CanBART: A Generative Foundation Model of Cancer Molecular Alterations for Synthetic Patient Generation and Genomic Profile Completion

Dmitrii Chebanov, Quaid Morris

Memorial Sloan Kettering Cancer Center

Abstract

Despite the rapid expansion of genomic profiling in oncology, real-world datasets remain limited in size and unevenly distributed, particularly for rare cancers and underrepresented patient populations. The number of available sequenced cases is often insufficient for robust machine learning, and even within existing cohorts, sequencing panels differ in gene coverage. As a result, crucial molecular features are frequently missing, making it difficult to compare cohorts, identify actionable alterations, or enroll patients into clinical trials. These data gaps limit the utility of precision oncology and drug discovery.

To address these challenges, we introduce CanBART – a generative foundation model trained on sequencing data from 144,000 patients. CanBART represents somatic alterations as tokenized sequences and learns to reconstruct missing genomic features, classify tumor types, and generate synthetic patient cohorts. The model employs a BART-style masked language architecture, enabling flexible inference, imputation, and biologically informed data augmentation.

CanBART was applied to impute missing gene statuses in real-world datasets. As an illustrative use case, in the OncoPanel dataset (DFCI), it successfully predicted over one-third of missing mutations with high confidence. Using sampling strategies adapted from NLP, we generated “plausible patients” – synthetic genomic profiles that extend the training data with biologically coherent examples. These synthetic cohorts were biologically validated and used to train tumor-type classifiers, improving accuracy for two-thirds of cancer types, particularly rare ones represented by only 20-500 samples.

By unifying fragmented genomic data and enabling the generation of plausible profiles, CanBART provides a scalable tool for precision oncology – with applications in rare cancer research, multi-site data harmonization, and clinical trial optimization. Moreover, it generates realistic “virtual patients” that reproduce real-world co-alteration patterns and can be used to simulate clinical trial eligibility.

Introduction

Foundation models have revolutionized multiple fields, from natural language processing to biomedical research, by enabling the development of large-scale, pre-trained architectures that generalize across diverse tasks. In genomic research, Transformer-based architectures have demonstrated significant utility in molecular biology, including AlphaFold for protein structure prediction, Enformer for predicting gene expression and epigenetic signals from genomic sequence, and RNA-FM for modeling RNA structure and function from large-scale transcriptomic data (1-4). These foundation models have shown the ability to learn biologically meaningful representations from raw sequences, setting a precedent for the application of large-scale pretrained models in genomics and transcriptomics.

A central motivation for this study was to develop a foundation model in which the tokens correspond to molecular alterations found in individual cancer patients. The goal was to use such a model to perform classification, simulation, and property inference for genomically profiled patients, analogous to how foundation models in molecular biology operate on DNA, RNA, or protein sequences. This task is particularly timely: on one hand, genomic profiles play a crucial role in treatment selection, prognosis estimation, and drug discovery; on the other hand, a sufficiently large dataset of such profiles is now available through the MSK patient cohort. Inspired by natural language processing, we hypothesized that a generative architecture could be used not only for classification but also for creating realistic synthetic patient cohorts—helping to overcome the scarcity of data for rare cancer types. A further motivating precedent comes from the GDD-ENS study by Darmofal et al., which showed that classifiers trained exclusively on common cancers were able to predict the organ of origin for previously unseen rare cancers with 66% accuracy (5). This implies that rare and common cancers may share latent genomic patterns that can be exploited by learned representations.

To explore this further, we introduce CanBART, a foundation model trained using the BART architecture, where the input "sentences" are composed of molecular alterations identified in tumor profiling. The model is pre-trained using masked language modeling on over 144,000 cancer patient profiles and is subsequently fine-tuned for a downstream task—tumor classification—serving as a testbed for evaluating representational quality and generation-based augmentation.

Our study proceeds in three stages. First, we fine-tuned the pre-trained CanBART model on a classification task involving 9 common cancer types and a single pooled category for all rare cancers. We then analyzed the latent space learned by the model to assess whether rare cancers tend to cluster near common ones, which would support the hypothesis of shared representational structure across tumor types. Second, we leveraged this structure to generate plausible patients for underrepresented rare cancer types via masked sampling and demonstrate that augmenting training datasets with these synthetic patients significantly improves classification accuracy for rare cancers. Third, we show that the same generative mechanism can be used to impute missing gene statuses in individual profiles, helping to recover potentially undiagnosed molecular alterations.

The problem of tumor classification based on genomic profiling has direct clinical relevance, particularly for cancers of unknown primary (CUP) and ambiguous histologies. Improved classification informs therapy selection, clinical trial enrollment, and precision treatment planning. Equally important is the task of imputing missing gene statuses in real-world patient profiles. By inferring plausible missing alterations, CanBART can help recover diagnostically relevant mutations in underprofiled patients—those with limited gene panel coverage or sequencing failure. This functionality supports downstream decision-making and enhances patient matching for therapies and clinical trials. Beyond classification and imputation, our approach enables cross-platform data harmonization and co-alteration inference, making CanBART a flexible foundation for a range of clinical and translational applications.

In summary, CanBART enables data-efficient representation learning from real-world cancer genomics data and supports both inference and generation tasks. By leveraging latent similarity between common and rare cancer types, it provides a principled framework for improving rare cancer classification and modeling patient cohorts through synthetic generation.

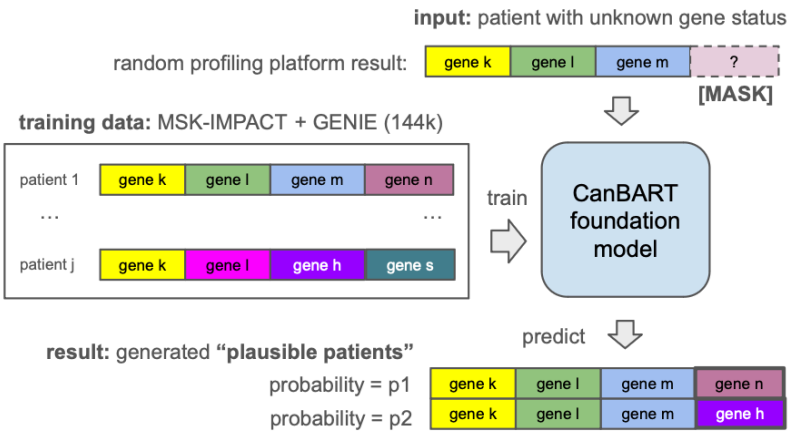
Methods

Data Representation

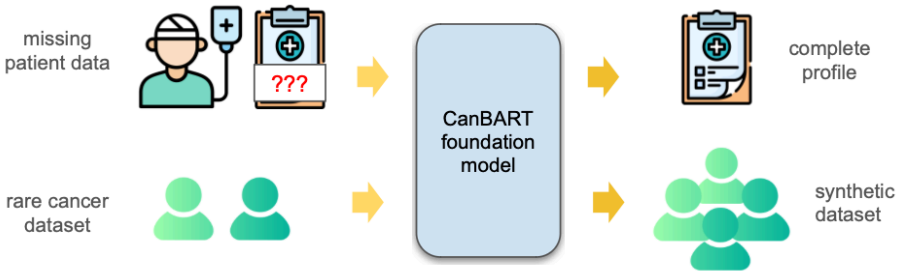
Our study was based on a dataset primarily derived from the MSK-IMPACT panel, comprising genomic profiling results from approximately 80,000 patients (6). These profiles included three main categories of alterations: mutations, copy number alterations (CNA), and structural variants (SV), assessed across up to 505 cancer-related genes. Additional clinical attributes such as patient sex, age, microsatellite instability (MSI), and tumor mutational burden (TMB) were also available for subsets of the cohort. To broaden the dataset, we incorporated samples from the AACR GENIE database, which included genomic profiles for an additional 64,000 patients (7). For consistency, we selected only the overlapping 505 genes between platforms to harmonize representation.

To adapt Transformer-based architectures for clinical genomic data, we designed a sentence-like representation of molecular alterations observed in patient tumor profiles. Instead of encoding profiles as binary feature vectors, which cannot account for multiple alterations affecting the same gene in a single patient, we treated them as sequences of alteration tokens sorted by chromosomal position. Each token consisted of the name of an altered gene and a suffix denoting one of four alteration types: mutation, hotspot, copy number alteration, or structural variant. (Fig. 1). This word-level representation preserved the biological context and structural information across the genome, enabling sequential modeling and generation, and allowing the application of well-established NLP techniques such as masked token prediction and autoregressive sampling.

a



b



c

		encoding
1.	mutations	m
2.	copy number alterations	a
3.	structural variants	v
4.	hotspots	mh

d

[BRAF**m**, CDK6**a**, APC**v**, TP53**mh**, MSI**i**, TMB**h**, SEX**f**, AGEov**65**, 14**q**]

e

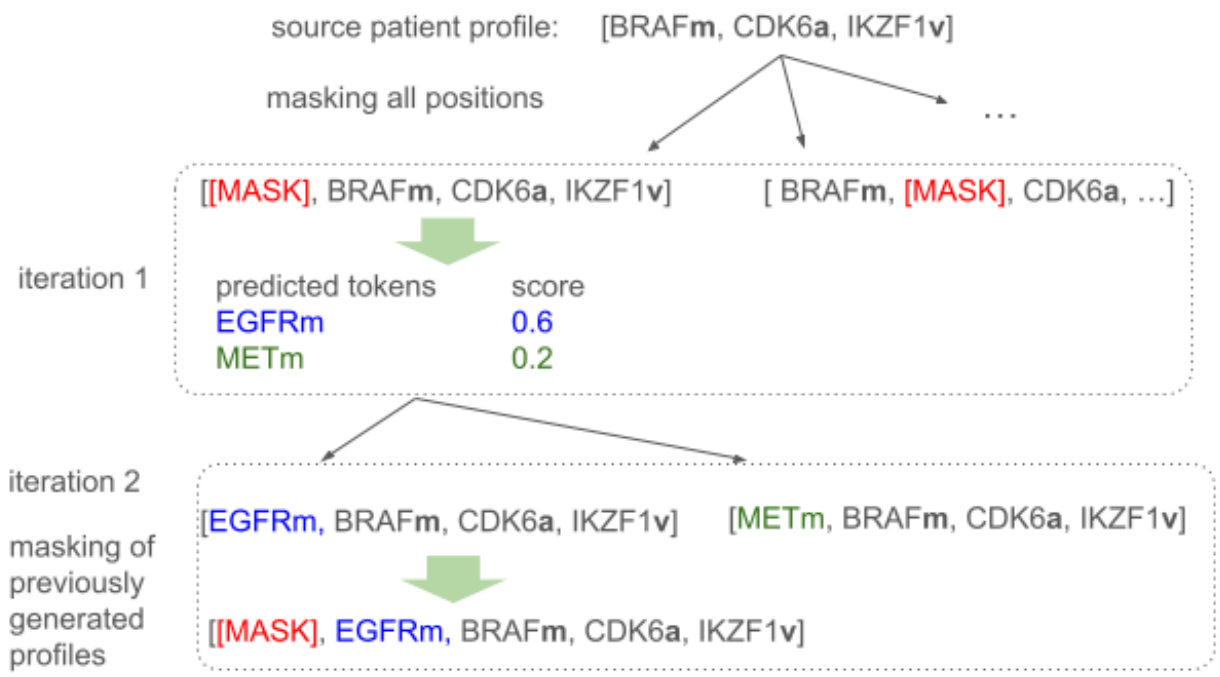


Fig. 1. *Model overview, token construction and patient profile encoding, and sampling algorithm.*
a) Schematic overview of the CanBART training and inference pipeline. The model is trained on genomic alteration profiles from real-world cancer cohorts (MSK-IMPACT, GENIE), using masked prediction tasks to learn alteration co-occurrence patterns. During inference, CanBART receives a partially masked patient profile and generates multiple plausible completions for the missing gene status, each assigned a confidence score (p1, p2).
b) Application of the model for genomic data completion and synthetic cohort generation. The model infers missing gene-level alterations in real patient profiles, enabling harmonization of incomplete genomic datasets. These inferred profiles are further sampled to generate plausible synthetic patients, particularly for rare cancer types with limited available data.
c) Molecular alterations included in the CanBART input representation and their corresponding suffixes used for token encoding. The model incorporates four alteration types: mutations (m), hotspot mutations (mh), copy number alterations (a), and structural variants (v)
d) Example of an encoded patient profile. Each gene is appended with a suffix denoting its specific alteration type. Additional patient-level features are included as separate tokens: microsatellite instability (MSI) status, tumor mutational burden (TMB), sex, age group (e.g., age over 65 years), and arm-level chromosomal alterations (e.g., 14q, 12p), which capture large-scale genomic CNAs.
e) Autoregressive sampling procedure used for generating plausible synthetic patient profiles. Starting from a real patient profile (top), we apply **position-wise masking** by inserting special mask tokens before and after each token in the sequence. This creates multiple candidate positions for token prediction.

During **Iteration 1**, the model predicts candidate tokens for the first masked position. For illustration, the model proposes two possible tokens with scores 0.6 and 0.2. At the same time, a new mask is inserted into the next available position in a left-to-right fashion.
In **Iteration 2**, the sequence is updated to include the previously selected token (e.g., the one with score 0.6), and a new masked position is evaluated. This **autoregressive decoding** process continues, updating the profile step by step.
At each iteration, **token probabilities are multiplied** to compute a **cumulative probability score** for the entire generated sequence. The sampling process proceeds up to a maximum of 50 iterations.
To control diversity and biological plausibility, we apply **top-p (nucleus) sampling**: at each masked position, the model selects from the smallest set of tokens whose **total probability mass** (i.e., sum of predicted scores) exceeds a predefined threshold $p = 0.75$. This ensures a balance between high-probability predictions and controlled randomness during generation.

Model Architecture

We based our approach on BART, a sequence-to-sequence Transformer architecture with a bidirectional encoder and autoregressive decoder (8). The model consisted of 3 encoder and 3 decoder layers, each with 1024-dimensional hidden states and 8 attention heads. A custom WordLevel tokenizer was trained on alteration tokens, resulting in a vocabulary of approximately 2,100 tokens.
The model was initially pre-trained using a masked language modeling (MLM) objective on over 144,000 tumor profiles from the MSK-IMPACT and AACR GENIE databases (Fig 1a, b). The MLM objective was designed to enable the model to learn combinatorial genomic patterns by reconstructing masked alterations in patient profiles represented as token sequences, and to build patient-level embeddings that support multiple downstream applications—similar to how foundation models are structured in NLP.
After pretraining, the model was fine-tuned for a tumor-type classification task. Patient profiles were mapped to 56 tumor types, each represented by at least 15 samples. Among these, nine cancer types with at least 6,000 samples each were designated as common, while the remaining 47 were considered rare and were analyzed in two separate experiments: first, as a single pooled rare cancer category contrasted against common cancers, in order to evaluate the model's ability to distinguish rare from common types and to explore the extent of shared genomic signals, and second, as 47 distinct rare cancer classes to assess fine-grained classification performance.

Plausible Patient Generation

To improve classification performance for rare cancers — especially those with limited data and insufficient training signal — we first sought to generate synthetic training examples for underrepresented tumor types. For this, we used the pre-trained model as a foundation, applying it without fine-tuning to generate new plausible profiles.

We implemented a masked autoregressive sampling strategy, inspired by text generation in NLP. Starting with a real patient profile, we iteratively masked one token at a time and sampled replacements from the model's output distribution. Sampling followed a nucleus (top-p) strategy to promote diversity and avoid mode collapse.
Each generated profile was scored by its cumulative generation probability, calculated as the product of token probabilities at each generation step, where the model predicts the likelihood of each token at the masked position. A maximum of 50 iterative steps was used for generating synthetic patient profiles. Generation was stopped once the cumulative probability fell below a predefined threshold. This threshold was selected empirically for each cancer type by evaluating classification accuracy across multiple values, as described in the next section.

Rare Cancer Classification with Synthetic Patients

We evaluated whether augmenting training data with synthetic rare cancer patients could improve classification performance. Rare cancers were defined as types with fewer than 6,000 samples and grouped into multiple size-based bins (e.g., 15-50, 50–150, 150–250, etc.). For each bin, experiments were conducted across varying numbers of cumulative probability thresholds, and—consequently—of the number of generated patients. The goal was to identify an optimal plausibility cutoff for each cancer type, based on classification accuracy.

Fine-tuning was performed using five-fold cross-validation. In each fold, a separate test set of real patient profiles was held out and not seen during model training. Accuracy was used as the evaluation metric and was reported per fold and averaged across all folds. The pre-trained model used as the base was trained on the full dataset prior to folding.

To ensure a fair comparison, we also tested a control condition where synthetic samples were replaced by duplicated real patients to keep dataset sizes equal.

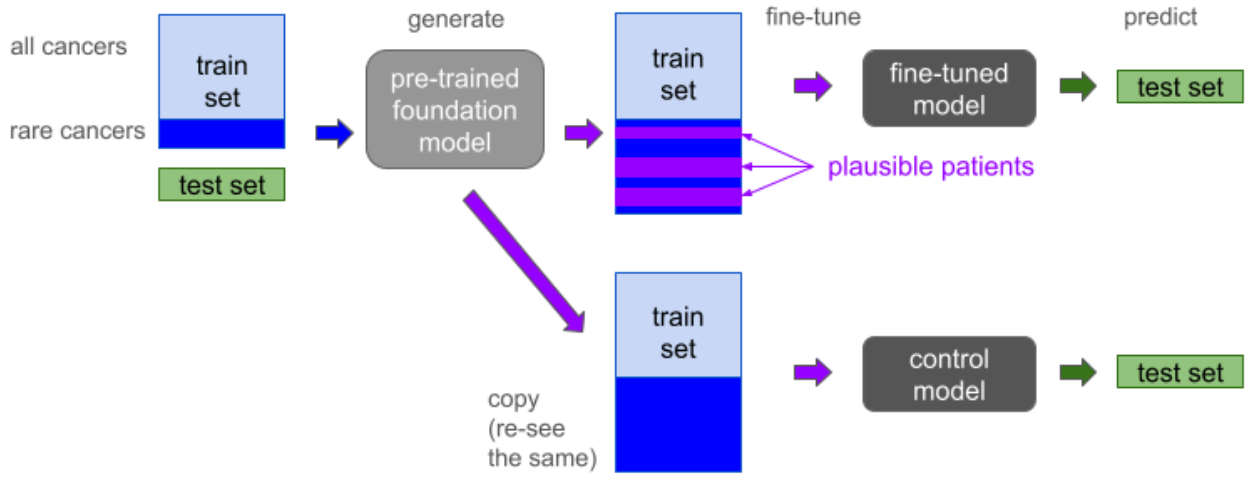


Fig. 2. Overview of the CanBART training pipeline for rare cancer classification.

The initial dataset is split into training and test sets (80/20), with rare cancer profiles explicitly marked within the training set. The training data is input into a pre-trained foundation model, which generates plausible patients (PPs) for the rare cancer types. Two parallel augmentation strategies are then applied:
(1) *Generative augmentation* — the training set is expanded by adding newly generated PPs for rare cancers;
(2) *Copy-based augmentation (control)* — the training set is expanded by duplicating existing rare cancer profiles to match the volume of PPs in (1).
Each augmented dataset is then used to fine-tune the model: resulting in a **fine-tuned model** (with synthetic data) and a **control model** (with duplicated data). Both models are evaluated on the same held-out test set to assess classification performance.

Missing Gene Imputation

To test the model's capacity to infer unobserved genomic alterations, we simulated incomplete panel testing by restricting profiles to the 278 genes shared with the DFCI panel and evaluating the model's ability to reconstruct the remaining 1,067 alterations across 181 genes from the MSK-IMPACT set. For this task, the pre-trained model was evaluated in a five-fold cross-validation setting, where in each fold the test set was treated as a simulated DFCI panel and excluded from training. Each missing gene was predicted in a binary classification setup, using the model's masked token prediction probability as a confidence score. Performance was evaluated using ROC-AUC across genes. This task simulates real-world scenarios in which genomic panels are incomplete. The model’s imputation ability supports clinical use cases such as treatment selection and trial matching.

Results

Rare Cancer Classification

We evaluated classification performance in the pooled rare versus common cancer task. The 10-class confusion matrix is shown in Fig. 3a, and Fig. 3b displays the results for high-confidence predictions only (probability > 0.75). To visualize the internal structure of the learned representations, we projected the encoder’s hidden states using UMAP (Fig. 4a) (9). This revealed that while some rare cancers cluster tightly—likely reflecting distinct molecular subtypes or well-represented rare types—many others are interspersed among common cancer clusters. These "signal-aligned" rare cancers appear adjacent to common types, suggesting latent proximity driven by shared genomic features and further supporting the idea that common cancers serve as anchors in the learned embedding space.

a

threshold = 0 (100% of patients covered)											0.71	- overall accuracy
		predicted										
		9	5	1	0	3	7	4	6	2	8	total samples	cancer type
real	9	75	6									42154	All rare
	5	19	64				6					23302	Non-Small Cell Lung Cancer
	1	9		78								17448	Colorectal Cancer
	0	13			74				7			14824	Breast Cancer
	3	20				70						10463	Glioma
	7	12	11	7		0	67					8159	Pancreatic Cancer
	4	22						66				6965	Melanoma
	6	19	7		11				49	9		6935	Ovarian Cancer
	2	7			6				9	69		6179	Endometrial Cancer
	8	25									70	6133	Prostate Cancer

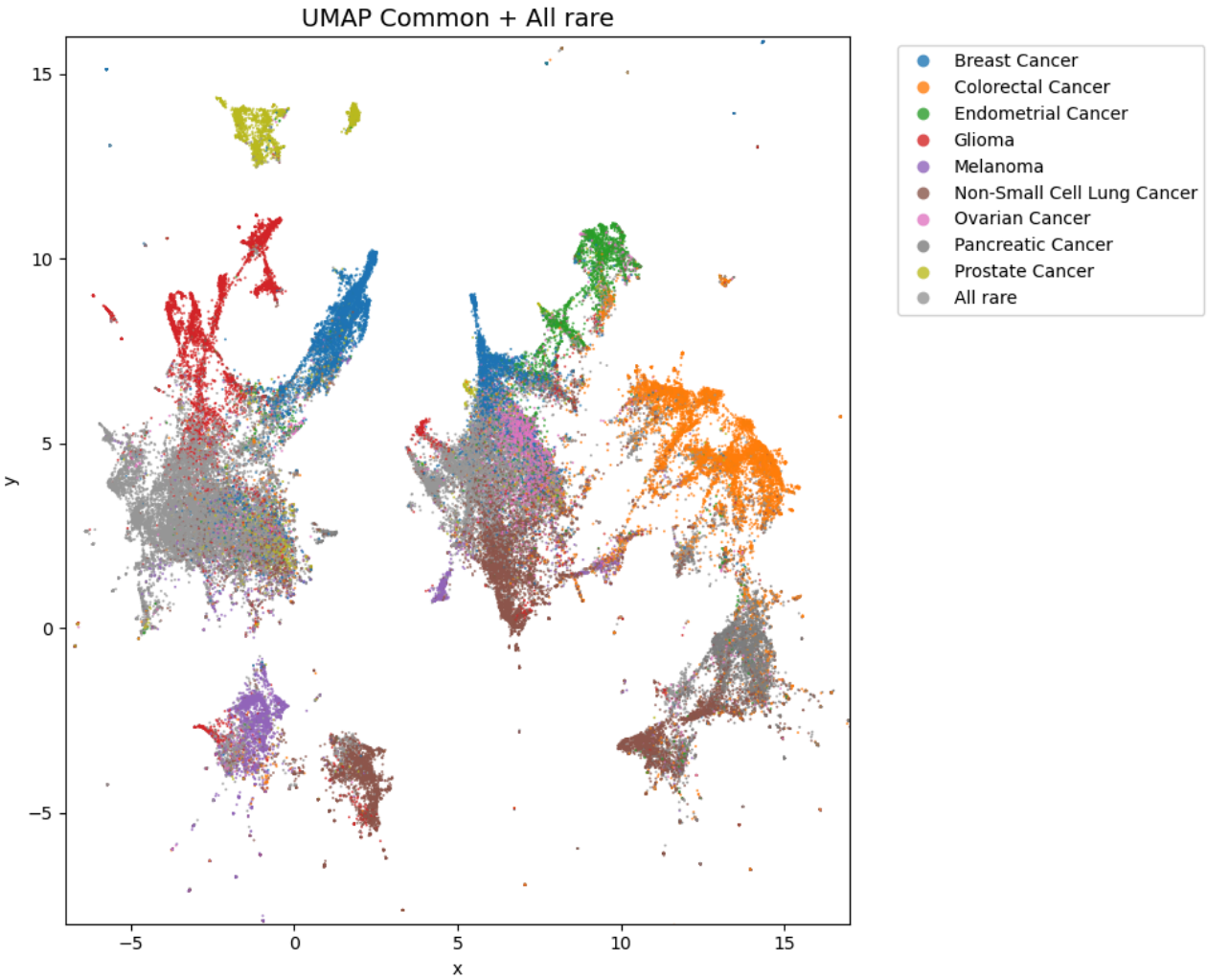
b

threshold = 0.75 (55% of patients covered)											0.88	- overall accuracy
		predicted										
		9	1	5	0	3	8	2	7	4	6	total samples	cancer type
real	9	87										21306	All rare
	1		94									12407	Non-Small Cell Lung Cancer
	5	8		85								11920	Colorectal Cancer
	0				92							9377	Breast Cancer
	3	10				87						7061	Glioma
	8	6					92					4267	Prostate Cancer
	2							91				3671	Melanoma
	7	7							84			3583	Endometrial Cancer
	4	10								85		3401	Pancreatic Cancer
	6	11			7			15			57	1924	Ovarian Cancer

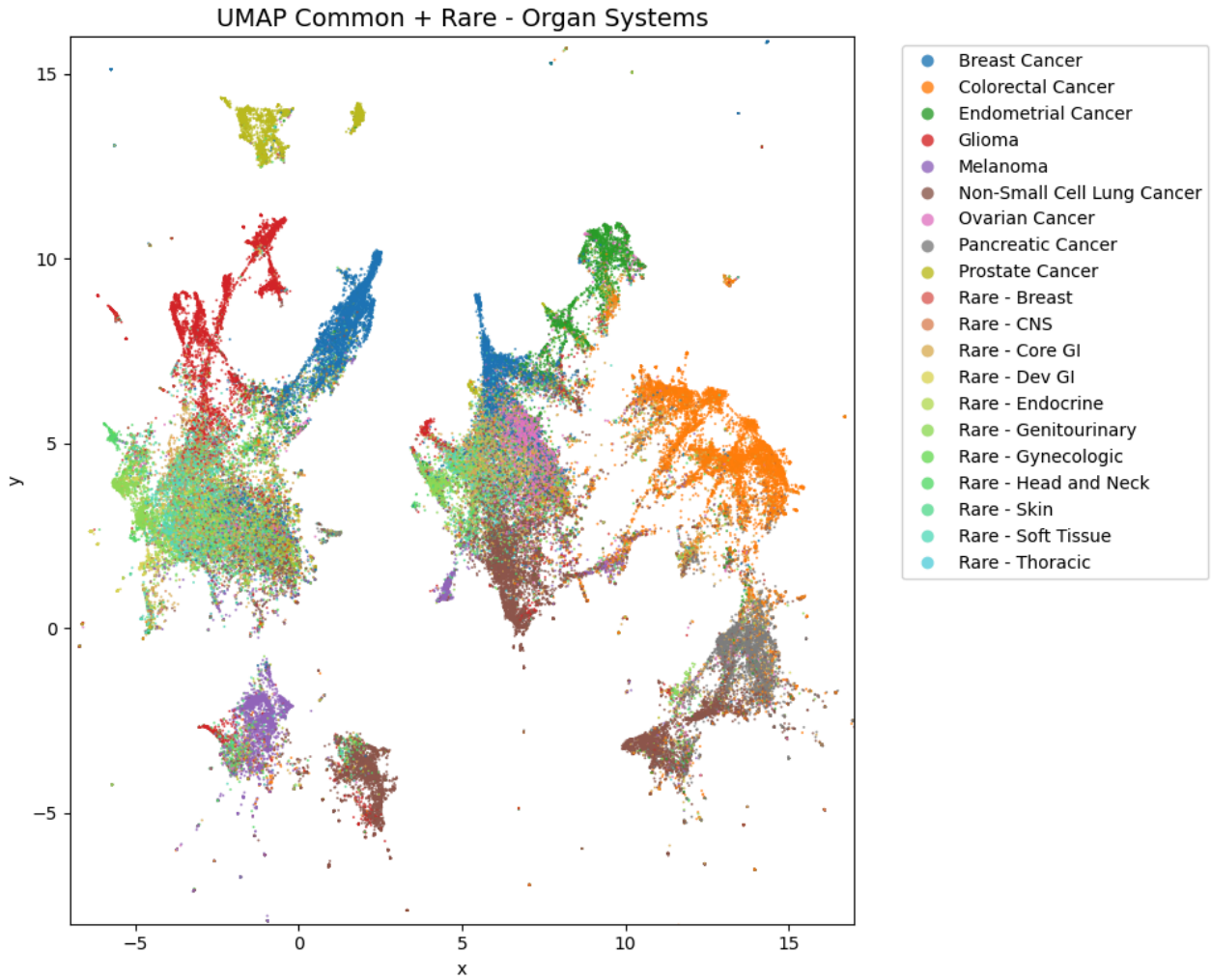
Fig. 3. Classification performance of CanBART on common vs. rare cancer groups.

Two confusion matrices are shown for the 9 most common cancer types versus the combined group of all rare cancers:
a — predictions for the entire test set.
b — predictions filtered to include only those with high confidence (softmax probability > 0.75). Values ≤ 5 were omitted for clarity.

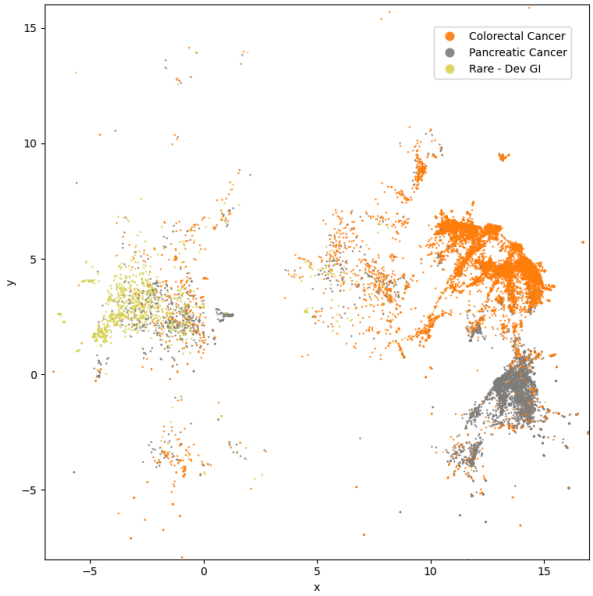
a


UMAP Common + All rare
Breast Cancer
Colorectal Cancer
Endometrial Cancer
Glioma
Melanoma
Non-Small Cell Lung Cancer
Ovarian Cancer
Pancreatic Cancer
Prostate Cancer
All rare
x
y
−5
0
5
10
15

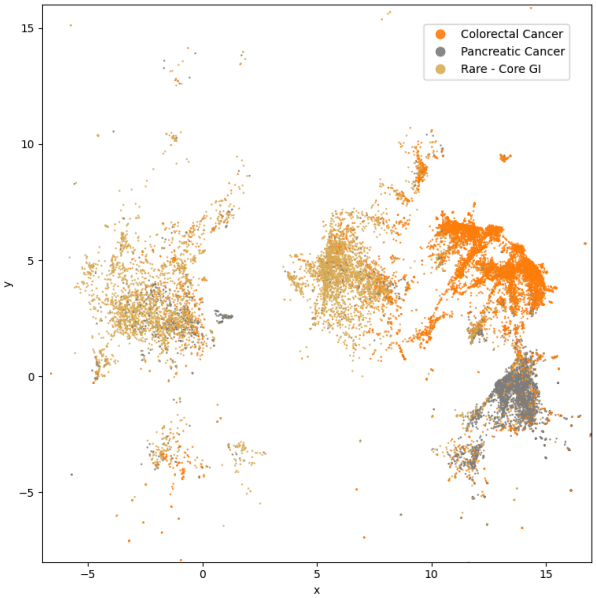
b


UMAP Common + Rare - Organ Systems
Breast Cancer
Colorectal Cancer
Endometrial Cancer
Glioma
Melanoma
Non-Small Cell Lung Cancer
Ovarian Cancer
Pancreatic Cancer
Prostate Cancer
Rare - Breast
Rare - CNS
Rare - Core GI
Rare - Dev GI
Rare - Endocrine
Rare - Genitourinary
Rare - Gynecologic
Rare - Head and Neck
Rare - Skin
Rare - Soft Tissue
Rare - Thoracic
x
y

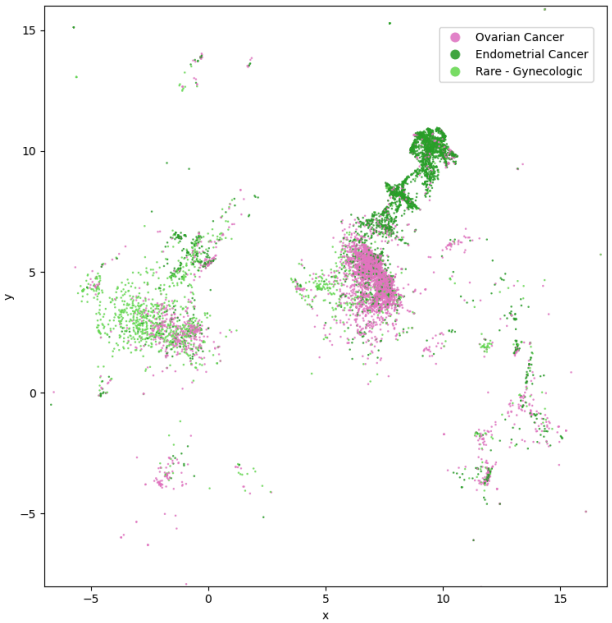
c


Colorectal Cancer
Pancreatic Cancer
Rare - Dev GI
x
y

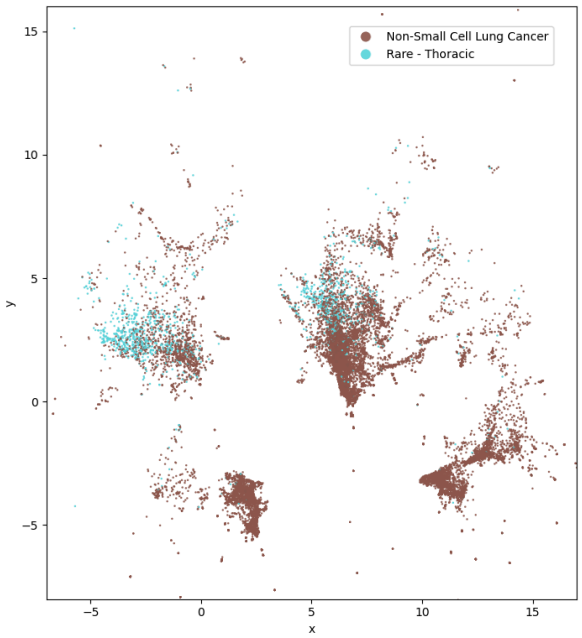
d


Colorectal Cancer
Pancreatic Cancer
Rare - Core GI
x
y

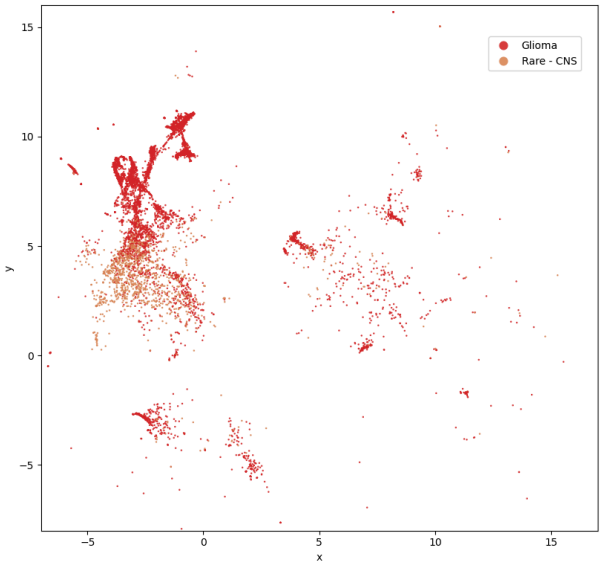
e


Ovarian Cancer
Endometrial Cancer
Rare - Gynecologic
x
y

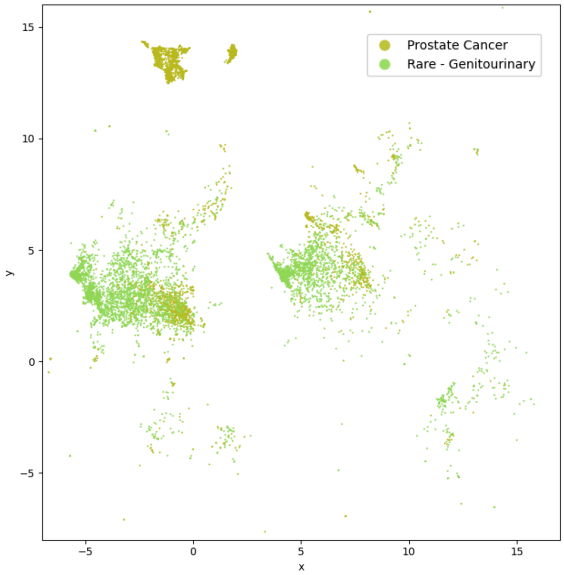
f


Non-Small Cell Lung Cancer
Rare - Thoracic
x
y

g


Glioma
Rare - CNS
x
y

h


Prostate Cancer
Rare - Genitourinary
x
y

i

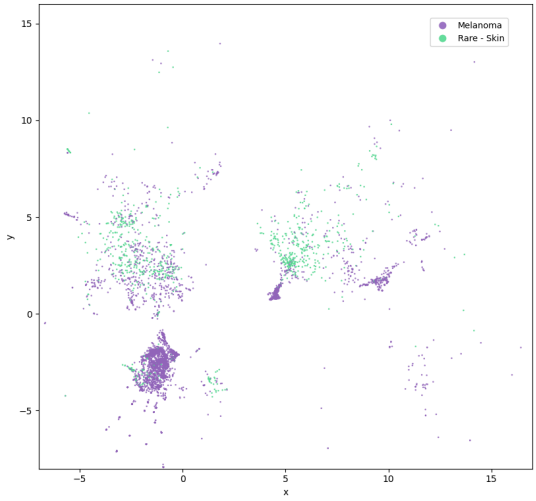


Fig 4. UMAP projections of patient embeddings for common and rare cancers, grouped by anatomical origin.

a) Two-dimensional UMAP of latent representations for the 9 most common cancer types (each with >6,000 samples, colored by class) and **all rare cancers aggregated as a single group** (grey). A subset of rare cancers clusters separately on the left, while others are interspersed across the map and overlap with each of the common types.

b) UMAP projection of the same space, but with rare cancers labeled and colored by **organ system** (11 groups in total), revealing structured heterogeneity within the rare cancer space.

c–i) Zoomed-in UMAP projections for specific **organ-system-level comparisons** between common cancers and anatomically related rare cancers:

- **c**: colorectal and pancreatic cancers, together with rare *gastrointestinal–developmental* tumors;
- **d**: colorectal and pancreatic cancers with rare *gastrointestinal–core* tumors;
- **e**: ovarian and endometrial cancers with rare *gynecologic* tumors;
- **f**: non–small cell lung cancer (NSCLC) with rare *thoracic* tumors;
- **g**: gliomas with rare *central nervous system (CNS)* tumors;
- **h**: prostate cancer with rare *genitourinary* tumors;

- **i**: melanoma with rare *cutaneous* tumors.

In most cases, rare tumors show **partial or substantial overlap** with common cancers from the same anatomical system. This suggests that shared molecular patterns may be captured by the

model and **transferred to rare cancers during training**, supporting the use of generative learning for biologically informed data augmentation.

We then evaluated CanBART's ability to classify 47 rare cancer types individually. The confusion matrix summarizing final performance after augmentation with synthetic patients is shown in Fig. 4b. A histogram of per-class changes in accuracy is provided in Fig. 4c: 28 cancer types showed improvement, 14 decreased, and 5 remained unchanged. The number of synthetic patients added per cancer type (averaged across five folds) is also indicated.

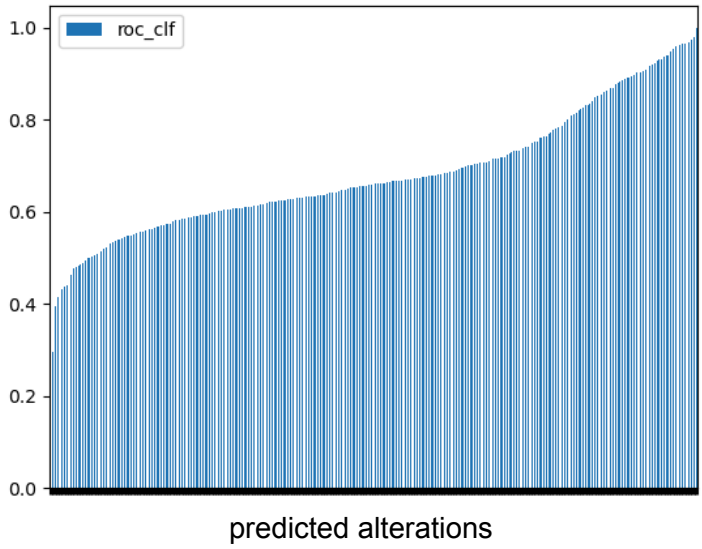
To control for dataset size, we conducted a parallel experiment using the same classification task across the 47 rare cancer types, where the number of training samples was augmented by simply duplicating existing real patients instead of generating new synthetic ones. The results showed that performance **decreased** for **30 cancer types**, with an average drop in accuracy of **39%** (*95% CI: 27–50%*). For **10 of the more common rare types**, accuracy slightly improved, though never by more than **6%**. For the remaining **7 types**, performance remained unchanged.

Optimal cumulative probability thresholds for selecting plausible patient numbers varied across cancer types and were determined empirically for each bin.

Missing Gene Imputation

We evaluated CanBART's ability to impute missing gene alterations in incomplete patient profiles. The model was tested on 1,067 alterations masked from the MSK-IMPACT panel. For each alteration, prediction confidence was scored using ROC-AUC. Among events observed in at least 15 patients, 33% were recovered with ROC-AUC ≥ 0.75 (Fig. 5). Performance decreased for extremely rare alterations, consistent with the hypothesis that sufficient representation is essential for imputation. These results highlight CanBART's potential to harmonize incomplete sequencing data and support decision-making in underprofiled clinical cases.

a



b

Alteration	ROC	# of samples with this alteration
CSDE1a	1	146
FGF19a	0.99	3030
ALOX12Ba	0.98	2304
TRIP13a	0.98	476
DCUN1D1a	0.98	476
DUSP4a	0.97	845
SYKa	0.97	1137
GPS2a	0.97	1310
ERFa	0.97	258
NCOR1a	0.97	1477

c

Alteration	ROC	# of samples with this alteration

EIF1AXm	0.86	2095
BCL10m	0.76	446
DCUN1D1m	0.76	274
SPENm	0.75	6419
MLLT1m	0.75	467
RRASm	0.75	283
CDC42m	0.74	173
DUSP4m	0.74	467
ERRFI1m	0.73	890
GNB1m	0.73	133

d

Alteration	ROC	# of samples with this alteration
TSHRv	0.92	28
MDM2v	0.88	116
MSI2v	0.85	28
NUP93v	0.84	35
WHSC1L1v	0.83	68
SETDB1v	0.83	32
CDK4v	0.81	60
CDKN1Av	0.8	26
TAP2v	0.78	24
DNAJB1v	0.78	66

e

Alteration	ROC	# of samples with this alteration
KNSTRNmh	0.98	35
SOX17mh	0.83	21
SMAD3mh	0.82	49

INHBAmh	0.8	32
PIK3R2mh	0.79	22
SYKmh	0.77	20
EIF1AXmh	0.76	142
EPHB1mh	0.75	25
PRDM14mh	0.73	15
SMARCD1mh	0.72	21

Figure 5. ROC AUC distribution and top-performing alterations in gene status completion.

a) Histogram of ROC AUC scores for individual molecular alterations in the gene status completion task, sorted in ascending order (left to right). To simulate a targeted panel setting, we constructed a dataset by removing genes not included in the DFCI panel from each sample, and trained the model to predict their status.
Alterations with fewer than 15 positive training examples were excluded from evaluation, consistent with the threshold used in the classification task.
Overall, 33% of alterations were predicted with a ROC AUC ≥ 0.75, indicating strong performance on a substantial subset of clinically relevant targets.

b–e) Top 10 predicted alterations by ROC AUC, stratified by alteration type:
b — copy number alterations (CNAs),
c — truncating mutations,
d — structural variants (SVs),
e — hotspot mutations.

Discussion

Generalization to Rare Cancers

This study presents CanBART, a Transformer-based foundation model trained on real-world genomic profiling data, designed to address key limitations associated with datasets for rare cancers and other underrepresented patient groups where available samples are insufficient for conventional modeling. By modeling molecular alterations as tokenized sequences and leveraging a BART-based architecture, CanBART bridges generative modeling in the clinical cancer genomics domain.

A core result of our study is that representations learned from common cancer types can generalize to rare ones. This was demonstrated in pooled classification and latent space visualization: many rare tumors clustered near common cancers, suggesting the presence of

transferable genomic signals. These signal-aligned rare cancers may share driver events, mutational patterns, or structural variants with more prevalent tumor types, allowing the model to anchor their representations through shared latent features. This supports prior hypotheses proposed in the GDD-ENS study and offers quantitative evidence that Transformer-based models can exploit implicit similarity across cancer types for data-efficient learning (5).

Synthetic Patient Augmentation

Importantly, this shared signal is not only visible in the latent structure—it also manifests functionally. Synthetic patients generated by CanBART, inheriting patterns learned from common tumors, led to measurable accuracy improvements when added to training data for rare cancer classification. This demonstrates that the generated profiles carry meaningful, cancer-specific signal rather than generic noise. In our setup, the classification task was not intended to yield a clinically deployable classifier, but rather to validate the informativeness of synthetic examples. Performance for some rare types remained modest, largely due to extremely small training cohorts.

However, the consistent gains achieved through generative augmentation justify the use of plausible patients in tasks where annotated datasets are scarce. These synthetic cohorts could support a wide range of downstream applications in settings where real-world data is limited. Crucially, the success of this approach confirms that the sentence-like representation of patient data, composed of molecular alterations and key clinical attributes is expressive enough to capture both tumor identity and the molecular portrait of individual cases. This implies that synthetic patients can participate meaningfully in digital experimentation and serve as useful proxies in planning clinical trials or modeling therapeutic response. They may be used to discover novel biomarkers by revealing consistent alteration patterns within rare subtypes, to simulate therapy response distributions in silico for trial planning, or to create virtual control arms for clinical studies. Moreover, they could serve as training data for predictive models in settings like progression-free survival, resistance profiling, or drug repurposing for underrepresented cancer populations. As models like CanBART become more refined, synthetic cohorts may also facilitate functional hypothesis generation, safety profiling, and prioritization of targets for experimental validation.

Model Limitations and Platform Heterogeneity

The observed classification performance should be interpreted in the context of several compounding factors. First, the dataset used for training included not only MSK profiles but also samples from dozens of institutions via the GENIE consortium (7). This increased heterogeneity makes the model more generalizable but also introduces substantial batch effects. Second, the platform diversity within GENIE is substantial, with over a hundred gene panels represented — some as small as 50 genes. This variability likely contributes to misclassification, particularly among patient profiles with minimal profiling coverage. Importantly, the strongest accuracy improvements from synthetic patient augmentation were observed in cancers with 20 to 500 samples, indicating that the generative model is most effective in filling low-data regimes.

However, we also observed some misclassification among more common cancers, likely due to signal entanglement introduced by augmentation. One potential solution may involve stratified training, with separate classifiers optimized for ultra-rare, rare, and common cancer groups to mitigate interference between signal-rich and signal-poor classes.

Missing Data Imputation

To further examine the biological grounding of our generative mechanism, we evaluated whether the same architecture could be used to predict unobserved molecular alterations. In a panel reduction setting, CanBART accurately recovered a substantial fraction of masked events, suggesting that the model internalized co-occurrence patterns from the training corpus. This not only supports the plausibility of the synthetic profiles used in classification tasks but also demonstrates the model's utility for gene-level inference. This capability is becoming increasingly important given the diversity of sequencing platforms in clinical use and the fragmentation of genomic coverage across institutions. Foundation models like CanBART offer a promising route toward cross-platform harmonization, enabling robust comparisons and unified representation across incomplete profiles.

Conclusion

Together, these results establish CanBART as a flexible foundation model for genomic oncology and computational drug discovery: capable of classifying rare tumors, simulating plausible cohorts, and imputing incomplete profiles using a shared generative engine. More broadly, our work illustrates how generative pretraining on real-world cancer data can support a spectrum of downstream applications without requiring handcrafted features or extensive per-task engineering. As genomic data collection continues to scale, foundation models like CanBART may become essential infrastructure for harmonized learning, interpretation, and synthetic data generation in precision oncology.

References

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2.
2. Avsec Ž, Agarwal V, Visentin D, Ledsam J, Grabska-Barwińska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. 2021;18(10):1190-1197. doi:10.1038/s41592-021-01252-x.
3. Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv*. 2022. Available from: https://arxiv.org/abs/2204.00300.

4. Fradkin P, Bhatia H, Linder J, Morris Q. Orthrus: towards evolutionary and functional RNA foundation models. *bioRxiv*. 2024. doi:10.1101/2024.10.10.617658. PMID:39416135.
5. Darmofal M, et al. Deep-learning model for tumor-type prediction using targeted clinical genomic features (GDD-ENS). *medRxiv*. 2024. doi:10.1101/2024.02.23.24303165.
6. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med*. 2017;23(6):703–713. doi:10.1038/nm.4333.
7. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov*. 2017;7(8):818–831. doi:10.1158/2159-8290.CD-17-0151.
8. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *ACL*. 2020;7871–7880. doi:10.18653/v1/2020.acl-main.703.
9. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*. 2018. Available from: https://arxiv.org/abs/1802.03426.

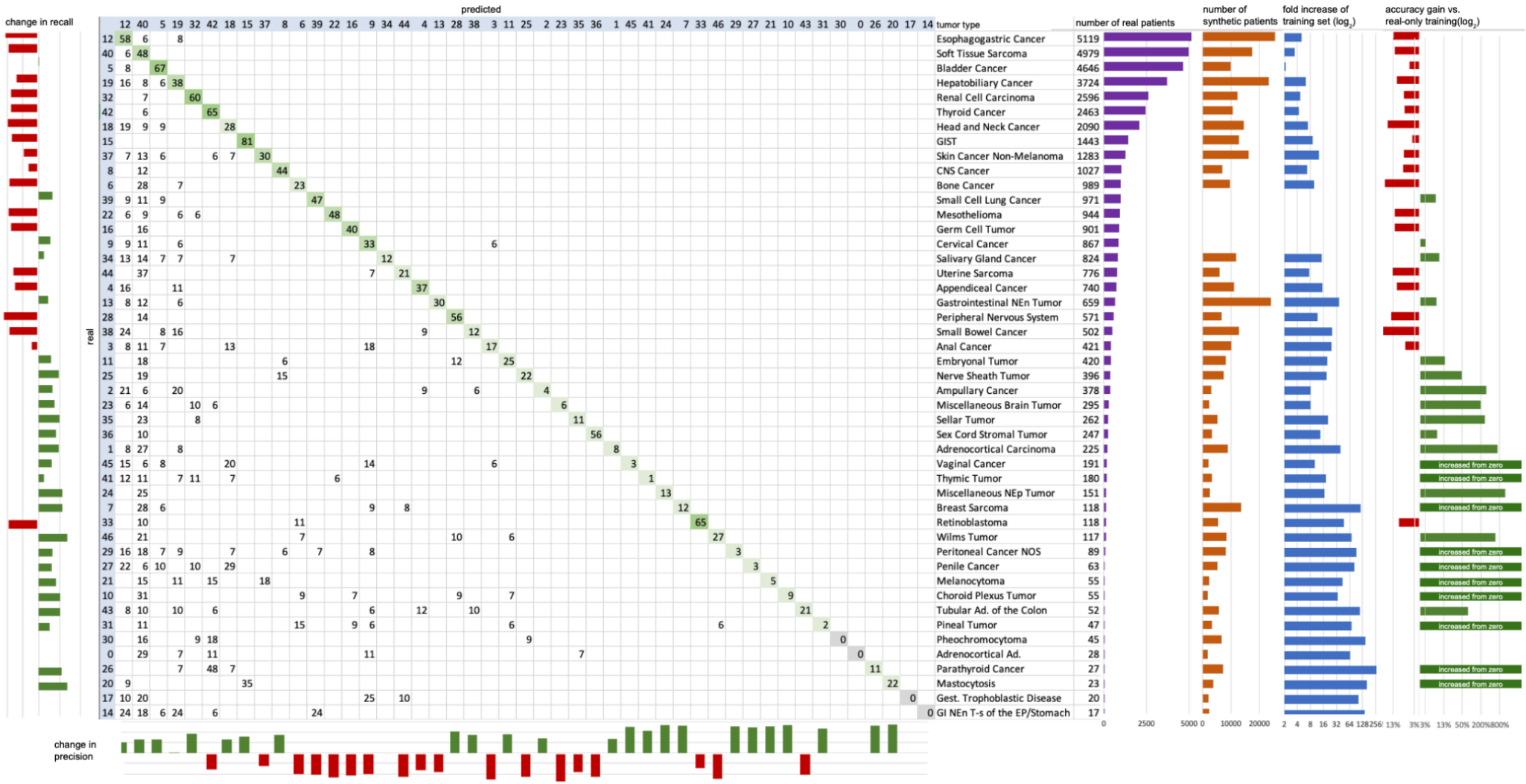


Fig. 6. Performance of CanBART in rare cancer classification after synthetic patient augmentation.
Confusion matrix showing predictions across **47 rare cancer types**, with true labels sorted top to bottom by the number of available real patients per class.

To the right of the matrix, four vertically aligned histograms provide additional context for each cancer type:

- **First column**: Number of real (original) patients per class in the training set.
- **Second column**: Number of synthetic patients generated for each cancer type through an optimization-guided augmentation strategy.
- **Third column**: Fold increase in training set size per cancer type after augmentation, shown on a **$\log_2$ scale**.
- **Fourth column**: **Change in classification accuracy**, measured as the difference from baseline accuracy achieved when training only on real patients (also shown on a $\log_2$ scale). Cases where accuracy increased from zero to a non-zero value are specially highlighted.

This visualization demonstrates how generative augmentation improves classification performance for ultra-rare cancers, many of which had very limited training data initially.