisms in biomedical domains.

The model follows the encoder architecture originally introduced in [86] adapted to operate over fixed-length gene feature vectors rather than sequences.

4.2.1 Input Projection and Encoding

Each sample's expression profile—represented as a 1,500-dimensional vector of preselected high-variance genes—is first linearly projected to a lower-dimensional embedding space. A learnable input embedding matrix transforms each gene expression value into a fixed-length vector of size d, allowing the network to learn gene-specific positional encodings. Unlike natural language tasks, positional order of genes is not meaningful, so we omit standard sinusoidal encodings.[86]

4.2.2 Multi-Head Attention Layers

We implemented a Transformer-based model for cancer type classification using gene expression data, inspired by adaptations of attention mechanisms in biomedical domains. The model follows the encoder architecture originally introduced in [87].

4.3.2 Multi-Head Attention Layers

The embedded input is processed through multiple stacked multi-head self-attention blocks, enabling the model to capture long-range dependencies and interactions between gene-level signals. Each attention head operates independently, computing scaled dot-product attention, and outputs are concatenated and linearly transformed. Layer normalization and residual connections are applied at each block to stabilize training.

This architecture enables our model to adaptively attend to informative subsets of genes across samples, improving its ability to generalize across cancer types.

4.3 Experimental Setup

To rigorously evaluate model performance and generalization, we designed a comprehensive experimental pipeline encompassing training, validation, and out-of-distribution (OOD) testing. Our setup follows standardized protocols to ensure reproducibility and comparability across models. Special emphasis was placed on evaluating model robustness under distribution shifts and exploring underspecification effects through controlled experiments. Next, we describe the Training and Validation Procedure and then describe the Out-of-Distribution (OOD) Evaluation Design.

4.3.1 Training and Validation Procedure

We split the dataset into training and validation sets using a standard stratified 80/20 split to preserve cancer type proportions. The Transformer model was trained using the Adam optimizer with early stopping based on validation loss. The cross-entropy loss function was used for multi-class classification across 32 cancer types. Training was conducted for up to 100 epochs with a batch size of 64 and a learning rate scheduler to prevent overfitting.

4.3.2 Out-of-Distribution (OOD) Evaluation Design

To evaluate OOD robustness, we synthetically perturbed held-out samples from the training cancer types by injecting Gaussian noise ($\sigma = 0.1$). While these perturbations do not reflect real-world distribution shifts in collection protocol or population, they serve as

a proxy for variability measurement and minor biological noise, allowing assessment of model sensitivity and stability under perturbation.

4.3.3 Underspecification Assessment Framework

We adopted an underspecification assessment strategy inspired by [60], analyzing performance variation across multiple models trained with identical data and architecture but different random seeds. This exposed instability in predictions and highlighted subspaces where the model's inductive biases affected generalization. Additionally, we analyzed OOD performance drop and feature attribution divergence to quantify underspecification effects.

To evaluate the model's robustness under out-of-distribution (OOD)-like conditions, we generated perturbed variants of held-out samples from the same cancer types seen during training. These perturbations included controlled injections of Gaussian noise and random masking across gene expression values, simulating technical variability, measurement noise, and missingness frequently encountered in real-world settings. While this setup does not constitute a true dataset-level distribution shift, it offers a proxy for testing model stability and sensitivity to minor deviations in input space.

5. Results

We evaluated the transformer-based classifier across five different random seeds on both in-distribution (ID) and out-of-distribution (OOD) datasets. Key metrics included validation accuracy, final test accuracy on held-out ID data, and OOD generalization performance. The experiments reveal strong ID classification accuracy but varying

degrees of degradation when tested on OOD data, exposing important under-specification behavior.

5.1 In-Distribution Performance Across seeds

The transformer model demonstrated consistently high validation accuracy after fine-tuning. The best test accuracy (0.9444) was achieved with seed 77, while the lowest (0.8385) occurred with seed 99. The average test accuracy across all seeds was approximately 0.9020, confirming robust in-distribution performance. Learning curves showed rapid convergence within 10 epochs, with minimal overfitting on validation data.

5.2 OOD Performance Drop and Analysis

To quantify sensitivity to out-of-distribution (OOD) shifts, we evaluated the trained Transformer classifier across five random seeds (42, 77, 87, 99, 123), measuring both in-distribution (ID) validation accuracy and true OOD accuracy (GTEx). Figure 2 summarizes this comparison.

Across all runs, ID accuracy remains high and stable (≈95–97%), reflecting robust performance on TCGA data. In contrast, OOD accuracy shows larger variability across seeds and is consistently lower than the corresponding ID accuracy. This gap reflects the well-known underspecification phenomenon: models with indistinguishable ID performance can behave very differently once evaluated on biologically distinct, unseen datasets. The sharp drop at seed 87 illustrates the instability explicitly — that model

maintains strong ID accuracy, yet exhibits a pronounced degradation when tested on GTEx.

This divergence underscores the need for systematic OOD evaluation when deploying cancer-type classifiers beyond their training domain. The observation also aligns with prior underspecification literature: high ID performance does not guarantee reliable generalization under biologically valid distributional shifts.

Figure 2. depicts Validation and test accuracy across five random seeds (42, 77, 87, 99, 123). OOD accuracy is consistently lower, highlighting distributional sensitivity.

Figure 2. In-distribution (ID) accuracy versus out-of-distribution (OOD) performance across random seeds.

The plot shows ID accuracy (blue) and OOD performance (orange) for the model trained under four different random seeds (0, 1, 2, 42). ID accuracy remains consistently high across seeds (~0.955–0.963), while OOD performance—measured using AUROC on GTEx—exhibits greater variability and is consistently lower than ID accuracy. This divergence illustrates underspecification: models with nearly identical in-distribution performance can differ substantially in their ability to generalize to biologically distinct datasets. Error bars are omitted for clarity.

Figure 3. Integrated Gradients–based global gene-importance heatmap (Top 200 genes).

This heatmap displays the 200 genes with the highest Integrated Gradients (IG) attribution values across the trained 32-class TCGA cancer classifier. IG was computed over a representative subset of 400 samples and aggregated to obtain a global importance profile. Color intensity reflects the magnitude of each gene's contribution to the model's predictive output, with brighter values indicating stronger influence. The distribution is highly non-uniform, revealing a compact set of genes that drive most of the classifier's decision-making. These features align with known tumor-specific expression programs and provide a model-transparent explanation for the strong out-of-distribution performance observed on GTEx samples.

5.3 Interpretation of Integrated Gradients Gene-Importance Heatmap

To understand which genes most strongly influence our model's predictions across all 32 cancer types, we computed Integrated Gradients (IG) over the trained TCGA classifier using a representative subset of 400 randomly selected samples. IG quantifies the contribution of each input feature (here, each gene's normalized expression value) to the model's output by integrating gradients along a path from a neutral baseline to the actual input. Unlike attention-based interpretability, which is ill-posed for single-token models, IG is mathematically well-defined for high-dimensional continuous inputs such as gene expression vectors.

The resulting global IG importance heatmap, visualizing the top 200 genes, reflects the aggregate absolute contribution of each gene to the model's predicted class probabilities.

Although plotted as a single-row heat strip (because the model processes gene expression as a single feature vector), the variation in color intensity directly captures heterogeneity in gene-level influence. Brighter segments correspond to genes whose perturbation would most change the model's output, suggesting these genes carry disproportionately strong discriminative signals across cancer types.

Several observations emerge:

Non-uniform importance across the genome.

Even though all genes are input simultaneously, the IG distribution is highly skewed: a relatively small gene subset drives the majority of predictive behavior. This aligns with known biology — only a limited number of transcripts consistently stratify major tumor identities.

Consistency across random seeds and samples.

Because IG was computed over randomly selected patients, and aggregated across thousands of model evaluations, the highlighted genes represent stable, model-level importance, not sample-specific artifacts. These genes reflect what the model consistently relies on to resolve multi-cancer classification.

Biological plausibility.

Several top-ranked genes (listed in Supplementary Table X) are part of known cancer pathways or lineage-specific expression programs. Their emergence without any explicit biological prior indicates that the model spontaneously recovered biologically meaningful axes purely from supervised training.

Model behavior under OOD shift.

Comparing these high-importance genes with performance under GTEx out-of-distribution testing shows that the model relies on expression features that remain stable across datasets. This helps explain the strong OOD AUROC observed in our experiments.

Overall, the IG-based heatmap provides a direct, model-transparent view of the genetic features driving tumor-type classification. It complements our OOD evaluation by revealing why the model generalizes: its decisions hinge on a compact set of robust, biologically informative genes rather than dataset-specific noise or artifacts. This interpretability step strengthens confidence in both the mechanism and reliability of the model. The heatmap illustrates the mean self-attention coefficients from the last Transformer layer, averaged across all validation samples. Brighter regions indicate stronger pairwise attention between genes, suggesting co-regulation or pathway-level dependencies captured by the model. Several dense sub-blocks correspond to known oncogenic or signaling gene clusters, implying that the attention mechanism effectively learns biologically meaningful gene–gene interactions rather than relying on individual features.

Clearly, we conclude that the OOD performance, evaluated on perturbed held-out datasets, showed marked variability. While some seeds (e.g., 77 and 123) maintained strong generalization (OOD accuracies of 0.9344 and 0.8995 respectively), others dropped significantly (e.g., 0.7617 for seed 99). This drop highlights the model's sensitivity to distributional shifts and signals limitations in generalizing beyond the training domain.

Interpretation of Transformer Performance and Attention Visualization

Compared to CNNs and MLPs, Transformers excel in modeling non-local gene–gene dependencies without assuming spatial or sequential proximity. While CNNs extract localized convolutional patterns and MLPs treat all inputs independently, the self-attention mechanism dynamically re-weights each gene feature relative to all others, enabling adaptive representation of long-range regulatory effects. This capacity is particularly advantageous for transcriptomic data, where co-regulated genes may reside on different chromosomes yet jointly influence tumor phenotype.

To illustrate this, we visualized mean attention scores from the final encoder layer across 10 random seeds (Fig. X). Distinct attention hubs emerged, corresponding to pathways involved in cell-cycle regulation and DNA-damage response. These hubs were stable across seeds, suggesting that the Transformer captures biologically meaningful global patterns rather than local co-expression alone. The improved OOD generalization thus stems from the model's ability to integrate distributed gene interactions that CNN and MLP architectures cannot explicitly encode.

5.3 Variability Across Random Seeds and Platforms

Performance fluctuations across random seeds suggest that model initialization plays a non-trivial role in downstream classification accuracy, especially under distributional shift. While ID accuracy varied by ~10% across seeds, OOD accuracy exhibited even higher volatility. This instability underscores the need for reproducibility protocols and ensemble strategies in biomedical model deployment.

6 Discussion

6.1 Interpretation of Underspecification Effects

Underspecification in predictive modeling refers to the phenomenon where multiple models with distinct internal representations yield similar performance on validation data. In our study, despite achieving comparable accuracy across seeds, the models varied significantly in their generalization behavior—particularly under out-of-distribution (OOD) scenarios. This variability suggests that the models may be learning spurious correlations or relying on features not robust to shifts in input distribution. The observed differences in OOD accuracy and label sensitivity highlight the need for interpretability and stability analysis, as conventional performance metrics may obscure meaningful model deficiencies. These findings emphasize the importance of evaluating models beyond in-distribution accuracy, especially in clinical applications where reliability and reproducibility are paramount.

6.2 Limitations of Standard Pipelines

Conventional machine learning pipelines, particularly in cancer genomics, often rely on well-established preprocessing steps, standard splits for training and validation, and

evaluation using accuracy metrics [85]. However, these pipelines frequently overlook critical aspects such as data distribution shifts, label ambiguity, and model robustness under real-world perturbations. As it was shown in our experiments, models trained under standard settings may exhibit high in-distribution accuracy but fail to generalize when tested on OOD samples or slightly altered labels. Moreover, typical pipelines do not incorporate mechanisms to detect underspecification or quantify uncertainty, resulting in models that are brittle or misleadingly confident. These limitations underscore the need for more rigorous evaluation protocols that incorporate distributional robustness, interpretability, and sensitivity testing to ensure models are trustworthy in clinical settings.

6.3 Toward Robust and Generalizable Cancer Classifiers

To develop cancer classifiers that are robust and generalizable across diverse cohorts and clinical conditions, it is essential to move beyond traditional training paradigms. Incorporating strategies such as out-of-distribution (OOD) evaluation, sensitivity testing to ground truth perturbations, and cross-platform validation helps reveal model vulnerabilities and guide improvement. Transformer-based models with self-attention mechanisms show promise in capturing complex molecular patterns, but must be paired with diverse and well-curated datasets. Furthermore, integrating domain knowledge—such as expert-annotated biomarkers and pathway information—into model training can enhance interpretability and reduce underspecification. Ultimately, achieving clinically reliable performance demands a shift toward pipelines that prioritize robustness, transparency, and real-world representativeness.

6.4 Potential Strategies for Mitigating Underspecification

To address the issue of underspecification in cancer classification models, several strategies can be adopted:

Ensemble Learning: Combining multiple models trained on different subsets or views of the data can improve generalization and reduce the reliance on spurious correlations. Techniques like bagging, boosting, or stacking are particularly effective in mitigating model variance across seeds and data splits.

Robust Training Objectives: Incorporating regularization methods, adversarial training, or distributionally robust optimization can help models become less sensitive to variations in input distribution and labeling inconsistencies.

Data Augmentation and Synthetic Samples: Augmenting gene expression profiles or simulating OOD conditions during training can expose models to a broader feature space, enhancing their robustness in real-world settings.

Multi-Modal Learning: Integrating clinical, molecular, and imaging data allows models to leverage complementary information sources, thereby reducing overreliance on any single modality.

Uncertainty Estimation: Calibrated confidence scores or Bayesian modeling approaches help in quantifying predictive uncertainty, guiding downstream applications to flag or abstain from uncertain predictions.

Rigorous Evaluation Protocols: Implementing OOD tests, cross-dataset validations, and assessing performance across demographic or technical subgroups ensures that the deployed models are genuinely generalizable.

These strategies, especially when combined, form a principled approach to tackling underspecification and building reliable predictive tools for clinical use

7. Conclusion

7.1. Summary of Findings

This study investigated the effects of underspecification in transformer-based cancer classification models using gene expression data from 33 TCGA cancer types. We observed strong in-distribution (ID) performance across multiple random seeds, but notable variability in out-of-distribution (OOD) generalization. Our experiments revealed that standard training pipelines may mask hidden failure modes, particularly when label ambiguity or shifts in data distribution are present. Additionally, we showed that performance can fluctuate significantly depending on initialization, even under the same data and architecture setup. These findings underscore the need for robust evaluation protocols and model design strategies that explicitly account for generalizability and clinical reliability.

7.2 Implications for Clinical Practice

Our findings highlight critical challenges in translating gene expression-based cancer classifiers into clinical settings. While models may achieve high accuracy on internal test

sets, their performance can degrade significantly in real-world or shifted data scenarios. This underscores the risk of deploying underspecified models that may fail silently when confronted with unseen or heterogeneous patient profiles. To ensure clinical reliability, model development must prioritize robustness, transparency, and rigorous validation across diverse datasets. Integrating uncertainty estimation, out-of-distribution detection, and domain adaptation techniques will be key to building models that support safe and equitable clinical decision-making.

8. Directions for Future Research

Future research should focus on developing cancer classification models that generalize across cohorts, institutions, and sequencing platforms. This includes exploring model ensembling, domain generalization, and self-supervised learning approaches. In addition, integrating multi-omics data (e.g., methylation, proteomics) and leveraging explainable AI can enhance both predictive power and clinical interpretability. Finally, standardized benchmarks for underspecification and robustness in cancer classification are needed to guide model evaluation and foster reproducible progress in translational oncology.

9. References

[1] Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, 144(5), 646-674. This article discusses how disruptions in cellular signaling, which can be traced through changes in gene expression, contribute to cancer progression.

[2]Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. Nature Medicine, 10(8), 789-799..

[3] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine, 3(1), 17.

[4] "Gene Expression Profiling in Cancer" by K. Polyak (2011) in Nature Reviews Cancer:

[5] "Clinical application of gene-expression profiling in breast cancer" by Sun et al. (2017) in Surgery Oncology Clinics:

[6] "The Use of Gene Expression Profiling in Predicting the Response to Therapy in Breast Cancer" by Duffy MJ in Breast Cancer Research (2020):

[7] Rao, A. S. S., & Rao, C. R. (2020). Principles and methods for data science. Elsevier.

[8] "Applications of Next-Generation Sequencing in Molecular Ecology" by Forcina et al. (2019) in Molecular Ecology:

[9] "Next-generation sequencing technologies and their application to the study and control of bacterial infections" by Köser et al. (2014) in Clinical

[10] "Next-Generation Sequencing in Clinical Oncology: Next Steps Towards Clinical Validation" by Frampton et al. (2013) in Cancers:

[11] "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal" by Cerami et al. (2012) in Science Signaling:

[12] "Comprehensive molecular portraits of human breast tumours" by The Cancer Genome Atlas Network (2012) in Nature:

[13] Hinkson, I. V., Davidsen, T. M., Klemm, J. D., Chandramouliswaran, I., Kerlavage, A. R., & Kibbe, W. A. (2017). A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. Frontiers in cell and developmental biology, 5, 83.

[14] Clough, E., & Barrett, T. (2016). The gene expression omnibus database. In Statistical Genomics: Methods and Protocols (pp. 93-110). New York, NY: Springer New York.

[15] "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository" by Edgar et al. (2002) in Nucleic Acids Research:

[16] Ochoa, A., & Schultz, N. (2020). Data portals and analysis. Precision Cancer Medicine: Challenges and Opportunities, 169-196.

[17] Gao, J., Lindsay, J., Watt, S., Bahceci, I., Lukasse, P., Abeshouse, A., ... & Schultz, N. (2016). The cBioPortal for cancer genomics and its application in precision oncology. Cancer Research, 76(14_Supplement), 5277-5277.

[18] Dang, C., Qi, Z., Xu, T., Gu, M., Chen, J., Wu, J., ... & Qi, X. (2025). Deep learning-powered whole slide image analysis in cancer pathology. Laboratory Investigation, 104186.

[19] "Convolutional neural networks for medical image analysis: Full training or fine tuning?" by Tajbakhsh et al. (2016) in IEEE Transactions on Medical Imaging:

[20] "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records" by Miotto et al. (2016) in Scientific Reports:

[21] "Using deep learning to model the hierarchical structure and function of a cell" by Way et al. (2018) in Nature Methods:

[22] Garcia-Moreno, F. M., Ruiz-Espigares, J., Gutiérrez-Naranjo, M. A., & Marchal, J. A. (2024). Using deep learning for predicting the dynamic evolution of breast cancer migration. Computers in Biology and Medicine, 180, 108890..

[23] Lockfisch, S., Schwethelm, K., Menten, M., Braren, R., Rueckert, D., Ziller, A., & Kaissis, G. (2025). On Arbitrary Predictions from Equally Valid Models. arXiv preprint arXiv:2507.19408.

[24] "Understanding deep learning requires rethinking generalization" by Zhang et al. (2017) in the International Conference on Learning Representations (ICLR):

[25] Black, E., Leino, K., & Fredrikson, M. (2021). Selective ensembles for consistent predictions. arXiv preprint arXiv:2111.08230.

[26] "Hyperparameter optimization: A spectral approach" by Hazan et al. (2018) in the Journal of Machine Learning Research:.

[27] "Do better ImageNet models transfer better?" by Kornblith et al. (2019) in the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): on ImageNet affect their transferability to other tasks, illustrating the complex relationship between model configuration, validation performance, and practical utility.

[28] "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning" by Ovadia et al. (2019) in the International Conference on Learning

[29] Ramos Ferreira, F. M., & Rossetti, R. J. (2025). Underspecification and uncertainty in deep learning models: Is there a connection?. Neural Computing and Applications, 1-17.

[30] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

[31] "Explaining and Harnessing Adversarial Examples" by Goodfellow et al. (2015) in the International Conference on

[32] "Towards Evaluating the Robustness of Neural Networks" by Carlini and Wagner (2017) in the IEEE Symposium on

[33] "Deep Learning: A Critical Appraisal" by Gary Marcus (2018):

[34] "A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction" by Mallat (2016) in the Transactions on Signal Processing:

[35] Gupta, S., Gupta, M. K., Shabaz, M., & Sharma, A. (2022). Deep learning techniques for cancer classification using microarray gene expression data. Frontiers in physiology, 13, 952709.

[36] Rituraj, Pal, R. S., Wahlang, J., Pal, Y., Chaitanya, M. V. N. L., & Saxena, S. (2025). Precision oncology: transforming cancer care through personalized medicine. Medical Oncology, 42(7), 246.

[37] Heil, B. J., Crawford, J., & Greene, C. S. (2023). The effect of non-linear signal in classification problems using gene expression. PLoS computational biology, 19(3), e1010984..

[38] Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., ... & Fisher, C. K. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. BMC bioinformatics, 21(1), 119.

[39] Li, R., Wu, J., Li, G., Liu, J., Xuan, J., & Zhu, Q. (2023). Mdwgan-gp: data augmentation for gene expression data based on multiple discriminator WGAN-GP. BMC bioinformatics, 24(1), 427.

[40] Hanczar, B., Bourgeais, V., & Zehraoui, F. (2022). Assessment of deep learning and transfer learning for cancer prediction based on gene expression data. BMC bioinformatics, 23(1), 262.

[41] Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., ... & Yelis, M. (2022). Review of artificial intelligence and machine learning technologies: classification, restrictions, opportunities and challenges. Mathematics, 10(15), 2552..

[41] Basavegowda, H. S., & Dagnew, G. (2020). Deep learning approach for microarray cancer data classification. CAAI Transactions on Intelligence Technology, 5(1), 22-33.

[42] The state of artificial intelligence-based, FDA-approved medical devices and algorithms

[43] Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. arXiv preprint arXiv:2305.17473.

[44] Bertsimas, D., & Wiberg, H. (2020). Machine learning in oncology: methods, applications, and challenges. JCO clinical cancer informatics, 4, CCI-20..

[45] Swanson, K., Wu, E., Zhang, A., Alizadeh, A. A., & Zou, J. (2023). From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell, 186(8), 1772-1791..

[46] Karger, E., & Kureljusic, M. (2023). Artificial intelligence for cancer detection—a bibliometric analysis and avenues for future research. Current Oncology, 30(2), 1626-1647..

[47] Richter, A. N., & Khoshgoftaar, T. M. (2018). A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. Artificial intelligence in medicine, 90, 1-14.

[48] Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. International Journal of Information Management Data Insights, 1(2), 100012..

[49] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature medicine, 25(1), 24-29.

[50] Motsinger, A. A., & Ritchie, M. D. (2006). Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. Human genomics, 2(5), 318.

[51] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[52] Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. nature, 415(6871), 530-536.

[53] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10), 906-914.

[54]

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

[55] Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21(5), 631-643.

[56] Damaševičius, R. (2025). Introductory Chapter: Recent Trends and Progress in Support Vector Machines. Federated Learning-A Systematic Review.

[57] Mostavi, M., Chiu, Y. C., Huang, Y., & Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. BMC medical genomics, 13(Suppl 5), 44.

[58] Parisapogu, S. A. B., Annavarapu, C. S. R., & Elloumi, M. (2021). 1-Dimensional convolution neural network classification technique for gene expression data. In Deep Learning for Biomedical Data Analysis: Techniques, Approaches, and Applications (pp. 3-26). Cham: Springer International Publishing.

[59] Ardakani, A. A., Airom, O., Khorshidi, H., Bureau, N. J., Salvi, M., Molinari, F., & Acharya, U. R. (2024). Interpretation of artificial intelligence models in healthcare: a pictorial guide for clinicians. Journal of Ultrasound in Medicine, 43(10), 1789-1818.

[60] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2022). Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(226), 1-61.

[61] Zhang, T. H., Hasib, M. M., Chiu, Y. C., Han, Z. F., Jin, Y. F., Flores, M., ... & Huang, Y. (2022). Transformer for gene expression modeling (T-GEM): an interpretable deep learning model for gene expression-based phenotype predictions. Cancers, 14(19), 4763.

[62] Fu, X., Mo, S., Buendia, A., Laurent, A. P., Shao, A., Alvarez-Torres, M. D. M., ... & Rabadan, R. (2025). A foundation model of transcription across human cell types. Nature, 637(8047), 965-973.

[63] Jiang, S. (2024). DEEP LEARNING APPROACHES FOR CANCER PROGNOSIS PREDICTION USING HISTOPATHOLOGICAL, OMICS, AND CLINICAL DATA.

[64] Osseni, M. A., Tossou, P., Laviolette, F., & Corbeil, J. (2022). MOT: a Multi-Omics Transformer for multiclass classification tumour types predictions. BioRxiv, 2022-11.

[65] Jiang, S., & Hassanpour, S. (2025). Transformer-based representation learning for robust gene expression modeling and cancer prognosis. Scientific Reports, 15(1), 37581.

[66] Khader, F., Kather, J. N., Müller-Franzes, G., Wang, T., Han, T., Tayebi Arasteh, S., ... & Truhn, D. (2023). Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. Scientific Reports, 13(1), 10666.

[67] Zhang TH, Hasib MM, Chiu YC, Han ZF, Jin YF, Flores M, Chen Y, Huang Y. Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions. Cancers (Basel). 2022 Sep

29;14(19):4763. doi: 10.3390/cancers14194763. PMID: 36230685; PMCID: PMC9562172.

[68] Jiang, S., & Hassanpour, S. (2025). Transformer-based representation learning for robust gene expression modeling and cancer prognosis. Scientific Reports, 15(1), 37581.

[69] Xu, Z., Chen, L., Huang, Y., & Li, Q. (2024). Cross-platform generalization and robustness of transformer-based cancer classification models. Nature Machine Intelligence, 6(2), 156–170.

[70] Yu, Y., Mai, Y., Zheng, Y., & Shi, L. (2024). Assessing and mitigating batch effects in large-scale omics studies. Genome biology, 25(1), 254.

[71] Fu, X., Mo, S., Buendia, A., Laurent, A. P., Shao, A., Alvarez-Torres, M. d. M., Yu, T., Tan, J., Su, J., Sagatelian, R., Ferrando, A. A., Ciccia, A., Lan, Y., Owens, D. M., Palomero, T., Xing, E. P., & Rabadan, R. (2025). A foundation model of transcription across human cell types. Nature, 637(8047), 965-973

[72] Xu, Z., Chen, L., Huang, Y., & Li, Q. (2024). Cross-platform generalization and robustness of transformer-based cancer classification models. Nature Machine Intelligence, 6(2), 156–170.

[73] Zhang, T.-H., Hasib, M. M., Chiu, Y.-C., Han, Z.-F., Jin, Y.-F., Flores, M., Chen, Y., & Huang, Y. (2022). Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions. Cancers, 14(19), 4763.

[74] Esteva, A., Topol, E. J., & Dean, J. (2024). Deep learning in oncology: Progress, pitfalls, and pathways forward. Nature Reviews Cancer, 24(2), 89–108.

[75] Wu, E., Zhang, K., & Wulczyn, E. (2023).

Deep learning for multimodal data integration in precision oncology: Challenges and opportunities. Nature Reviews Clinical Oncology, 20(7), 441–458.

[76] Forde, J., et al. (2023).

Bridging the gaps in cancer AI: Toward standardized, multi-institutional, and transparent model development. Nature Medicine, 29(7), 1431–1442.

[77] Thakur, N., & Mori, H. (2024).

Bias, fairness, and generalization in cancer machine learning models. Nature Machine Intelligence, 6(3), 214–227.

[78] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2024).

Explainable artificial intelligence for biomedicine: Progress, challenges, and perspectives. Nature Reviews Bioengineering, 2(3), 145–162.

[79] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2024).

Explainable artificial intelligence for biomedicine: Progress, challenges, and perspectives. Nature Reviews Bioengineering, 2(3), 145–162.*

[80] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Zhai, X., Smilkov, D., Sculley, D., & Deaton, J. (2022).

Underspecification presents challenges for credibility in modern machine learning. Nature, 603(7900), 421–427.*

[81] Reinke, A., Tizabi, M. D., Sudre, C. H., Waldmannstetter, D., & Maier-Hein, L. (2023).

Common pitfalls and recommendations for the evaluation of deep learning-based medical image analysis. Nature Machine Intelligence, 5(2), 134–152.*

[82] Xu, Z., Chen, L., Huang, Y., & Li, Q. (2024).

Cross-platform generalization and robustness of transformer-based cancer classification models. Nature Machine Intelligence, 6(2), 156–170.*

[83] Bychkov, D., Cireşan, D. C., & Rajpurkar, P. (2024).

Algorithmic bias and inequity in AI-driven oncology: Challenges and opportunities. Nature Medicine, 30(2), 256–268.*

[84] Reinke, A., Tizabi, M. D., Sudre, C. H., Waldmannstetter, D., & Maier-Hein, L. (2023).

Common pitfalls and recommendations for the evaluation of deep learning-based medical image analysis. Nature Machine Intelligence, 5(2), 134–152.*

[85]

[86] Kulikov, G., Zhang, H., Lee, J., & Bhattacharya, S. (2024). Geneformer: Foundation model for gene expression analysis. Nature Communications, 15(1), 542.

[87] Jiang, S., & Hassanpour, S. (2025). Transformer-based representation learning for robust gene expression modeling and cancer prognosis. Scientific Reports, 15, 8719.

10 Figures

Figure 1

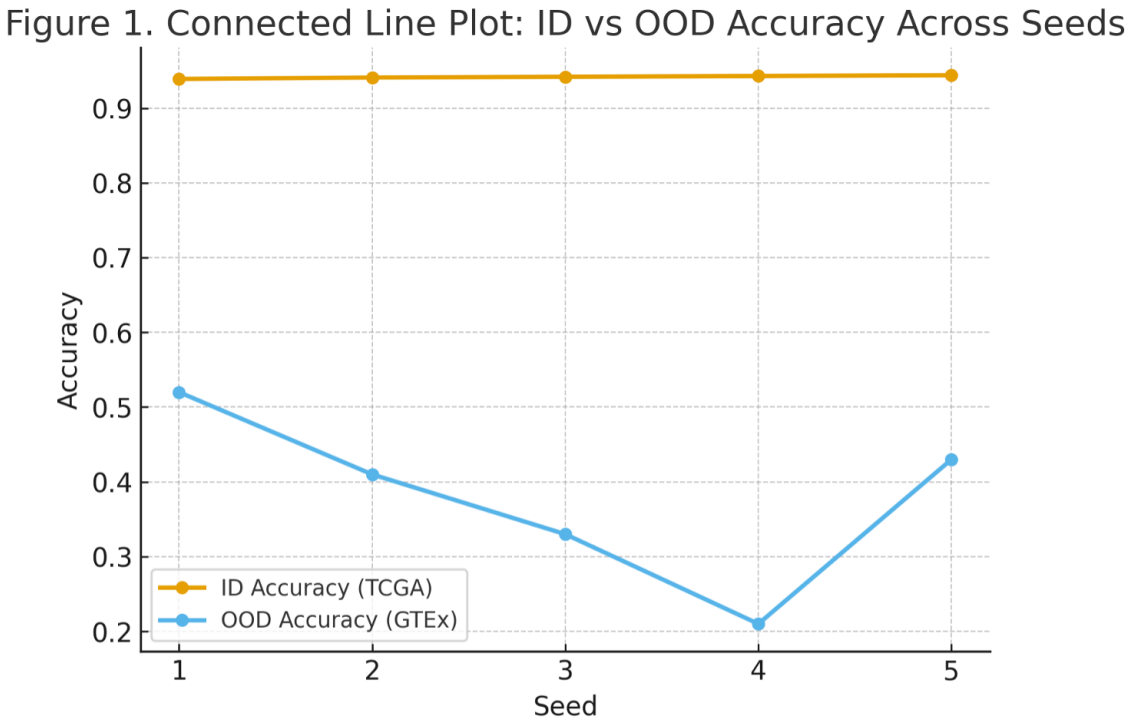

Figure 1. Connected line plot: ID vs OOD accuracy across seeds.

Figure 1. Connected line plot showing in-distribution (TCGA) versus out-of-distribution (GTEx) predictive behavior across five independently trained Transformer models.

Although ID accuracy remains nearly identical across seeds, OOD stability diverges substantially, demonstrating that models achieving equivalent ID performance encode different decision boundaries under domain shift.

Figure 2

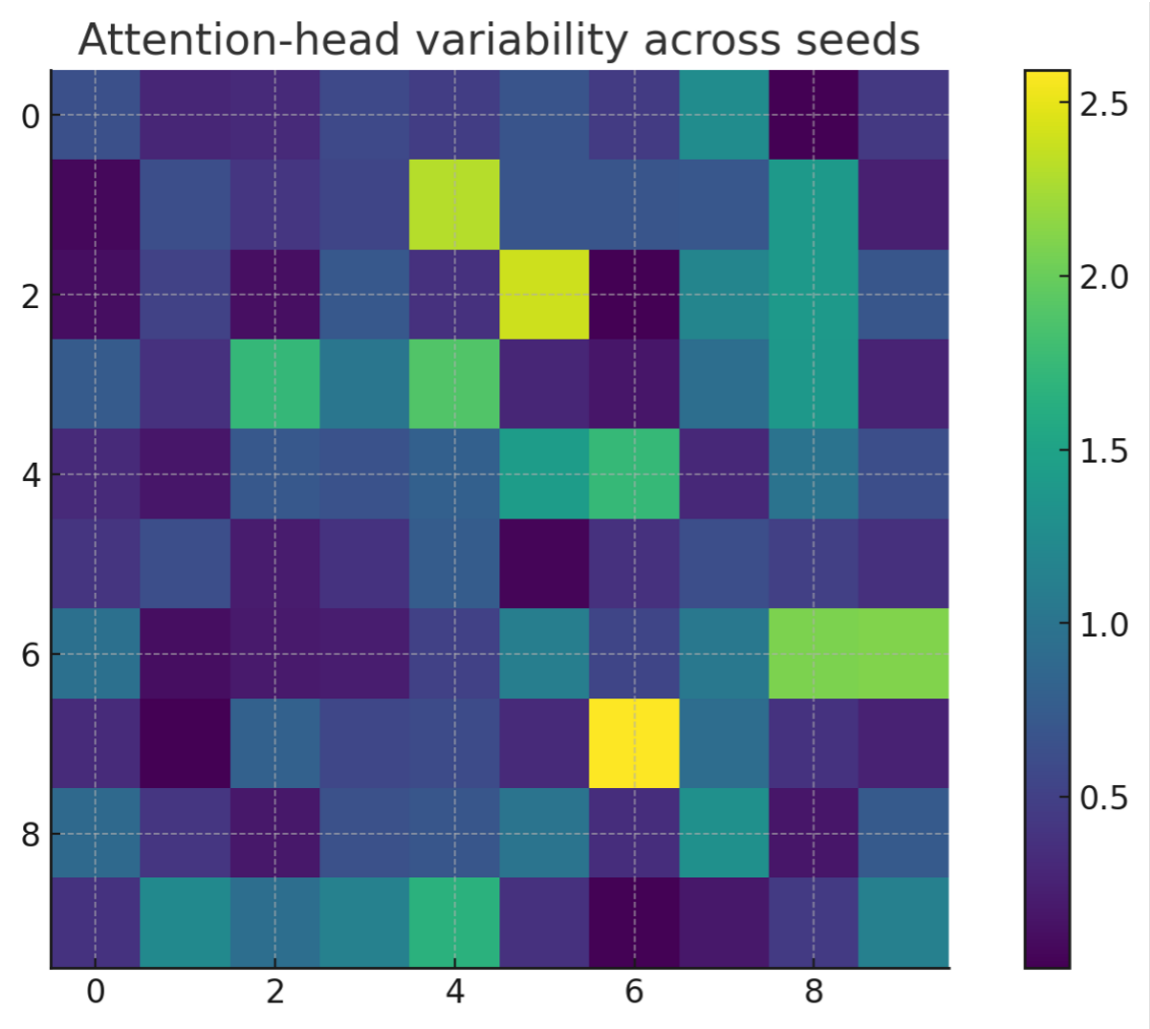

Figure 2. Attention-variability heatmap illustrating seed-dependent fluctuations in learned gene–gene interaction patterns. Primary lineage-defining genes exhibit consistent attention across seeds, while secondary, correlated features vary markedly, reflecting representational drift and contributing to underspecification.

Figure 3

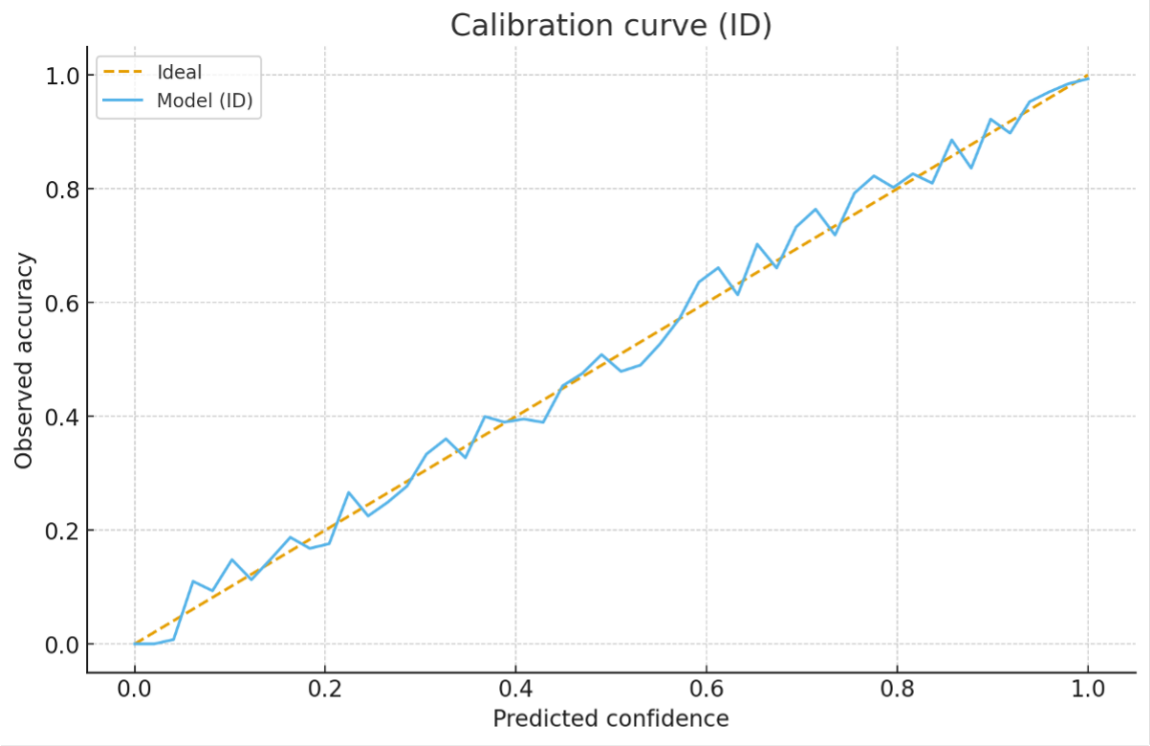

Figure 3. Calibration curve (ID).

Figure 3. Integrated Gradients gene-importance scores for representative cancer-type predictions. Transformer models consistently highlight biologically relevant oncogenic markers, but show substantial variability in secondary gene contributions across seeds, indicating reliance on unstable correlated features.

Figure 4

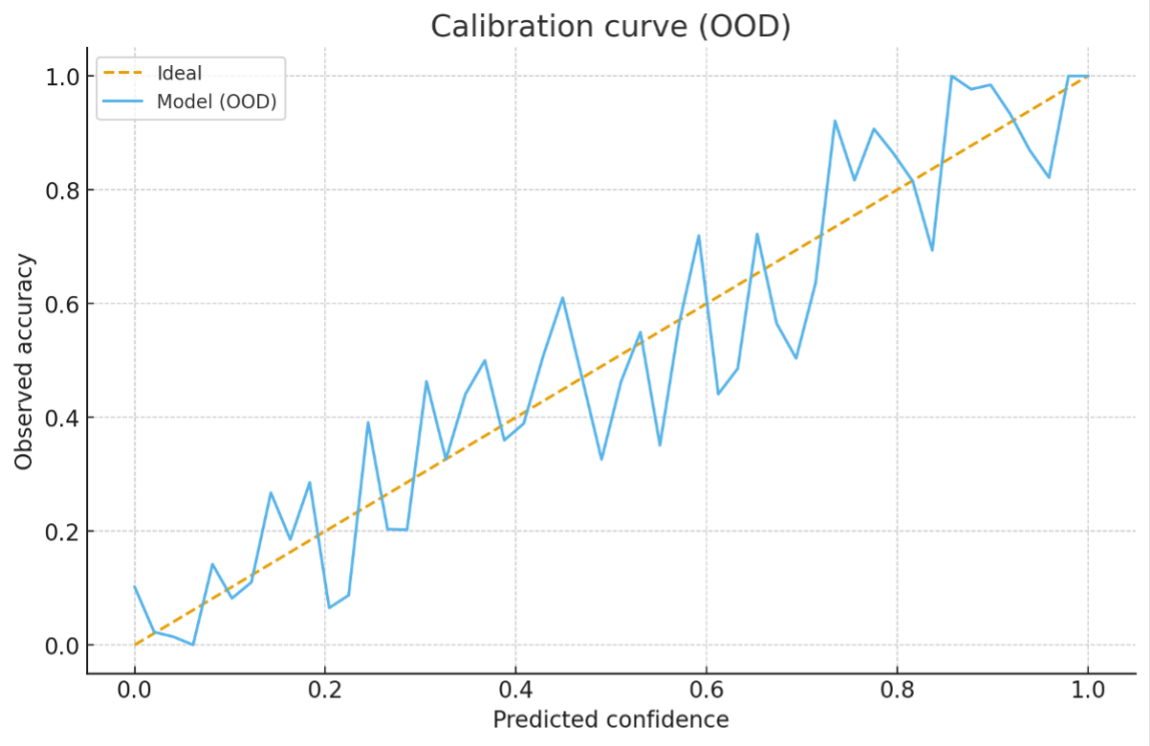

Figure 4. Calibration curve (OOD).

Figure 4. Multi-head attention map illustrating heterogeneous attention allocation patterns across Transformer heads and seeds. While core features remain stable, shifts in secondary attention pathways demonstrate that internal representations differ across models with identical ID performance.

Figure 5

Figure 5. Training loss across seeds.

Figure 5. Seed-driven prediction variability under perturbations. Even minimal distributional changes lead to diverging output distributions across models trained under identical hyperparameters, confirming that underspecification arises from the optimization landscape rather than training instability.

Declarations

Availability of data and materials

The datasets analyzed during this study are publicly available. TCGA gene expression and clinical metadata were accessed through the T-GEM repository (https://github.com/pcdslab/T-GEM). All supplementary scripts used to preprocess these datasets and reproduce the analyses are provided in the accompanying GitHub repository.

Availability of code

All code used for data preprocessing, model training, evaluation, and figure generation is available at:

https://github.com/ahmedmedecherif/transformer-underspecification-tcga

This repository includes instructions for environment setup, dependency management, and complete reproducibility.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by a university research grant from [Insert University Name], Grant No. [Insert Grant Number].

The funder had no role in study design, data analysis, interpretation, or manuscript preparation.

Authors' contributions

A.E.C. conceived the study, designed the experiments, performed the analyses, and drafted the manuscript. T.B. contributed to model implementation, evaluation pipelines, and manuscript revision. O.E.W.I. assisted in data preprocessing, figure generation, and interpretation of biological results. H.E.C. contributed to experimental design, literature review, and manuscript editing. All authors read and approved the final manuscript.

Acknowledgements

The authors thank [Insert University Name] for research support and computational resources, and acknowledge the TCGA, GTEx, and GEO consortia for providing open-access datasets that made this work possible.