Foundation AI Models for the Prediction of Therapeutic Response to Next Generation Immune Checkpoint Inhibitors

Ryan Dalton[1,*], Eshed Margalit[1,*], Keith Mitchell[1,*], Michela Meister[1,*], Daniel Millman[1], Maede Zolanvari[1], Lucas Cavalcante[1], Joy S. Tea[1], Dexter Antonio[1], Maxime Dhainaut[1], Chloe Delepine[2], Dulce Ovando[1], Francis Fernandez[1], Aaron Salm[1], Angela V. Hafner[2], Joseph E. Grossman[2], Dhan Chand[2], Ronald W. Alfa[1], Daniel Bear[1], Emily Corse[1], Lacey Padrón[1]

1: Noetik Inc., South San Francisco, CA
2: Agenus, Inc., Lexington, MA
*These authors contributed equally to this work

Abstract

Identifying the patients most likely to benefit from specific therapies remains a central challenge in cancer treatment. Patient selection is particularly challenging in immuno-oncology, where therapeutic response depends on a dynamic interplay between the immune system and the tumor microenvironment (TME) that is difficult to capture using traditional biomarker or histology-based approaches. As a result, there is a critical unmet need for platforms capable of representing patient and disease heterogeneity using routinely available clinical materials to drive patient selection and therapeutic success. We recently developed TARIO-2, a multimodal AI 'world model' designed as a general simulator of biology and capable of inferring spatial and molecular tumor biology from hematoxylin and eosin (H&E) pathology images alone. To evaluate TARIO-2's clinical utility, we used it to retrospectively predict patient responses to botensilimab (BOT), a multifunctional Fc-enhanced anti-CTLA-4 antibody, plus balstilimab (BAL), an anti-PD-1 antibody, in the C-800-01 phase 1b trial. In the setting of microsatellite stable metastatic colorectal cancer with no active liver metastases (MSS mCRC NLM), TARIO-2 identified a subgroup of patients with substantially improved outcomes across multiple clinical endpoints. Supportive response-enrichment signals were also observed in ovarian cancer and sarcoma. In benchmark analyses within MSS mCRC NLM, TARIO-2 outperformed leading H&E-based external pathology AI models. These findings support use of TARIO-2 in extracting clinically relevant spatial biology from routine pathology images. Prospective validation will help to determine whether H&E-based multimodal AI models may help match patients to therapies, enrich clinical trials for likely responders, and support more efficient development of immuno-oncology agents across tumor types.

Introduction

AI is helping to engender an age of pharmacological abundance in which novel molecules can be designed and synthesized with unprecedented speed. Yet even with well-designed drugs it remains difficult to predict efficacy. This disconnect between drug behavior in preclinical studies and drug responses in patients speaks to a critical knowledge gap between chemistry and biology and underscores that the rules of human biology will remain unclear until we model them directly.

The disconnect is evident in clinical oncology, where despite considerable advances in targeted therapies, identifying the right treatment for patients remains challenging. As a result, many promising therapies only benefit a small subset of patients, and clinicians and drug developers often lack the practical tools to identify those patients in advance. This is particularly true for cancer immunotherapies, which have demonstrated durable impacts in even treatment-refractory cancers, but whose impacts are uneven, reflecting both the complex and multifaceted mechanisms of these drugs, as well as a variety of known resistance mechanisms. This creates a need for scalable platforms that can model human biology to enable better patient selection, stronger trial design, and more efficient development of new cancer therapies.

To satisfy this need in a clinical setting, these platforms will need to navigate a steep tradeoff: they must capture the complex molecular, cellular, and spatial architecture of the tumor microenvironment (TME) without relying on specialized assays or scarce clinical tissue. Spatial transcriptomic (SpT) technologies are capable of measuring tumor and immune system complexity, but these technologies require substantial laboratory and computational resources; moreover, the data they create are influenced by processing and storage artifacts. These challenges hamper the direct, large-scale deployment of SpT, making it infeasible in clinical practice. In contrast to SpT, H&E pathology images represent one of the most widely available sources of patient tumor information. These images are generated for nearly every patient with cancer and contain visual information about tumor structure, immune infiltration, stromal organization, necrosis, and other features of the TME. But much of this information is difficult to quantify by human review alone.

Resolving this tension between data availability and data richness may be possible through the usage of computational paradigms that treat routine clinical materials not merely as static images, but as partially hidden representations of complete biological systems. This is precisely the value of ‘world models’, AI models built to predict the complete or future state of a system given only information about its current or partially-hidden state. Here we introduce TARIO-2, a world model designed to understand patient-level biology. TARIO-2 was trained on multimodal data from thousands of human cancer samples, including custom generated H&E stains, multiplex immunofluorescence, and SpT data. Critically, at inference time, TARIO-2 requires only an H&E image as input–the same image that is already collected for every cancer patient as part of routine diagnostic care. From this input, TARIO-2 constructs a spatial representation of the TME that includes, but is not limited to, an inferred spatial transcriptome (iSpT): expression values for 18,694 genes, at single-cell resolution. Through these core capabilities TARIO-2 can be used to perform broad simulations of TME biology, including synthetic interventions such as the deletion or overexpression of arbitrary genes or cells. Thus, TARIO-2 overcomes the key methodological gaps outlined above, using a universally-available data modality, the H&E stain, to create a systems-level representation of the TME.

To evaluate the clinical utility of TARIO-2, we applied it retrospectively to identify patients responding to next-generation immune checkpoint inhibitor (ICI) BOT+BAL in the fully enrolled C-800-01 phase 1b trial. BOT is a multifunctional Fc-enhanced anti–CTLA-4 antibody with

distinct mechanisms of action from other anti-CTLA-4 drugs, designed to enhance anticancer immunity by activating cancer-fighting immune cells and depleting the regulatory immune cells that suppress them. BAL is an anti–PD-1 antibody pharmacologically comparable to approved PD-1 inhibitors intended to maintain the immune response against cancer. BOT+BAL has demonstrated durable anticancer activity in several immunologically 'cold' and ICI-refractory tumors, including MSS mCRC NLM, sarcoma, and ovarian cancer [1-4]. Across these indications, among treatment-refractory patients with advanced disease, 19–23% of patients demonstrated objective response (i.e., the percentage of patients whose tumors shrink by a predefined amount), with clinically meaningful survival outcomes also exhibited. Given BOT+BAL's differentiated mechanism of action, clinical activity is not strongly associated with traditional biomarkers such as PD-L1 expression or tumor mutational burden. These qualities make BOT+BAL well suited for the evaluation of a broader TME-based strategy to identify which patients are most likely to benefit.

In the present work, we assessed whether TARIO-2-derived features from routine pretreatment H&E images could identify patients with improved clinical responses to BOT+BAL treatment. We focused on best overall response and overall survival across MSS mCRC NLM, ovarian cancer, and sarcoma cohorts, and benchmarked TARIO-2 against selected external H&E-based pathology AI models. Although these analyses are preliminary and were performed using a small population of treated patients (n=113), we found that TARIO-2 significantly outperformed existing pathology foundation models in prediction tasks across multiple cancers and for multiple clinical endpoints. This retrospective study establishes a proof-of-concept for TARIO-2–guided patient selection, and sets the stage for a paradigm in which patients are rapidly matched to treatments on the basis of spatial-molecular signatures inferred from routine histology.

Methods

To train a foundation model that understands human tumor biology, we first generated a rich, multimodal dataset profiling thousands of patient samples. We sourced formalin-fixed, paraffin-embedded human tumor samples across 19 cancer indications (**Figure 1A**) and created Tissue Microarrays (TMAs) by taking 12-24 cores from each sample. Cores were chosen according to a proprietary computational pipeline operating on H&E images from the samples. On these TMAs, and using adjacent serial 4 μm sections, we generated data across three modalities: H&E, multiplex immunofluorescence using a 16-plex panel, and whole-transcriptome spatial transcriptomics (SpT) using the NanoString CosMx platform. We then computationally aligned data from serial sections at high resolution, such that all multimodal data were paired and spatially aligned.

Using the paired H&E and SpT data, we trained a self-supervised foundation model, TARIO-2. **(Figure 1B)**. TARIO-2 is an autoregressive transformer that predicts SpT from a combination of aligned H&E images and, optionally, SpT from the same sample. During training, TARIO-2 leverages the paired Noetik dataset to learn the relationship within and between both modalities.

At inference time, the model is able to predict SpT from H&E alone, allowing prediction of inferred SpT (iSpT) from H&E samples that have no paired SpT ground truth.

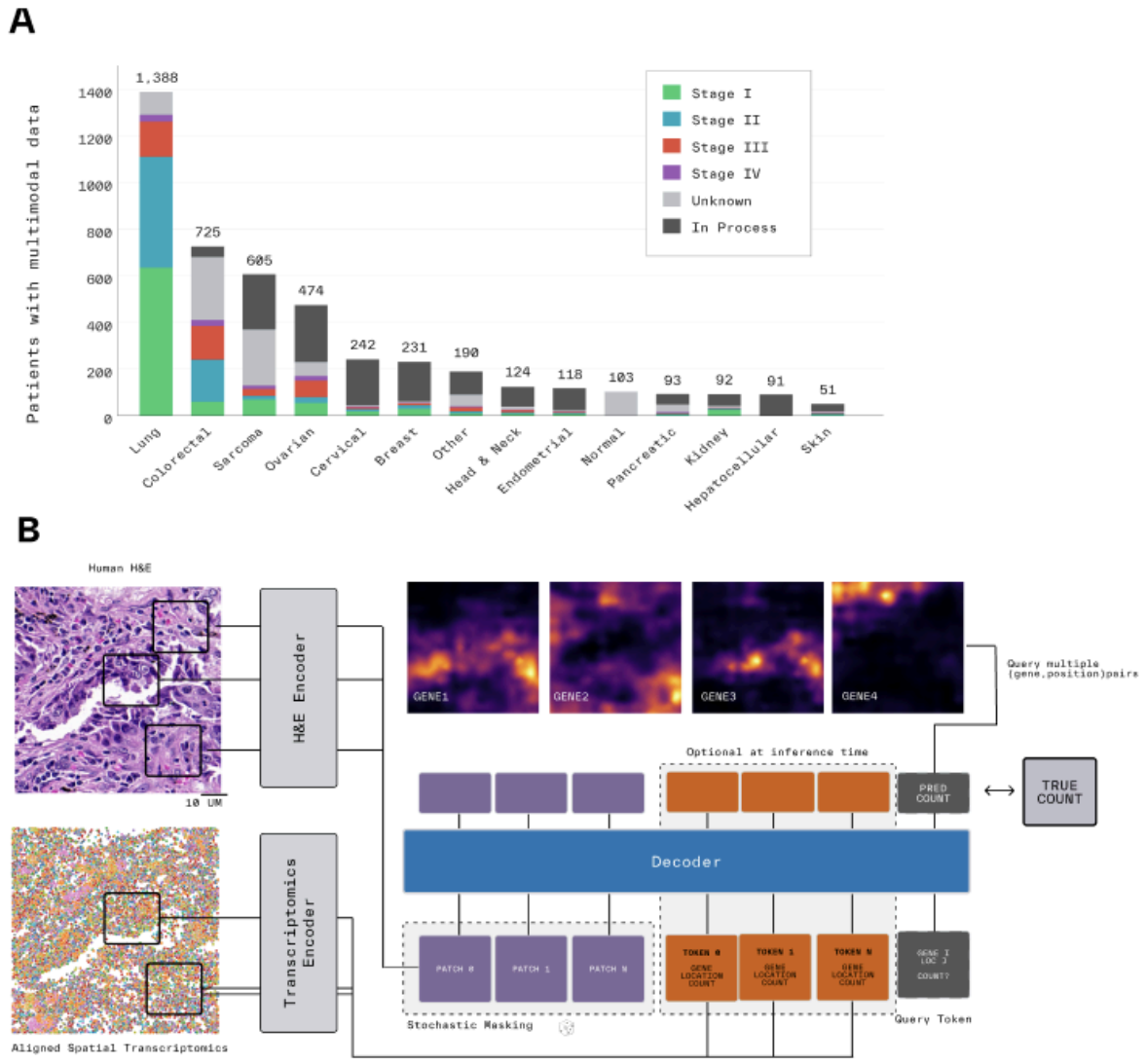

Figure 1: TARIO-2 is a foundation model trained on paired H&E and Spatial Transcriptomics from thousands of patient samples. *(A) Noetik's foundation cancer dataset on Tissue Microarrays (TMAs, patent pending, 12-24 cores per patient sample) summarized by cancer type and stage. This dataset currently includes >5000 patients, each with all 3 spatial data types, including >725 CRC, >605 sarcoma, & >474 ovarian cancer patients. Foundation models were not trained on the prediction cohort in this study. (B) TARIO-2 is a foundation model that is trained on aligned tiles of human H&E images and spatial RNA transcripts. Each modality is encoded into a sequence of tokens with a modality-specific encoder; then the sequences from both modalities are concatenated. During training, a transformer-based decoder is used to predict later tokens in the sequence from earlier tokens. Training does not include clinical samples. At inference time, H&E from pretreatment samples is fed to the model, which predicts dense spatial patterns of gene expression (iSpT) for each gene in the transcriptomics panel.*

To determine whether iSpT faithfully simulate patient-level tumor biology, we evaluated their accuracy on a held-out cohort of 213 patients whose data was not used for model training. Prediction accuracy was computed as the Spearman correlation between the spatial pattern of

ground truth transcript detections and TARIO-2 predictions (**Figure 2A**). Ground-truth spatial patterns vary in the amount of spatial structure present; when genes are either weakly expressed or expressed uniformly across the tissue, there is no meaningful spatial variation to predict. To overcome this issue, we quantified ground-truth spatial structure using Moran's I, a statistical measure used to calculate spatial autocorrelation, and then sorted genes by this measure. We found that model predictions were most accurate for genes with more spatially structured expression, and decreased gradually as genes with weaker spatial structure were included in the quantification (**Figure 2B**). Because spatial Spearman correlations capture the similarity of spatial patterns within samples, but not differences in expression magnitude across samples, we also evaluated the ability of TARIO-2 to rank patients by overall expression magnitude, irrespective of specific spatial patterns. To do so, we measured the correlation between total transcripts aggregated at the gene-level between ground-truth CosMx and iSpT. Herein we refer to this measure as "bulk" correlation; we found that bulk correlation was reliably greater than ϱ = 0.5 for almost all genes across all tested cancer types (**Figure 2C**). This capability underlies the ability of TARIO-2 to identify patients in which a given gene is expressed at high or low levels, which may contribute to the ultimate classification of best overall response (BOR) or overall survival (OS).

We next sought to use TARIO-2 to retrospectively identify patients likely to benefit from BOT+BAL, using only their pretreatment H&E images. To do so we worked with samples from the C-800-01 Phase 1b Trial (NCT03860272) that evaluated BOT±BAL across advanced, treatment-refractory solid tumors (data cutoff of December 13, 2025; **Figure 3A**). We used 113 pretreatment H&E images from efficacy-evaluable (received ≥1 post-baseline 6-week imaging scan) patients treated with BOT+BAL, spanning MSS mCRC NLM (n=53), ovarian cancer (n=27), and sarcoma (n=33). All H&E images were evaluated by a pathologist to assess tumor content and to annotate tumor regions. To ensure the robustness of predictive models, the evaluation pipelines for BOR and OS were treated as independent analytical streams, each subject to its own rigorous, locked quality control protocols. As a result, there are small discrepancies in patient inclusion for the MSS mCRC NLM and ovarian cancer cohorts for BOR versus OS analyses. After all quality control was completed, we had a final analysis set of 49 (BOR) and 47 (OS) samples for MSS mCRC NLM, 23 (BOR) and 24 (OS) samples for ovarian cancer, and 31 (BOR and OS) samples for sarcomas (**Figure 3B**).

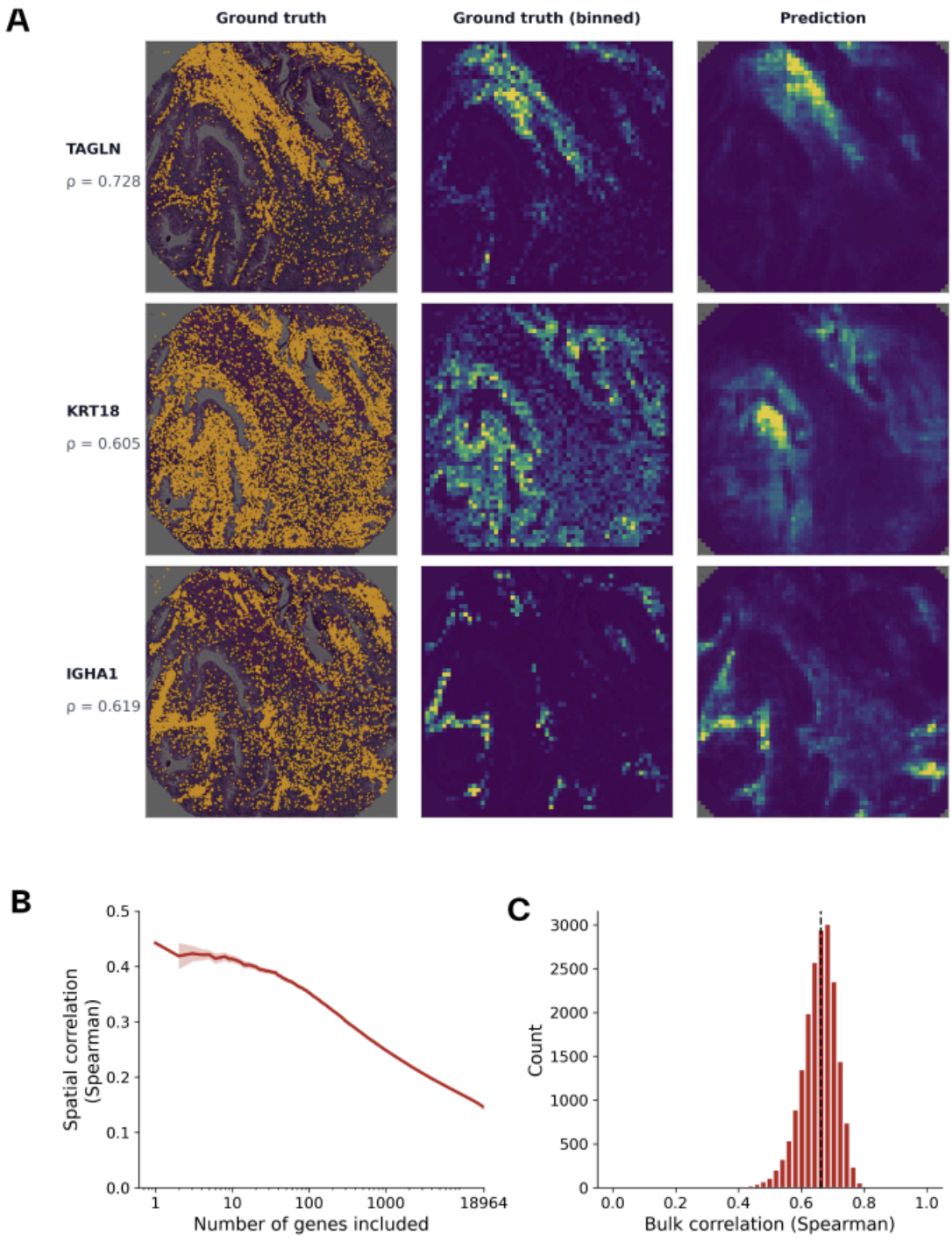

Figure 2: TARIO-2 accurately predicts subcellular spatial expression of the whole transcriptome.
(A) Comparison between raw transcripts (yellow, left column), binned transcript counts (center column), and TARIO-2 predictions (right column) across multiple genes for the same sample (B) Quantification of TARIO-2 accuracy on held-out patient samples. Spatial correlation as a function of the number of genes included, where genes are ranked by spatial structure (Moran's I). Shaded regions: 95% CI across samples. (C) Correlation between spatially-averaged TARIO-2 predictions and total transcript detections across held-out samples.

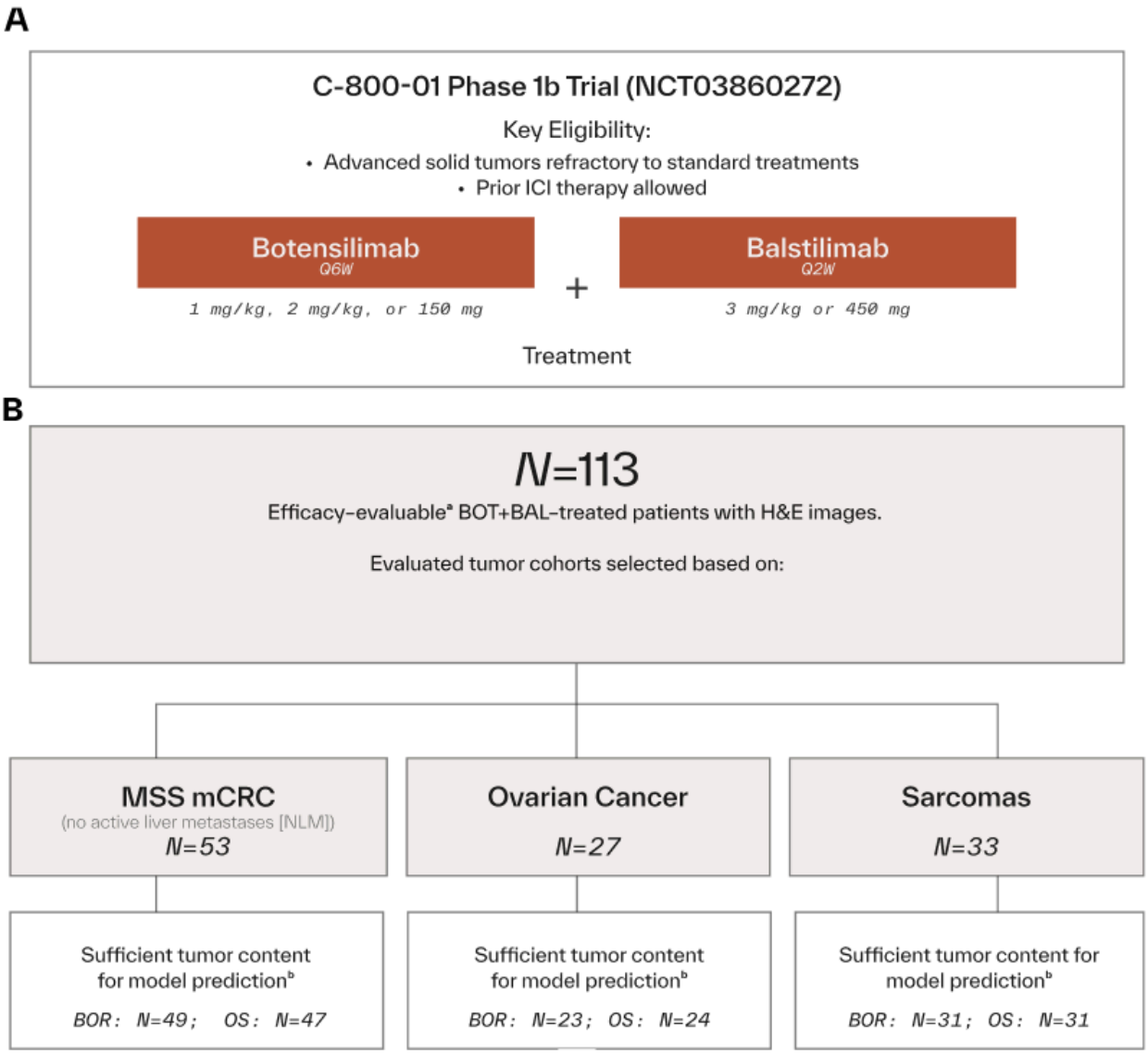

Figure 3: BOT + BAL clinical trial design and sample selection for this study. *(A) Overview of C-800-01 Phase 1b trial design and key eligibility criteria. (B) CONSORT diagram illustrating sample selection for the current study. All samples consist of H&E images. Data cut-off date: December 13, 2025. [a]Efficacy-evaluable patients received ≥1 post-baseline 6-week imaging scan. [b]Differences in quality control resulted in 2 additional patients in MSS mCRC NLM BOR studies and 1 additional patient in ovarian cancer OS studies.*

We then created iSpT data by running TARIO-2 inference on all H&E images at a regular spacing of 35 μm. While many featurization approaches were determined in our preliminary studies to be sufficient to predict BOT+BAL response, in the present work we simply aggregated the predictions of each gene by averaging that gene's predicted expression within pathologist-defined tumor regions. We next used these features to train models for two separate prediction tasks: OS and BOR. The BOR prediction task was formulated as a binary

classification problem, with responders defined as those with an investigator-assessed partial or complete response (PR or CR) according to the RECIST 1.1 criteria.

For OS prediction, we trained Gradient Boosting models with Cox proportional hazard loss; for the BOR prediction task we trained support vector classifiers (SVC). Any hyperparameter tuning was performed within training folds. In both prediction tasks, models were evaluated using leave-one-group-out (LOGO, used for OS prediction) or 50 repeats of Monte Carlo cross validation (MCCV, used for BOR prediction). Accuracy was assessed using predictions on out-of-fold data. For BOR, accuracy is summarized as median area under the receiver operating characteristic (AUROC), and for OS prediction, accuracy is summarized as median C-index. For both AUROC and C-index, a value of 0.5 denotes a chance classifier and a value of 1.0 denotes a perfect classifier. All endpoints were predicted independently, with no data leakage between endpoints. In order to test for significance, performance on real data was compared to permutation distributions created by shuffling data labels.

Results

TARIO-2 stratified outcomes in MSS mCRC NLM

In the MSS mCRC NLM cohort, TARIO-2 derived features from pretreatment H&E images stratified both OS and BOR following BOT+BAL treatment. The OS model achieved a median C-index of 0.67, which was significant compared with the permutation control distribution (p=0.04; **Figure 4A**).

To assess the clinical relevance of these results, patients were ranked by cross-validated TARIO-2 survival scores and divided into model-selected subgroups comprising the top 30% of patients (i.e those with the lowest predicted risk of death scores using held-out data) and the remaining 70% of patients. This cutoff was predefined to identify a clinically meaningful subgroup enriched for clinical benefit while preserving sufficient sample size for analysis and based roughly on the objective response rates in the C-800-01 clinical trial, which were approximately 20% across the cohorts evaluated here and approached 30% in select tumor histologies (angiosarcoma) [2, 3, 4, 8]. Median overall survival was not reached in the best 30% cohort, versus 13.3 months in the remaining 70%, corresponding to a hazard ratio of 0.18 (**Figure 4B**).

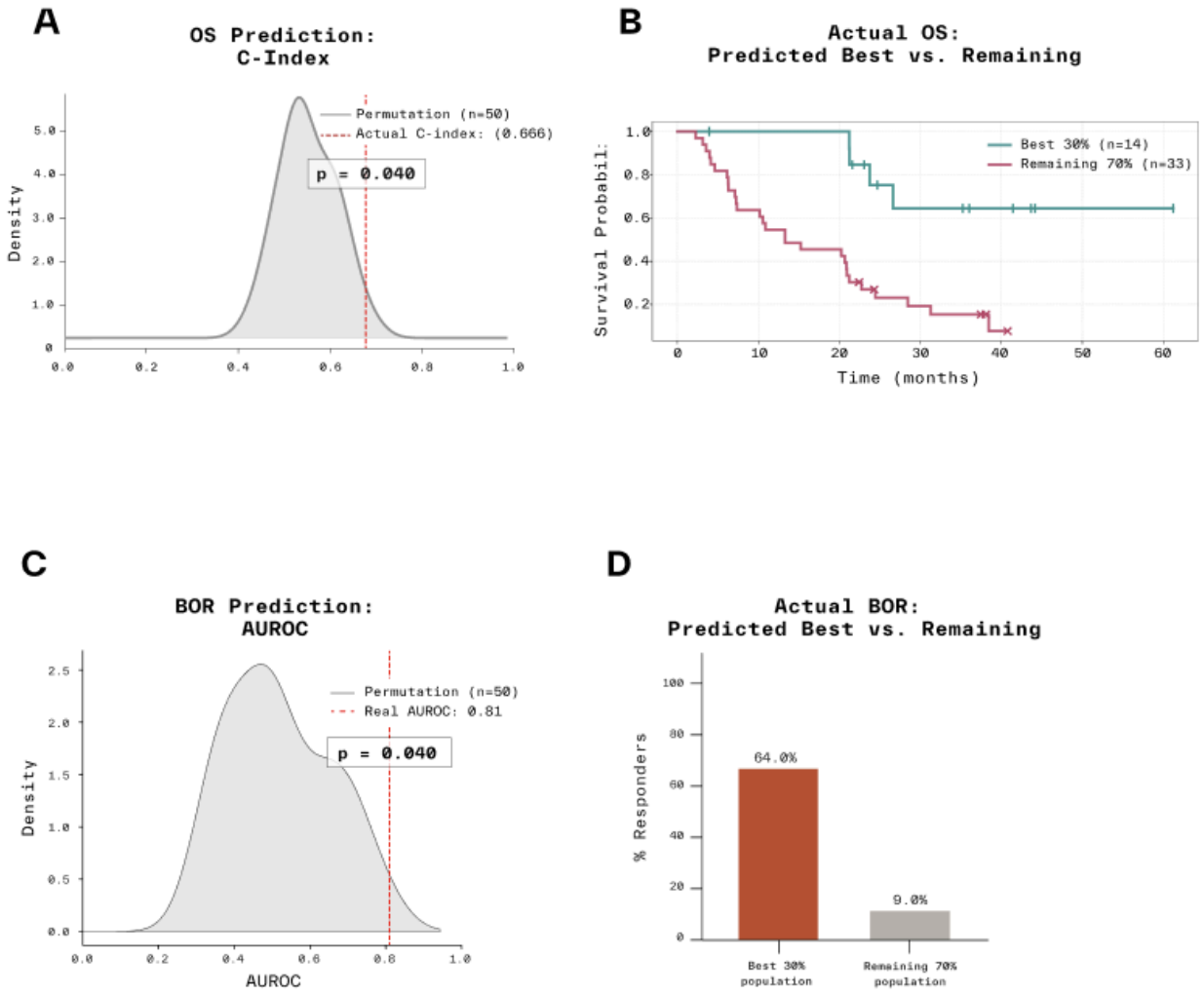

Figure 4: TARIO-2 accurately predicts OS and Response from H&E images alone for the MSS mCRC NLM Cohort, and identifies patient subsets with improved outcomes. *(A) Results of OS prediction task. Null distribution computed as a result of permuting OS labels is shown in grey. Red line indicates actual C-index achieved with TARIO-2 model on true OS labels. Actual C-index reaches a p-value of 0.04 by a 2-sided test relative to null distribution. (B) For all patients used in OS prediction tasks, a median risk score was calculated across out-of-fold predictions. Based on these scores, patients were partitioned into two groups: best 30% (blue) and remaining 70% (red). Actual OS was then compared for these partitions. The best 30% population has a hazard ratio of 0.18 as compared to the remaining patients, and median OS was not reached. The remaining population has a median OS of 13.3 months. (C) Results of BOR prediction relative to null distribution achieved by permuting response labels. Classifiers reached a median AUROC of 0.81 and p-value = 0.04 (D) Using median predicted responses on out-of-fold data, patients were stratified into best 30% and remaining 70% groups. Within the best 30% group the actual response rate was 64%, compared to a 9% actual response rate among the remaining 70% group.*

TARIO-2 also stratified BOR in the MSS mCRC NLM cohort. The response model achieved a median AUROC of 0.81, which was significant compared with the permutation control distribution (p=0.04; **Figure 4C**). When patients were ranked by cross-validated TARIO-2 response scores, the model-selected top 30% subgroup demonstrated an objective response rate of 64%, compared with 9% in the remaining cohort (**Figure 4D**). Baseline demographics, clinical characteristics, and conventional biomarkers were not statistically significantly different

between the model-selected and remaining groups, suggesting that the observed enrichment was not readily explained by measured baseline features alone, although interpretation is limited by sample size (**Table 1**).

MSS CRC, no liver mets	Best 30% by Predicted Response ($N = 14$)	Remaining 70% by Predicted Response ($N = 35$)
Responders, N (%)	9 (64%)***	3 (9%)
PFS hazard ratio	0.46*	2.16
Median PFS (months)	9.8*	4.1
Median Age, years (range)	60 (48–82)	57 (36–81)
Female, N (%)	5 (36%)	18 (51%)
ECOG of 1 at baseline, N (%)	7 (50%)	24 (69%)
Median prior lines of therapy (range)	4 (1–10)	4 (1–9)
Prior PD-(L)1, N (%)	2 (14%); $N = 14$	11 (31%); $N = 35$
PD-L1 TPS – median (range); mean; N	0 (0–20); mean=2.8; $N = 13$	0 (0–5); mean=0.2; $N = 33$
PD-L1 CPS – median (range); mean; N	5 (0–90); mean=12.8; $N = 13$	3 (0–20); mean=5.6; $N = 33$

Table 1: Demographics & outcomes among model-selected cohort. *Patients were stratified into groups based on their median predicted BOR. Demographic and disease characteristics were then compiled for these groups. Categorical variables were compared with two-sided Fisher's exact tests (with missing values dropped per group), continuous variables with two-sided Mann-Whitney U tests, and time-to-event endpoints (OS, PFS) with two-sided log-rank tests between the best 30% and remaining 70% groups, with displayed hazard ratios derived from Cox models comparing each group against the other cohort (best vs. remaining, remaining vs. best) (* p<0.05, ** p<0.01, *** p<0.001).*

Supportive signals observed in ovarian cancer and sarcoma

TARIO-2 was also evaluated in ovarian cancer and sarcoma cohorts treated with BOT+BAL. Given the smaller sample sizes, these analyses were considered exploratory.

In ovarian cancer, the TARIO-2 response model achieved a median AUROC of 1.00, although this did not reach statistical significance versus the permutation distribution (p=0.12; **Figure 5A**). The ability to achieve perfect AUROC with a p-value that doesn't reach significance highlights the critical need for permutation testing when analyzing small cohorts. The model-selected best 30% cohort had an actual objective response rate of 83%, as compared to the remaining 70% cohort, which had a response rate of 6% (**Figure 5B**). Models trained to predict OS achieved a median C-index of 0.73, which reached a p-value of 0.10 compared to the permutation distribution (**Figure 5C**). As with MSS mCRC NLM, no differences in baseline characteristics between these groups reached statistical significance (**Table S1**).

In Sarcoma, BOR models achieved a median AUROC of 0.83, which was not statistically significant, with a p-value of 0.20 compared to the permutation distribution (**Figure 5D**). The model-selected best 30% cohort had an actual objective response rate of 56%, as compared to the remaining 70% cohort, which had a response rate of 5% (**Figure 5E**). The sarcoma OS model achieved a median C-index of 0.65, which reached a p-value of 0.18 compared to the permutation distribution (**Figure 5E**). While sample sizes are too small to draw definitive

conclusions, a comparison of the baseline characteristics, demographics, histologies, and biomarkers of these patients revealed that there were no statistically significant differences in other features (**Table S2**).

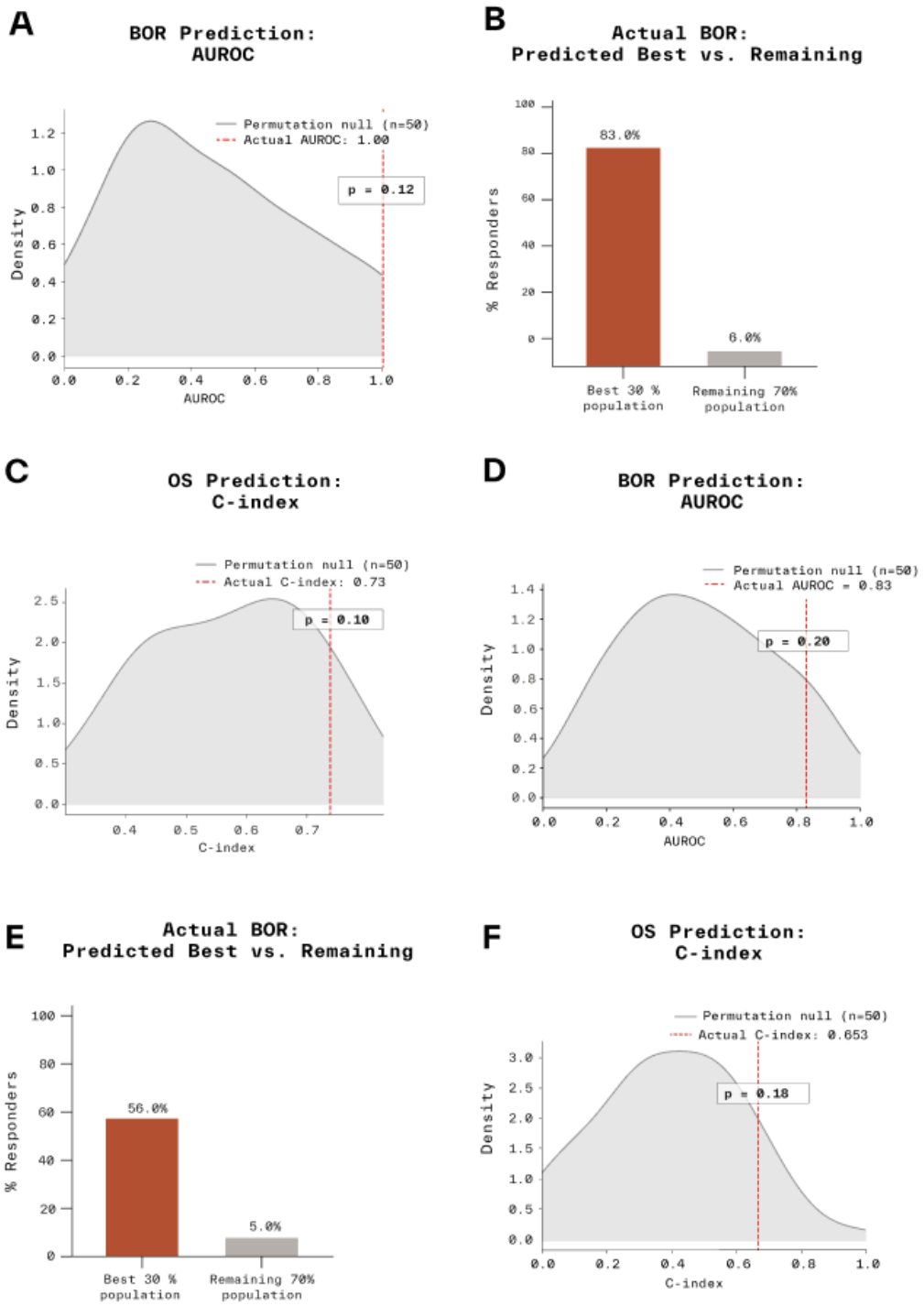

Figure 5: best overall response & OS prediction: ovarian cancer and sarcoma: *(A) Results of BOR prediction in ovarian cancer. Median AUROC = 1 and p-value = 0.12 from two-sided permutation test. (B) Patients were stratified*

into groups based on their median predicted response using held-out data, yielding a best 30% population with an actual response rate of 83%, compared with a 6% response rate for the remaining 70% population. (C) Results of OS prediction in ovarian cancer. C-index = 0.728 and p-value = 0.1. (D) Results of BOR prediction in sarcoma. Median AUROC of 0.83 and p-value of 0.2. (E) Best 30% population, stratified as above, has an actual response rate of 56%, compared with a 5% response rate for the remaining 70% population. (F) Results of OS prediction in sarcoma. C-index = 0.653 and p-value = 0.18.

TARIO-2 outperformed selected pathology foundation models in MSS mCRC NLM cohort

We next sought to understand whether the ability to predict therapeutic response to BOT+BAL is unique to TARIO-2, potentially reflecting its understanding of the relationship between H&E and SpT, or if recently-developed models trained on H&E only might also be able to predict outcomes. To test this, we compared our accuracy with TARIO-2 in MSS mCRC NLM, our largest cohort, to accuracy using two open-source H&E foundation models that are trained on large H&E-only datasets: GigaPath [5] and H-Optimus-0 [6]. We chose these models as the most high-performing recent H&E models with permissive licenses. Both models are trained on large patient cohorts: GigaPath is trained on data from over 30,000 patients, and H-Optimus-0 is trained on data from over 500,000 whole-slide images from an undisclosed number of patients.

Using the same tumor-region aggregation and prediction framework applied to TARIO-2, we next evaluated prediction accuracy on embeddings from each external model. We found that TARIO-2 outperformed these external H&E foundation models on both BOR and OS prediction tasks (**Table 2**). We also evaluated all models using attention-based multiple instance learning (abMIL), a popular technique for predicting labels from whole slide images, and the approach recommended by the H-Optimus-0 authors. The abMIL models were trained to 20 epochs and evaluated across 5 repeats of 3-fold cross validation. None of the models achieved performance above chance for either BOR or OS using abMIL. Finally, we considered aggregation methods with more representational capacity. For GigaPath we fine-tuned the slide encoder using the official implementation to predict BOR. The slide encoder was trained for 5 epochs and then evaluated on 15 folds, by running 3 sets of 5-fold cross-validation and reporting the median AUROC across these 15 folds. For TARIO-2 and H-Optimus-0 we fit an instance of TransMIL [7] to predict BOR or OS. As with abMIL, we trained TransMIL to 20 epochs and evaluated across 5 repeats of 3-fold cross validation. In this setting we find that TARIO-2 outperforms both H&E foundation models, although results are weaker than the tumor-region aggregation described above. In summary, across these analyses, TARIO-2 maintained stronger endpoint-stratification performance than competing models, supporting the value of its multimodal training strategy linking routine H&E images to inferred spatial and molecular tumor biology.

Aggregation Method	Model	BOR	OS
Mean Pooling in Tumor Regions	TARIO-2	**0.81***	**0.66***
	GigaPath	0.68	0.58
	H-Optimus-0	0.44	0.54
abMIL	TARIO-2	0.50	**0.52**
	GigaPath	**0.53**	0.46
	H-Optimus-0	**0.53**	0.41
TransMIL or slide encoder	TARIO-2	**0.60**	**0.58**
	GigaPath	0.55^	0.47
	H-Optimus-0	0.50	0.43

Table 2: Model Performance by Aggregation Method. Bold *indicates the best score within each aggregation method block. Underline indicates the best model overall across all aggregation methods. ^computed using GigaPath slide encoder. *statistically significant (two-sided permutation test) at α= 0.05*

Discussion

This retrospective analysis provides an early clinical proof point for TARIO-2 as a platform for extracting clinically relevant tumor biology from routine pretreatment H&E pathology images. In patients treated with BOT+BAL, TARIO-2 identified subgroups associated with improved clinical outcomes, most notably in MSS mCRC NLM, where the model-selected subgroup showed higher response rates and improved OS compared with the remaining cohort. Supportive response-enrichment signals were also observed in ovarian cancer and sarcoma, suggesting that the platform may capture relevant TME features across multiple immunotherapy-resistant tumor types.

BOT+BAL represents an important test case for this approach. The combination was designed to extend immunotherapy activity into tumors that have historically been resistant to conventional checkpoint inhibition, and its activity is not fully explained by traditional biomarkers such as PD-L1 expression or tumor mutational burden; this creates a need for biomarker strategies that capture broader features of the TME. TARIO-2 addresses this need by using routinely available pretreatment H&E images to infer spatial and molecular features of the TME that may be associated with treatment benefit.

The practical importance of this approach is that it does not require additional tissue, specialized spatial assays, or complex molecular testing for every patient. H&E slides are already generated as part of standard cancer care and are widely available across clinical trials and real-world datasets. If prospectively validated, TARIO-2 could provide a scalable strategy to support patient selection, enrich clinical trials for likely responders, and improve confidence in development decisions for BOT+BAL and potentially other immuno-oncology agents with complex mechanisms of action.

The benchmarking results further support the potential value of TARIO-2's multimodal training strategy. In MSS mCRC NLM, TARIO-2 outperformed selected external H&E-based pathology foundation models for endpoint stratification. This suggests that linking routine histology to inferred spatial tumor biology may provide information beyond what is captured by H&E-only

image models. For drug development, this distinction is important. The goal is not simply to classify images, but to convert routine pathology into a biologically informed representation that can help identify which patients are more likely to benefit from a given therapy.

Beyond informing patient selection for BOT+BAL and other therapies, TARIO-2 may have broader applications across drug discovery and development. A multitude of drugs have been found in clinical trials to be effective for very small numbers of patients; at times the lack of clear biomarkers has been an impediment to full utilization of these drugs. It stands to reason then that there are patients today who would benefit from existing drugs if novel biomarkers could be discovered. We propose that TARIO-2 identifies what might be conceptualized as 'latent integrative biomarkers'. These latent integrative biomarkers are suited to achieve the task of matching patients to therapies in settings where traditional biomarkers cannot.

Furthermore, by building a relationship between latent representations of biology and therapeutic response, TARIO-2 extends naturally from matching patients to existing therapeutics, to guiding the design of novel therapies further upstream. Currently, early drug discovery efforts make heavy use of computational chemistry approaches, as well as model systems such as mouse, primate, and cell culture, to build mechanistic hypotheses of drug activity prior to using those drugs in a clinical setting. However, while these models are invaluable for isolating baseline pharmacology, they frequently fail to capture the complex, multicellular networks that dictate therapeutic success in real patients. We expect that TARIO-2, in combination with our recent work projecting model organisms into human spatial transcriptomic space, may provide a clear route to developing what might be considered 'hybrid model systems' capable of simulating biology and testing emerging hypotheses in tight feedback loops. Taken together, world models of biology like TARIO-2 may enable improved stratification for existing therapeutic approaches while promoting a tighter coupling between drug development and precision care, reducing the costs and time associated with drug development while improving patient outcomes.

Conclusion

Here we present TARIO-2, a multimodal 'world model' of biology designed as a general simulator of biology and capable of inferring clinically relevant spatial and molecular tumor biology from routine H&E pathology images. TARIO-2 was used to predict iSpT on pre-treatment images from patients treated with BOT+BAL, a promising combination immunotherapy which has shown efficacy across several hard-to-treat tumor types. Using rigorous modeling approaches and permutation-based controls on a preliminary, retrospective population of 113 patients, TARIO-2 demonstrated statistically significant predictive performance for BOR and OS in MSS mCRC NLM, with supportive trends observed in sarcoma and ovarian cancer; in MSS mCRC NLM, TARIO-2 also outperformed selected external foundation models.

These findings provide an initial clinical proof point for TARIO-2 as a scalable H&E-based biomarker platform. Prospective validation is needed to confirm model thresholds, clinical utility, and generalizability, including in earlier-line MSS CRC settings. More broadly, scaling TARIO-2

across H&E and clinical outcome datasets may enable more efficient patient-to-therapy matching, improved clinical trial enrichment, and broader application of AI-enabled pathology across oncology drug development, bringing immediate benefits to the patients of today while helping to prepare for the patients of tomorrow.

References

1. Chand D, et al. *Cancer Discov*. 2024;14(12):2407–2429.
2. Bullock AJ, et al. *Nat Med*. 2024;30(9):2558–2567.
3. Wilky BA, et al. *J Clin Oncol*. 2025;43(11):1358-1368.
4. Porter R, et al. *J Immunother Cancer*. 2025;13(12):e013222.
5. Xu, H., et al. *Nature.* 2024. 630, 181–188
6. Saillard, C.,et al. 2024. *H-optimus-0*. GitHub.
7. Shao, Z., et al. NeurIPS. 2021;34:2136–2147.
8. Schlechter BM, et al. Poster presented at the ESMO Gastrointestinal Cancers Congress. Barcelona, Spain. 2025. Poster #8P.

Supplementary Tables

Ovarian	**Best 30% by Predicted Response** ($N = 6$)	**Remaining 70% by Predicted Response** ($N = 17$)
Responders, N (%)	5 (83%)**	1 (6%)
PFS hazard ratio	0.15**	6.79
Median PFS (months)	11.1**	1.4
Median Age, years (range)	66 (55–70)	59 (37–75)
ECOG of 1 at baseline, N (%)	3 (50%)	9 (53%)
Median prior lines of therapy (range)	2 (1–8)	4 (2–16)
Prior PD-(L)1, N (%)	1 (17%); $N = 6$	1 (6%); $N = 17$
PD-L1 TPS – median (range); mean; N	1 (0–15); mean=3.0; $N = 6$	1 (0–30); mean=3.2; $N = 17$
PD-L1 CPS – median (range); mean; N	7 (1–30); mean=10.7; $N = 6$	5 (0–65); mean=10.8; $N = 17$
Histology		
HGS & Serous Adeno, N (%)	5 (83%)	14 (76%)
Clear Cell, N (%)	1 (17%)	3 (18%)

Table S1: Demographics & outcomes among model-selected cohort for Ovarian Cancer. *Patients were stratified into groups based on their median predicted BOR. Demographic and disease characteristics were then compiled for these groups. Categorical variables were compared with two-sided Fisher's exact tests (with missing values dropped per group), continuous variables with two-sided Mann-Whitney U tests, and time-to-event endpoints (OS, PFS) with two-sided log-rank tests between the best 30% and remaining 70% groups, with displayed hazard ratios derived from Cox models comparing each group against the other cohort (best vs. remaining, remaining vs. best) (* p<0.05, ** p<0.01, *** p<0.001).*

Sarcoma	**Best 30% by Predicted Response** ($N = 9$)	**Remaining 70% by Predicted Response** ($N = 22$)
Responders, N (%)	5 (56%)**	1 (5%)
PFS hazard ratio	0.68	1.46
Median PFS (months)	5.6	1.4
Median Age, years (range)	64 (40–78)	62 (31–80)
Female, N (%)	4 (44%)	18 (82%)
ECOG performance status = 1 at baseline, N (%)	5 (56%)	12 (55%)
Median prior lines of therapy (range)	2 (1–6)	3 (1–10)
Prior PD-(L)1, N (%)	1 (11%)	1 (5%)
PD-L1 TPS – median (range); mean; N	2 (0–95); mean=18.6; $N = 9$	0 (0–10); mean=2.2; $N = 20$
PD-L1 CPS – median (range); mean; N	5 (0–98); mean=19.2; $N = 9$	0 (0–20); mean=3.9; $N = 20$
Histology		
Angiosarcoma, N (%)	5 (55%)	5 (23%)
LMS, N (%)	2 (22%)	10 (45%)
Other, N (%)	2 (22%)	7 (32%)

Table S2: Demographics & outcomes among model-selected cohort for Sarcoma. *Patients were stratified into groups based on their median predicted BOR. Demographic and disease characteristics were then compiled for these groups. Categorical variables were compared with two-sided Fisher's exact tests (with missing values dropped per group), continuous variables with two-sided Mann-Whitney U tests, and time-to-event endpoints (OS, PFS) with two-sided log-rank tests between the best 30% and remaining 70% groups, with displayed hazard ratios derived from Cox models comparing each group against the other cohort (best vs. remaining, remaining vs. best) (* p<0.05, ** p<0.01, *** p<0.001).*