A Deep Generative Mixture Model for Enhancing Circulating Tumor DNA Estimation

Anonymous Authors[1]

Abstract

Circulating tumor DNA (ctDNA) provides a minimally invasive measure of tumor burden, but estimating ctDNA fraction from plasma is difficult when tumor-derived molecules are rare relative to the cell-free DNA background. We introduce a deep generative mixture model that estimates sample-level ctDNA fraction from read-base events in whole-genome sequencing (WGS). Each event is modeled as a noisy observation from healthy-derived or tumor-derived components, mixed by the sample-specific ctDNA fraction. To capture batch effects and patient-specific background noise, sample embeddings are learned jointly with a decoder that uses read-level covariates, sequence context, and tumor-catalog information to model allele probabilities. Self-normalized inverse-probability weighting handles stratified genome-wide sampling, and one trained model supports tumor-informed and tumor-agnostic inference. We apply the framework to serial plasma WGS from a stage III colorectal cancer cohort.

1. Introduction

Circulating tumor DNA (ctDNA) consists of tumor-derived fragments within cell-free DNA (cfDNA), which is otherwise dominated by DNA from healthy tissues, and can be identified through somatic mutations and other tumor-associated alterations (Wan et al., 2017).

Because ctDNA can be measured from plasma, it provides a minimally invasive biomarker for detecting and tracking tumor burden (Wan et al., 2017). Key applications include minimal residual disease detection, treatment-response monitoring, recurrence surveillance, and early cancer detection (Pantel & Alix-Panabières, 2019; 2025; Phallen et al., 2017). Across these settings, the central inference problem is to estimate a minute tumor-derived component within a much larger healthy-derived cfDNA background (Bettegowda et al., 2014).

The challenge is not only the rarity of ctDNA, but also the heterogeneity of the background. At low ctDNA fractions, tumor-derived observations must be separated from sequencing error, DNA damage, germline variation, clonal hematopoiesis, and technical noise. These processes vary across sequence contexts, labs, protocols, sequencing runs, patients, and samples, so the same apparent variant evidence can carry different meaning in different cfDNA samples (Christensen et al., 2023; Stoler & Nekrutenko, 2021).

Existing whole-genome sequencing (WGS)-based ctDNA methods address sensitivity by aggregating weak evidence across many loci, either using patient-specific mutation catalogs or tumor-agnostic genome-wide signals (Zviran et al., 2020; Widman et al., 2024; Nordentoft et al., 2024; Reinert et al., 2019; Christensen et al., 2019; Wan et al., 2020; Zhu et al., 2025). These approaches establish the value of genome-wide evidence, but leave an important modeling question: how should noisy read-base observations be weighted when both tumor signal and unlabelled sample-specific background factors vary?

We approach this as a generative modeling problem. We develop a deep generative mixture model for sample-level ctDNA estimation from WGS data, where read-base events (matches and mismatches mapped to the reference genome) are modeled as observations from healthy-derived and tumor-derived components mixed by the sample-specific ctDNA fraction. The model uses the deep generative decoder idea of learning sample representations jointly with decoder parameters (Schuster & Krogh, 2023) within an event-level mixture likelihood. Event-specific covariates and sample-specific latent representations parameterize allele probabilities, allowing sample-specific background structure to be separated from the ctDNA mixture proportion. When available, tumor mutation catalogs are included as position-level conditioning variables, enabling tumor-informed and tumor-agnostic inference with the same model.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

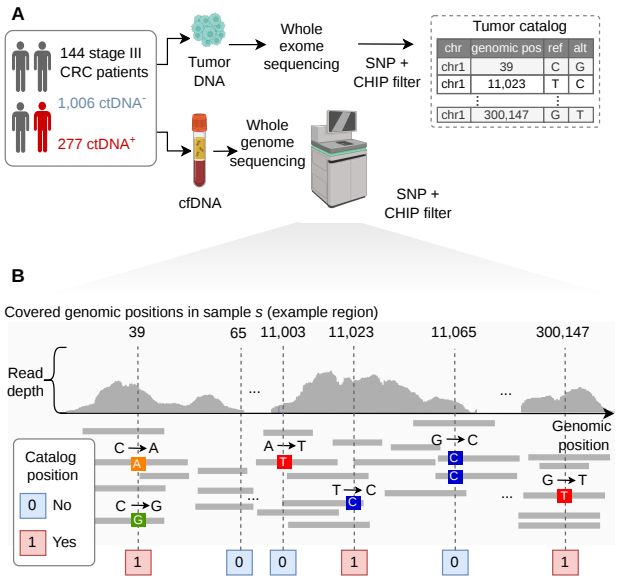

Figure 1. **Data pipeline for event-level ctDNA modeling.** Tumor sequencing defines patient-specific mutation catalogs, while matched cfDNA WGS provides read-base events with covariates, position-level catalog membership C_{si}, and expected alternate-allele encodings a_{sj}.

2. Materials and methods

2.1. Clinical Data

The analysis is based on serial plasma WGS data from the stage III colorectal cancer cohort described by Frydendahl et al. (Frydendahl et al., 2024). The dataset contains 1,283 plasma samples from 144 patients, sequenced at approximately 20× genome-wide depth. Samples were collected across clinically relevant time points, including pre-operative, post-operative, adjuvant chemotherapy, post-adjuvant chemotherapy, and surveillance draws. Using the ctDNA status labels associated with the cohort, 277 samples were ctDNA-detected and 1,006 were non-detected. For model evaluation, samples were assigned to stratified training (80%) and test (20%) sets at the patient level, ensuring that all samples from a given patient were kept within the same split. Stratification used patient-level covariates, including age, sex, disease-stage information, and ctDNA amount. The data generation pipeline is summarized in Figure 1.

2.2. Data and Extracted Covariates

Consider S samples indexed by $s \in \{1, \ldots, S\}$. From each sample, we sample genomic positions indexed by i using a stratified design. Positions are first partitioned by membership in the patient-specific tumor mutation catalog: $C_{si} = 1$ if $i \in \text{catalog}(s)$ and $C_{si} = 0$ otherwise. For stratum $c \in \{0, 1\}$, let $P_{sc,\text{total}}$ denote the number of eligible positions in sample s and let $P_{sc,\text{keep}}$ denote the number retained by the sampler. The stratum-specific retention fraction is $q_{sc} = P_{sc,\text{keep}}/P_{sc,\text{total}}$, for $c \in \{0, 1\}$. Thus, q_{s1} is the retained fraction of catalog positions and q_{s0} is the retained fraction of non-catalog positions. Retained positions receive inverse-probability weights (IPW) $w_{si} = 1/q_{s,C_{si}}$ to correct for the stratum-specific sampling design.

After position subsampling, sample s contains m_s read-base events indexed by $j \in \{1, \ldots, m_s\}$, with total event count $M = \sum_{s=1}^{S} m_s$. In event-level expressions, i denotes the retained genomic position overlapped by event j; all events overlapping position i share the same C_{si} and w_{si}.

Each event is associated with a covariate vector $x_{sj} \in \mathbb{R}^d$, which captures local sequence context and technical variables such as fragment length, local GC content, base quality, distance to read end, mapping quality, homopolymer run length, read depth, replication timing, and related regional or read-level features. Catalog positions additionally carry the expected tumor alternate allele a_{sj}.

Further details on data preprocessing are provided in Appendix A.

2.3. Model Architecture and Sample Representation

Each sample s is associated with a learned latent embedding $z_s \in \mathbb{R}^r$, capturing sample-specific biological, technical, and background-noise characteristics, and a scalar logit parameter $\eta_s \in \mathbb{R}$, representing the sample-specific ctDNA fraction in log-odds space.

The ctDNA fraction is obtained as $\alpha_s = \sigma(\eta_s) = 1/(1 + e^{-\eta_s})$, with $\alpha_s \in (0, 1)$.

For each event, covariates, sample embedding, position-level catalog indicator, and expected tumor alternate allele are processed by a shared backbone, $h_{sj} = \text{backbone}_\theta([z_s; x_{sj}; C_{si}; a_{sj}])$, where $h_{sj} \in \mathbb{R}^k$ (Figure 2).

The output parameterization follows a logit-delta design with dual cancer shifts: the healthy-component head f_h maps h_{sj} to $\text{logits}_{sj}^{(h)} = f_h(h_{sj})$. Two additional neural heads, Δf_{marg} and Δf_{cat}, produce the marginal and catalog-specific cancer shifts $\Delta_{sj}^{\text{marg}} = \Delta f_{\text{marg}}(h_{sj})$ and $\Delta_{sj}^{\text{cat}} = \Delta f_{\text{cat}}(h_{sj})$. The total cancer shift is $\Delta_{sj}^{\text{tot}} = \Delta_{sj}^{\text{marg}} + C_{si}\Delta_{sj}^{\text{cat}}$, giving cancer logits $\text{logits}_{sj}^{(c)} = \text{logits}_{sj}^{(h)} + \Delta_{sj}^{\text{tot}}$ (Figure 2).

2.4. Likelihood with Inverse-Probability Weighting

For $b \in \{A, C, G, T\}$, the healthy and cancer component distributions are given by $p_h(y = b \mid x_{sj}, C_{si}, a_{sj}, z_s) =$

$\mathrm{softmax}(\mathrm{logits}_{sj}^{(h)})_b$ and $p_c(y = b \mid x_{sj}, C_{si}, a_{sj}, z_s) = \mathrm{softmax}(\mathrm{logits}_{sj}^{(c)})_b$, respectively. This leads to the observed allele likelihood marginalizing over healthy- and cancer-derived origins

$$p(y_{sj} \mid x_{sj}, C_{si}, a_{sj}, z_s, \alpha_s) = (1 - \alpha_s)\, p_h(y_{sj}) + \alpha_s\, p_c(y_{sj}).$$

Using IPW, the weighted data term is the self-normalized (Hájek-style) estimator (Hájek, 1971):

$$\mathcal{L}_{\mathrm{data}} = \frac{\sum_{s=1}^{S} \sum_{j=1}^{m_s} w_{si}\, \mathrm{NLL}_{sj}}{\sum_{s=1}^{S} \sum_{j=1}^{m_s} w_{si}},$$

where $\mathrm{NLL}_{sj} = -\log[(1 - \alpha_s) p_h(y_{sj}) + \alpha_s p_c(y_{sj})]$.

2.5. Training and Inference

Sample embeddings are regularized by a Gaussian mixture prior with $K = 5$ components. To stabilize low-fraction estimates, we place a Beta prior on α_s with $\alpha_s \sim \mathrm{Beta}(1, 19)$.

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{data}} + \beta\, \mathcal{L}_{\mathrm{prior}} + \gamma\, \mathcal{L}_{\alpha}.$$

We optimize four parameter groups with different update frequencies: backbone and head parameters per batch, GMM parameters per batch, logit parameters η_s per batch, and embeddings z_s per epoch using DGD-style accumulation.

For a new sample, decoder and GMM parameters are frozen, while z_s and η_s are initialized and optimized using the same weighted objective. The model supports two inference modes: (1) *tumor-informed* mode using the sample-specific catalog to activate both marginal and catalog-specific shifts, and (2) *tumor-agnostic* mode setting all $C_{si} = 0$ to rely only on marginal shifts.

2.6. Synthetic Validation Setup

We used a controlled synthetic cohort matching the event-level model inputs as a proof of concept for optimization and frozen-decoder inference. The cohort contained 24 samples: 18 for training and six held out for inference. The retained dataset contained 57,510 read-base events, including 21,516 catalog events from 150 positions and 35,994 non-catalog events from 899 sampled positions. Reads were generated from healthy or tumor components according to the true α_s. Catalog tumor reads were alternate-enriched, while non-catalog tumor reads carried weaker marginal signal through C-to-T-like contexts and tumor-shifted fragment length, replication timing, and accessibility. Remaining technical covariates were sampled independently from simple parametric distributions. Non-catalog positions were subsampled and inverse-probability weighted.

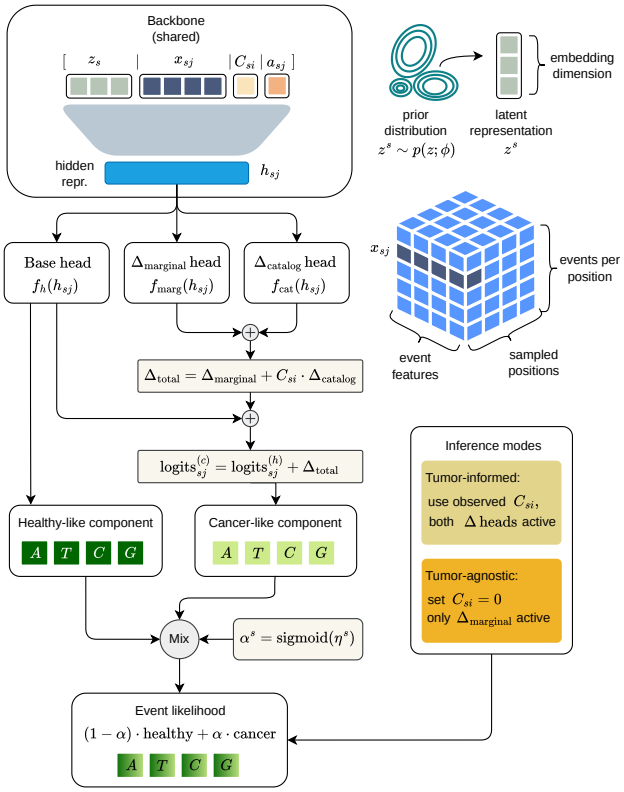

Figure 2. **Overview of the logit-delta ctDNA mixture model with dual-mode inference.** Event-level features, sample embedding, catalog indicator, and expected tumor alternate allele are mapped to healthy-like and cancer-like allele probabilities. The cancer-like component is parameterized as a logit shift from the healthy-like component, with a catalog-specific shift active only when $C_{si} = 1$. The two components are mixed by the sample-level ctDNA fraction $\alpha^s = \mathrm{sigmoid}(\eta^s)$. In tumor-informed mode, the observed catalog indicator is used; in tumor-agnostic mode, C_{si} is set to zero for all retained positions.

3. Results and Discussion

3.1. Synthetic Validation Results

We used the synthetic data to test whether the likelihood and inference procedure behave as expected in a controlled setting.

On the training split, the fitted sample-specific mixture logits followed the simulated ctDNA fractions with high correlation (Figure 3). This provides a controlled proof of concept that the weighted event-level mixture likelihood can capture sample-level tumor-fraction differences when its modeling assumptions are satisfied.

The same trained decoder supported both inference modes, but their behavior differed sharply on held-out samples (Figure 4). When catalog information was available, frozen-decoder inference preserved the ordering of the simulated ctDNA fractions across the validation samples. When catalog information was removed, the sparse marginal signal in this synthetic dataset was insufficient for reliable tumor-

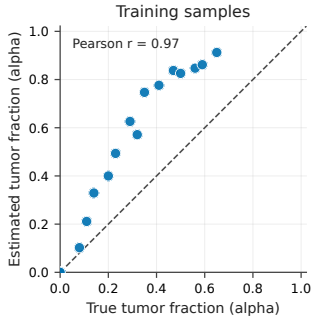

Figure 3. **Training-sample tumor-fraction recovery in synthetic data.** Estimated sample-level mixture fractions after training are plotted against the true simulated ctDNA fractions. The dashed line denotes identity. The learned sample-specific logits recover the rank ordering of the training samples with Pearson correlation $r = 0.97$.

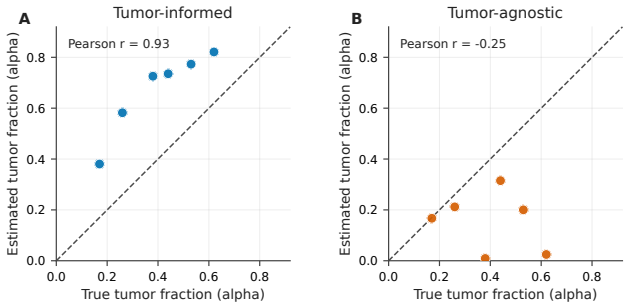

Figure 4. **Frozen-decoder inference in tumor-informed and tumor-agnostic modes on held-out synthetic samples.** For each held-out sample, decoder and prior parameters were fixed and only the sample embedding and mixture logit were optimized. In tumor-informed mode, the observed catalog indicator and expected alternate allele were used, activating both marginal and catalog-specific shifts. In tumor-agnostic mode, catalog indicators and alternate-allele encodings were set to zero, leaving only the marginal shift active. Tumor-informed inference followed the true simulated fractions ($r = 0.93$), whereas tumor-agnostic inference was weak in this small synthetic cohort ($r = -0.25$).

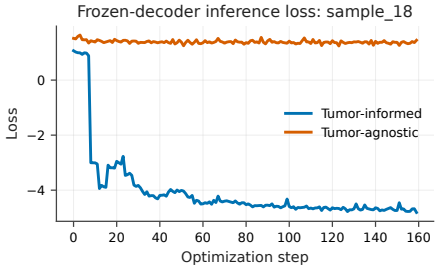

Figure 5. **Representative frozen-decoder inference loss.** Optimization traces are shown for one held-out synthetic sample. Tumor-informed inference sharply reduces the weighted objective and converges to a lower loss, whereas tumor-agnostic inference remains nearly flat at a higher loss. The difference reflects the additional catalog-specific evidence available in tumor-informed mode.

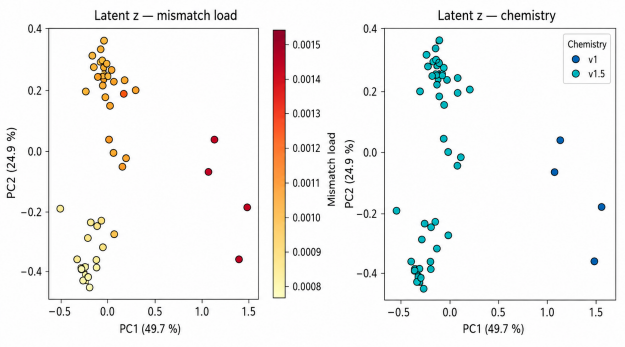

Figure 6. **Preliminary latent representation structure in real training data.** PCA of learned sample embeddings colored post hoc by mismatch load (left) and sequencing chemistry (right). Neither variable was included as a model input.

fraction recovery under this controlled setup. This is consistent with the design of the simulation: catalog alternate reads provide a strong patient-specific signal, while the tumor-agnostic component must rely on a much weaker genome-wide shift. For a representative held-out sample, tumor-informed frozen-decoder inference rapidly reduced the weighted objective, whereas the tumor-agnostic trace remained nearly flat at a higher loss (Figure 5).

As an implementation sanity check, catalog expected-alternate reads were assigned high posterior cancer responsibility on average (0.91), whereas catalog non-alternate reads had lower cancer responsibility (0.35). The catalog-specific delta was exactly gated off for non-catalog events, with $\max \|C_{si}\Delta_{sj}^{\text{cat}}\| = 0$ among events at $C_{si} = 0$ positions. The small synthetic experiment required two stabilizing choices: initializing η_s from a catalog alternate-fraction estimate and using an intermediate phase in which η_s was held fixed while the delta heads learned the correct component orientation. These choices prevent label switching in the small synthetic cohort and should be revisited for large-scale real-data training.

3.2. Preliminary Real-Data Latent Structure

As a preliminary real-data check, we trained the model on 46 training patients. PCA of the learned embeddings z_s showed structure associated with sequencing chemistry and mismatch load, which were used only for post hoc coloring (Figure 6). This suggests that the DGD-style embeddings can capture latent technical variation, although the result remains preliminary.

Together, these results show how event-level genomic observations can be coupled with learned sample representations for weak-signal aggregation and discovery of latent technical structure. Future work will evaluate ctDNA accuracy on held-out real samples and test whether the learned representations improve robustness across sequencing protocols and patient cohorts.

Impact Statement

This paper presents methodological work in machine learning for cancer genomics. If validated in larger real-data studies, improved ctDNA quantification methods could support cancer detection and treatment monitoring. The present results are preliminary and do not constitute a clinical assay. ctDNA analysis also raises privacy considerations for genomic data and should be conducted with appropriate consent and data protection measures.

References

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., Antonarakis, E. S., Azad, N. S., Bardelli, A., Brem, H., Cameron, J. L., Lee, C. C., Fecher, L. A., Gallia, G. L., Gibbs, P., Le, D., Giuntoli, R. L., Goggins, M., Hogarty, M. D., Holdhoff, M., Hong, S. M., Jiao, Y., Juhl, H. H., Kim, J. J., Siravegna, G., Laheru, D. A., Lauricella, C., Lim, M., Lipson, E. J., Marie, S. K. N., Netto, G. J., Oliner, K. S., Olivi, A., Olsson, L., Riggins, G. J., Sartore-Bianchi, A., Schmidt, K., Shih, I. M., Oba-Shinjo, S. M., Siena, S., Theodorescu, D., Tie, J., Harkins, T. T., Veronese, S., Wang, T. L., Weingart, J. D., Wolfgang, C. L., Wood, L. D., Xing, D., Hruban, R. H., Wu, J., Allen, P. J., Schmidt, C. M., Choti, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Papadopoulos, N., and Diaz, L. A. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.*, 6(224): 224ra24, 2014. doi: 10.1126/scitranslmed.3007094.

Christensen, E., Birkenkamp-Demtroder, K., Sethi, H., Shchegrova, S., Salari, R., Nordentoft, I., Wu, H. T., Knudsen, M., Lamy, P., Lindskrog, S. V., Taber, A., Balcioglu, M., Vang, S., Assaf, Z., Sharma, S., Tin, A. S., Srinivasan, R., Hafez, D., Reinert, T., Navarro, S., Olson, A., Ram, R., Dashner, S., Rabinowitz, M., Billings, P., Sigurjonsson, S., Andersen, C. L., Swenerton, R., Aleshin, A., Zimmermann, B., Agerbaek, M., Lin, C. H. J., Jensen, J. B., and Dyrskjot, L. Early detection of metastatic relapse and monitoring of therapeutic efficacy by ultra-deep sequencing of plasma cell-free DNA in patients with urothelial bladder carcinoma. *J. Clin. Oncol.*, 37(18):1547–1557, 2019. doi: 10.1200/JCO.18.02052.

Christensen, M. H., Drue, S. O., Rasmussen, M. H., et al. DREAMS: deep read-level error model for sequencing data applied to low-frequency variant calling and circulating tumor DNA detection. *Genome Biol.*, 24(1):99, 2023. doi: 10.1186/s13059-023-02920-1.

Frydendahl, A., Nors, J., Rasmussen, M. H., Henriksen, T. V., Nesic, M., Reinert, T., Afterman, D., Lauterman, T., Kuzman, M., Gonzalez, S. R., Glavas, D., Smadback, J., Maloney, D., Levatic, J., Yahalom, M., Ptashkin, R. N., Tavassoly, I., Donenhirsh, Z., White, E., Ravi, K., Alon, U., Nordentoft, I., Lindskrog, S. V., Dyrskjot, L., Jaensch, C., Love, U. S., Andersen, P. V., Thorlacius-Ussing, O., Iversen, L. H., Gotschalck, K. A., Zviran, A., Oklander, B., and Andersen, C. L. Detection of circulating tumor DNA by tumor-informed whole-genome sequencing enables prediction of recurrence in stage III colorectal cancer patients. *Eur. J. Cancer*, 211:114314, 2024. doi: 10.1016/j.ejca.2024.114314.

Hájek, J. Comment on 'An essay on the logical foundations of survey sampling, part one'. In Godambe, V. P. and Sprott, D. A. (eds.), *Foundations of Statistical Inference*, pp. 236. Holt, Rinehart and Winston, Toronto, 1971.

Nordentoft, I., Lindskrog, S. V., Birkenkamp-Demtroder, K., Gonzalez, S., Kuzman, M., Levatic, J., Glavas, D., Ptashkin, R., Smadbeck, J., Afterman, D., Lauterman, T., Cohen, Y., Donenhirsh, Z., Tavassoly, I., Alon, U., Frydendahl, A., Rasmussen, M. H., Andersen, C. L., Lamy, P., Knudsen, M., Polak, P., Zviran, A., Oklander, B., Agerbaek, M., Jensen, J. B., and Dyrskjot, L. Whole-genome mutational analysis for tumor-informed detection of circulating tumor DNA in patients with urothelial carcinoma. *Eur. Urol.*, 86(4):301–311, 2024. doi: 10.1016/j.eururo.2024.05.014.

Pantel, K. and Alix-Panabières, C. Liquid biopsy and minimal residual disease—latest advances and implications for cure. *Nature Reviews Clinical Oncology*, 16(7):409–424, 2019.

Pantel, K. and Alix-Panabières, C. Minimal residual disease as a target for liquid biopsy in patients with solid tumours. *Nature Reviews Clinical Oncology*, 22(1):65–77, 2025.

Phallen, J., Sausen, M., Adleff, V., Leal, A., Hruban, C., White, J., Anagnostou, V., Fiksel, J., Cristiano, S., Papp, E., Speir, S., Reinert, T., Orntoft, M. B. W., Woodward, B. D., Murphy, D., Parpart-Li, S., Riley, D., Nesselbush, M., Sengamalay, N., Georgiadis, A., Li, Q. K., Madsen, M. R., Mortensen, F. V., Huiskens, J., Punt, C., Van Grieken, N., Fijneman, R., Meijer, G., Husain, H., Scharpf, R. B., Diaz, L. A., Jones, S., Angiuoli, S., Orntoft, T., Nielsen, H. J., Andersen, C. L., and Velculescu, V. E. Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.*, 9(403): eaan2415, 2017. doi: 10.1126/scitranslmed.aan2415.

Reinert, T., Henriksen, T. V., Christensen, E., Sharma, S., Salari, R., Sethi, H., Knudsen, M., Nordentoft, I., Wu, H. T., Tin, A. S., Rasmussen, M. H., Vang, S., Shchegrova, S., Johansen, A. F. B., Srinivasan, R., Assaf, Z., Balcioglu, M., Olson, A., Dashner, S., Hafez, D., Navarro, S., Goel, S., Rabinowitz, M., Billings, P., Sigurjonsson, S., Dyrskjot, L., Swenerton, R., Aleshin, A., Laurberg,

S., Madsen, A. H., Kannerup, A. S., Stribolt, K., Krag, S. P., Iversen, L. H., Sunesen, K. G., Lin, C. H. J., Zimmermann, B. G., and Andersen, C. L. Analysis of plasma cell-free DNA by ultradeep sequencing in patients with stages I to III colorectal cancer. *JAMA Oncol.*, 5(8):1124–1131, 2019. doi: 10.1001/jamaoncol.2019.0528.

Schuster, V. and Krogh, A. The deep generative decoder: MAP estimation of representations improves modelling of single-cell RNA data. *Bioinformatics*, 39(9):btad497, 2023. doi: 10.1093/bioinformatics/btad497.

Stoler, N. and Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.*, 3(1):lqab019, 2021. doi: 10.1093/nargab/lqab019.

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., Pacey, S., Baird, R., and Rosenfeld, N. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer*, 17(4):223–238, 2017. doi: 10.1038/nrc.2017.7.

Wan, J. C. M., Heider, K., Gale, D., et al. ctDNA monitoring using patient-specific sequencing and integration of variant reads. *Sci. Transl. Med.*, 12(548):eaaz8084, 2020. doi: 10.1126/scitranslmed.aaz8084.

Widman, A. J., Shah, M., Frydendahl, A., et al. Ultrasensitive plasma-based monitoring of tumor burden using machine-learning-guided signal enrichment. *Nat. Med.*, 30(6):1655–1666, 2024. doi: 10.1038/s41591-024-03040-4.

Zhu, G., Rahman, C. R., Getty, V., et al. A deep-learning model for quantifying circulating tumour DNA from the density distribution of DNA-fragment lengths. *Nat. Biomed. Eng.*, 9(3):307–319, 2025. doi: 10.1038/s41551-025-01370-3.

Zviran, A., Schulman, R. C., Shah, M., et al. Genome-wide cell-free DNA mutational integration enables ultrasensitive cancer monitoring. *Nat. Med.*, 26(7):1114–1124, 2020. doi: 10.1038/s41591-020-0915-3.

A. Data Processing and Feature Construction

Aligned paired-end cfDNA reads were converted into fragment-level and base-level model inputs separately for each sample and chromosome. Candidate fragments were defined from properly paired, primary, non-duplicate reads mapping to the same chromosome with opposite orientation. For each pair, only the read-overlap interval was used, and a base was retained only when both reads supported the same nucleotide with sufficient quality. This overlap consensus reduces sequencing-error contributions before downstream feature extraction.

For every retained fragment, we extracted a fixed-width representation centered on the fragment and augmented it with 5 bp of flanking reference sequence. The resulting arrays contain aligned reference, read-one, and read-two tracks, base qualities, and gap indicators. Fragment length was computed from the outer alignment coordinates, and fragments of length 300 bp or longer were discarded.

Reads and bases were filtered using mapping quality ≥ 20, base quality ≥ 30, proper-pair status, primary alignment, and the absence of duplicate or QC-fail flags. Positions were additionally removed if they overlapped genomic blacklist regions, donor-specific blacklists, patient-specific low-power or problematic regions, or patient-specific germline SNV/indel blacklists. The remaining high-confidence overlap bases define the set of available modeling opportunities.

Each valid base was represented by sequence, fragment, and genomic context. Sequence-array inputs encode the reference and paired-read alignments with quality and gap tracks. Tabular features summarize trinucleotide context, GC content, sequence complexity, relative position in the fragment, distance to fragment ends, nearby indels or clipping, and the number of other mismatches on the fragment. When available, we also included replication timing, mappability, repeat and microsatellite annotations, and ATAC/DNase accessibility features. Accessibility annotations were summarized in 10 bp bins using both binary overlap indicators and distances to the nearest open region. Missing annotation values were encoded explicitly, and distance features were transformed as $\log(1 + x)$.

The preprocessing is base-resolution within fragment overlaps, so each sample contains millions of fragments and typically tens to hundreds of millions of candidate bases after filtering. Intermediate chromosome-level outputs were written as compressed Parquet chunks and then merged per sample. For model training, each sample was stored as NumPy array bundles containing numerical features, categorical features, and precomputed sequence-context tensors. Sample-level sequencing QC summaries, including duplication, insert-size, and coverage metrics, were joined from external metadata tables. Numerical normalization statistics were estimated on the training split only, and categorical variables were mapped to integer vocabularies with an explicit unknown category.

B. Implementation Details

The pipeline was implemented in Python 3.12 using PyTorch 2.8.0 and NumPy 2.3.2. Columnar data handling used Polars 1.35.2 and PyArrow 16.1.0, and BAM processing used pysam 0.23.3. Model training was run on CUDA-enabled Linux systems when available, with automatic mixed precision in `float16` and gradient scaling; otherwise, computation used `float32`. NumPy and PyTorch random seeds were fixed to 0 unless otherwise noted. Workflow execution was managed with GWF 1.7.2[1] and scheduled on a SLURM high-performance computing cluster. Project dependencies were managed with Poetry.

Data splits were performed at the patient level to avoid leakage between training and evaluation. One representative row per patient was used to define stratification labels, after which all samples from that patient were assigned to the same split. The main split used an 80/20 train/test ratio with multilabel stratification over age, sex, sampling-time category, and tumor fraction. Age and tumor fraction were discretized into 20 bins before stratification. All normalization and target-standardization statistics were computed from the training samples only and then applied unchanged to validation and test samples.

[1] https://anaconda.org/gwforg/gwf