www.nature.com/natmachintell / December 2025 Vol.7 No.12

nature machine intelligence

Cell-free RNA profiling with a language model

nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01148-x

A multimodal cell-free RNA language model for liquid biopsy applications

Received: 16 April 2025

Accepted: 30 October 2025

Published online: 10 December 2025

Check for updates

Mehran Karimzadeh[1,5], Aiden M. Sababi[1,5], Amir Momen-Roknabadi[1], Nae-Chyun Chen[1], Taylor B. Cavazos[1], Sukh Sekhon[1], Jieyang Wang[1], Rose Hanna[1], Alice Huang[1], Dang Nguyen[1], Selina Chen[1], Ti Lam[1], Kimberly H. Chau[1], Anna Hartwig[1], Lisa Fish[1], Helen Li[1], Babak Behsaz[1], Fereydoun Hormozdiari[1,2] ✉, Babak Alipanahi[1] ✉ & Hani Goodarzi[3,4] ✉

Cell-free RNA (cfRNA) profiling has emerged as a powerful tool for non-invasive disease detection, but its application is limited by data sparsity and complexity, especially in settings with constrained sample availability. We introduce Exai-1, a multimodal, transformer-based generative foundation model that integrates RNA sequence embeddings with cfRNA abundance data to capture biologically meaningful representations of circulating RNAs. By leveraging both sequence and expression modalities, Exai-1 captures a biologically meaningful latent structure of cfRNA profiles. Pretrained on over 306 billion tokens from 8,339 samples, Exai-1 enhances signal fidelity, reduces technical noise and improves disease detection by generating synthetic cfRNA profiles. We show that self-attention and variational inference are particularly important for the preservation of biological signals and contextual relationships. Additionally, Exai-1 facilitates cross-biofluid translation and assay compatibility through disentangling biological signals from confounders. By uniting sequence-informed embeddings with cfRNA expression patterns, Exai-1 establishes a transfer learning foundation for liquid biopsy, offering a scalable and adaptable framework for next-generation cfRNA-based diagnostics.

Blood surveillance technologies have advanced substantially, particularly in cancer detection, shifting from traditional tissue-based diagnostics to less invasive liquid biopsies. For example, in cancer, liquid biopsies analyse tumour biomarkers in biofluids to detect disease, leveraging circulating tumour DNA, circulating tumour cells, cell-free RNAs (cfRNAs) and other circulating biomolecules[1–6]. These approaches, in addition to applications for early detection, have also become powerful tools for minimal residual disease detection and treatment monitoring[7].

We previously identified a class of small RNAs (smRNAs) that are actively expressed in cancer cells but largely absent in healthy cells[8]. A substantial fraction of these cancer-emergent RNAs, which we named orphan non-coding RNAs (oncRNAs), are actively secreted by tumours into circulation at levels higher than circulating tumour DNA, which is mostly shed through cell death, making them strong indicators of cancer cell states. We have shown that their detection in blood opens new opportunities for early cancer diagnosis, minimal residual disease monitoring and cancer subtype stratification from small blood volumes[8–10].

A major challenge is detecting individual oncRNAs in small blood volumes due to the sparsity of tumour-released molecules, particularly in early stage disease. However, circulating oncRNAs do not appear in isolation but rather in distinct patterns and profiles—akin to a constellation of stars in a starry sky. This is precisely the type of problem in which artificial intelligence excels. To harness the diagnostic potential of oncRNAs, we developed ORION, a deep generative model that

[1]Exai Bio Inc., Palo Alto, CA, US. [2]University of California, Davis, Davis, CA, US. [3]University of California, San Francisco, San Francisco, CA, US. [4]Arc Institute, Palo Alto, CA, US. [5]These authors contributed equally: Mehran Karimzadeh, Aiden M. Sababi. ✉e-mail: fhormozd@ucdavis.edu; babaka@exai.bio; hani@arcinstitute.org

effectively maps cfRNA fingerprints in blood to disease presence and outcomes[10]. Training such models requires large datasets—often hundreds or thousands of samples—typically feasible only for common cancers. Detecting rare cancers, therefore, necessitates foundational models capable of generalization and few-shot learning, effectively recognizing disease patterns from a limited number of instances.

In recent years, large masked language models have emerged as capable foundation models, serving as powerful tools for zero-shot and few-shot learning. Models such as GPT-4 (ref. 11) leverage masked language modelling on a large number of training tokens to learn complex, context-dependent feature representations across diverse domains. These developments are transforming genomics[12], where transformer-based architectures now play a key role in uncovering regulatory patterns in DNA and RNA sequences.

Several large genomic foundation models have emerged, each tailored to specific aspects of molecular biology. DNABERT[13], Borzoi[14] and GET[15] focus on linking genetic sequences to regulatory functions. GROVER[16], a DNA foundation model, uses byte-pair tokenization to better capture the genomic sequence structure. Evo-2, a long-context transformer pretrained on all genomes, demonstrates strong performance in tasks such as gene essentiality prediction and synthetic sequence generation[17]. BigRNA[18], specifically trained on RNA-sequencing datasets, excels at predicting tissue-specific RNA expression.

Single-cell foundation models have also made prominent strides in transcriptomics. Models such as scBERT[19], Exceiver[20], GeneCompass[21], scFoundation[22], Geneformer[23], scGPT[24] and UCE[25] have demonstrated efficacy in tasks such as cell-type classification and gene expression imputation. Although these models have shown promise, their full potential in genomics is still being realized[26–28].

Despite these recent advances in transformer architectures, their application to liquid biopsy data remains largely unexplored, primarily due to the limited availability of large-scale datasets. One major challenge in cfRNA analysis is the imbalance between the high dimensionality of circulating RNA features and the small number of available clinical samples. cfRNA encompasses the entirety of circulating RNA species, the majority of which are <200 nt and are released in extracellular vesicles and lipoprotein complexes[29]. Traditional predictive models typically focus on specific supervised tasks with a restricted feature space, often failing to capture the complex interactions between RNA species originating from different cell types. This limitation stems from the scarcity of labelled data, making it difficult to develop models that generalize across diverse biological contexts.

Although cfRNA analysis presents unique challenges, there are inherent biological structures that can be leveraged to overcome them. Importantly, cfRNAs are not independent entities, but rather share information between their sequence and expression[9,30], for example, those that are generated from the same regulatory pathways. Their biotypes and secretion patterns are dictated by their sequence and structure, whereas their expression levels are shaped by cell-type specificity and regulatory mechanisms. Because of this interplay, circulating RNAs do not appear in isolation but rather in distinct patterns. As a result, unlike conventional biomarkers, which are often interpreted deterministically, cfRNA detection must be approached probabilistically, taking into account the broader context of other RNAs within the same sample. When viewed through this multimodal lens, the issue of data sparsity becomes more tractable.

More importantly, we have also developed an automated cfRNA assay capable of processing hundreds of liquid biopsy samples per batch, ensuring consistency and scalability. This expansive dataset provides a unique opportunity to train transformer-based masked language models, which excel at learning contextual dependencies through self-supervised learning. Here we introduce Exai-1, a multimodal transformer-based generative model designed specifically for cell-free smRNA data, trained and evaluated on the largest corpus of cfRNA profiles—over 13,000 plasma and serum samples. By integrating sequence- and structure-informed embeddings from RNA foundation model (RNA-FM) alongside cfRNA abundance data, Exai-1 unites two complementary modalities into a unified latent space, enabling it to extract biologically meaningful relationships between cfRNA sequences and their expression patterns. This multimodal framework allows Exai-1 to generalize across diverse downstream tasks, improving signal fidelity, mitigating noise and enhancing the ability to distinguish disease-associated cfRNA profiles. By incorporating sequence-driven knowledge with cfRNA abundance, Exai-1 provides a scalable and unified framework for understanding cfRNA biology and advancing liquid biopsy applications.

Results

Exai-1: training a multimodal foundation model on large-scale cfRNA data

To train a foundational model that is unbiased towards specific patient cohorts, it is essential to select features representing broad blood biology rather than any particular pathology. We, therefore, curated a comprehensive feature set from our in-house annotated collection of cfRNAs, comprising 9,491,734 features. Our selection focused specifically on tRNA, snoRNA, miRNA, yRNA, lncRNA and oncRNA biotypes. First, we retained only those cfRNAs expressed in at least 1% of the samples and subsequently applied a coefficient-of-variation cut-off to preserve highly variable features (Extended Data Fig. 1a–e). Given the large initial number of oncRNAs (394,175) after this selection, we used embeddings from a recently published RNA-FM[31] to both cluster these oncRNAs and provide sequence-informed initialization for Exai-1. Specifically, we extracted the top 32 principal components (PCs) from the RNA-FM embedding space—collectively explaining over 91% of the variation, and used these PCs to initialize Exai-1's feature embeddings for cfRNA features. We then selected 4,558 representative oncRNA features evenly distributed across this low-dimensional space (Methods and Extended Data Fig. 1e,f). The final feature set consisted of 704 tRNAs, 610 snoRNAs, 761 yRNAs, 716 miRNAs and 4,558 oncRNAs, each forming distinct clusters in the RNA-FM-based embedding space (Fig. 1a).

Next, we assembled a large cohort consisting of 13,014 blood samples (6,792 serum and 6,222 plasma) for training and evaluating Exai-1 (Fig. 1b). During Exai-1's pretraining, we selected 7,349 cfRNA features and applied masked language modelling, by randomly masking 25% of the features among all the samples of the mini-batch to encourage contextual learning (Fig. 1c). To incorporate sequence-derived information, we initialized the embedding layer parameters using PCs of cfRNA sequences from RNA-FM[32] (Methods). Each 32-dimensional-sequence-based embedding was then scaled by the corresponding cfRNA abundance in each sample, ensuring that both sequence context and expression levels were captured in a single, unified representation. These embeddings were fed into Exai-1's transformer encoder, which projects them onto a lower-dimensional latent space, where self-attention mechanisms model dependencies across cfRNA species. A variational layer then learns a latent Gaussian distribution, from which the transformer decoder samples to reconstruct the original cfRNA abundances.

Another critical design aspect that enables Exai-1 to learn sample-specific representations is the integration of special tokens for multitask learning. These include tokens for 'cancer detection' <CD>, 'tissue of origin' <TOO>, 'assay version' <AV> and sample 'biofluid type' <BT> (Fig. 1c). These special tokens are essential for enhancing the model's ability to incorporate additional contextual information alongside cfRNA abundances, enabling the model to generate biologically and clinically relevant embeddings. Token permutation or retraining without tokens consistently reduced the performance of downstream tasks, further indicating that these tokens encode meaningful biological and contextual information (Extended Data Fig. 2). Consequently, the inclusion of special tokens substantially broadens the applicability

https://doi.org/10.1038/s42256-025-01148-x

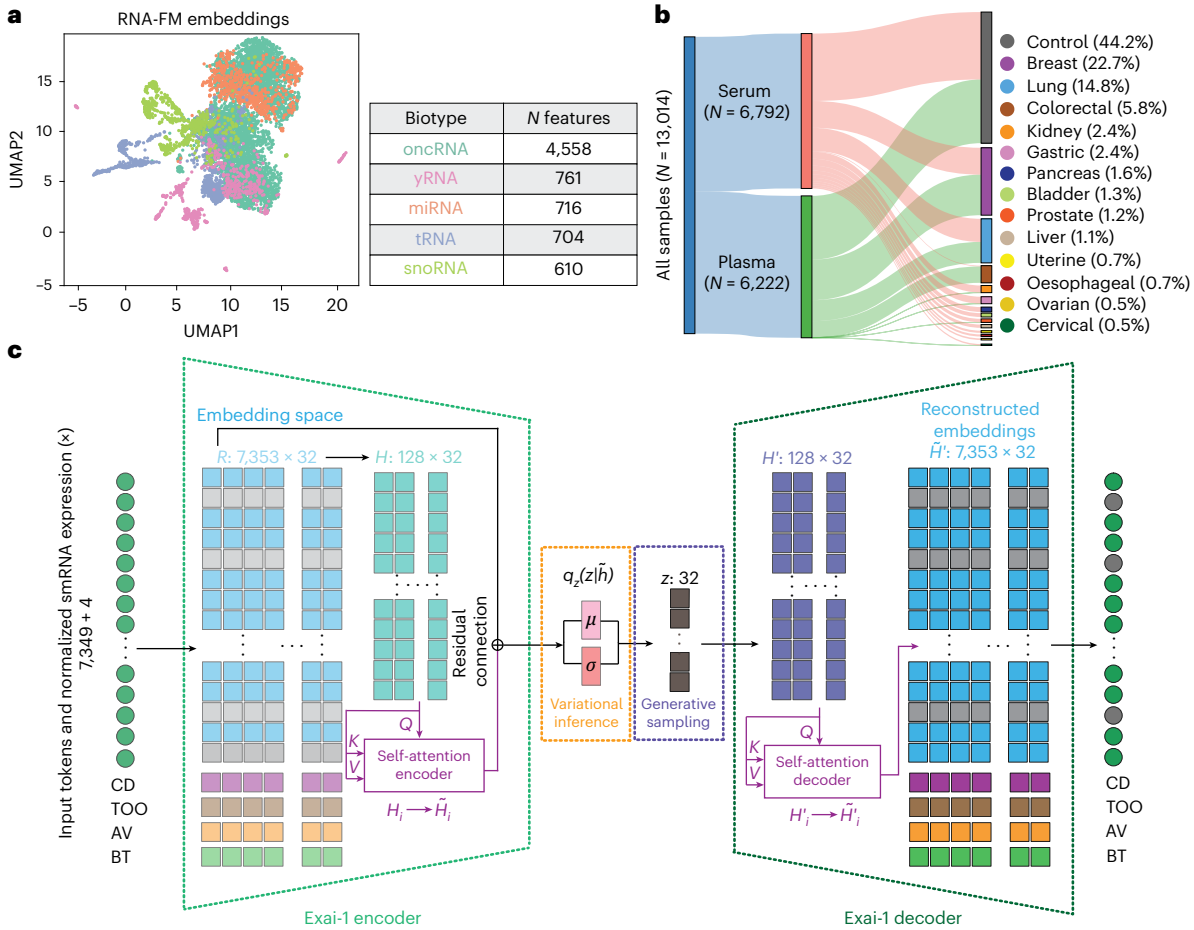


Fig. 1 | Feature selection, data composition and Exai-1 architecture. a, UMAP projection of the 7,349 selected cfRNA features, colour coded by biotype (miRNA, tRNA, snoRNA, oncRNA and yRNA), showing distinct clustering patterns. **b**, Sankey diagram illustrating the dataset composition, with 13,014 total blood samples stratified by the biofluid type (serum, $N = 6{,}792$; plasma, $N = 6{,}222$) and diagnostic categories (control, benign and cancer subtypes). **c**, Schematic of Exai-1's transformer-based architecture. cfRNA features, along with special tokens for contextual learning (CD, cancer detection; TOO, tissue of origin; AV, assay version; BT, biofluid type), are embedded into a high-dimensional space. A variational encoder applies self-attention to generate a latent representation, which is then used by the decoder through generative sampling to reconstruct masked cfRNA features. Task-specific multilayer perceptrons further refine embeddings for downstream applications.

and utility of Exai-1. Moreover, by applying self-attention to a compressed hidden space rather than the original high-dimensional input features, Exai-1 maintains a compact architecture with only 3.6 million trainable parameters, thereby minimizing the risk of overfitting.

Exai-1 enables high-fidelity cfRNA reconstruction and noise removal

As a masked language model, Exai-1 is trained primarily using a self-supervised objective to learn cfRNA abundances and their relationships. To evaluate its reconstruction performance, we conducted a test by randomly masking 25% of cfRNA features in held-out samples and comparing Exai-1's reconstructed values to the original measurements. In genomic datasets, the dataset average often serves as a challenging baseline that many models struggle to surpass[26,33]. Although this baseline yielded an R^2 value of 0.57 (95% confidence interval (CI), 0.55–0.59), Exai-1 significantly outperformed it, achieving an R^2 value of 0.89 (95% CI, 0.88–0.89; Fig. 2a,b). Even after subtracting each expected and observed feature value by its mean, or after z scoring, this observation remained valid (Extended Data Fig. 3). We also assessed whether reconstructed profiles preserve the biologically meaningful co-expression structure by computing the adjusted Rand index between feature clusters derived from the original and reconstructed matrices. Using different numbers of clusters from 2 to 64, we observed adjusted Rand index values above 0.78, indicating a strong concordance in clustering assignments, demonstrating that Exai-1 reconstructions retain the grouping of biologically related cfRNAs. This result underscores Exai-1's ability to accurately capture sample-specific cfRNA abundance profiles in blood and its capability to effectively fill in the gaps.

We performed ablation analyses to better understand the contribution of each component of Exai-1. Lack of self-attention, for example, significantly dropped the performance of classification tokens. On reconstruction tasks, we particularly observed a substantial drop when more than 50% of the held-out dataset tokens were masked. Similarly, a model without the variational component (vanilla transformer instead of a variational transformer), achieved the lowest performance in both reconstruction and token-specific tasks (Extended Data Fig. 4).

To assess whether the variational component learns a useful latent space (as opposed to collapsing), we monitored the Kullback–Leibler (KL) divergence term throughout training and summarized the learned latent variances across samples (Extended Data Fig. 5). The KL divergence shows an initial decrease followed by stabilization after ~500 epochs, with a modest upward drift later in training as the joint objective balances reconstruction, contrastive and token losses. Consistently, the distribution of latent variances concentrates near 1.0 for the majority of samples (>80%), indicating well-regularized, information-carrying latents. Additionally, we were able to train

https://doi.org/10.1038/s42256-025-01148-x

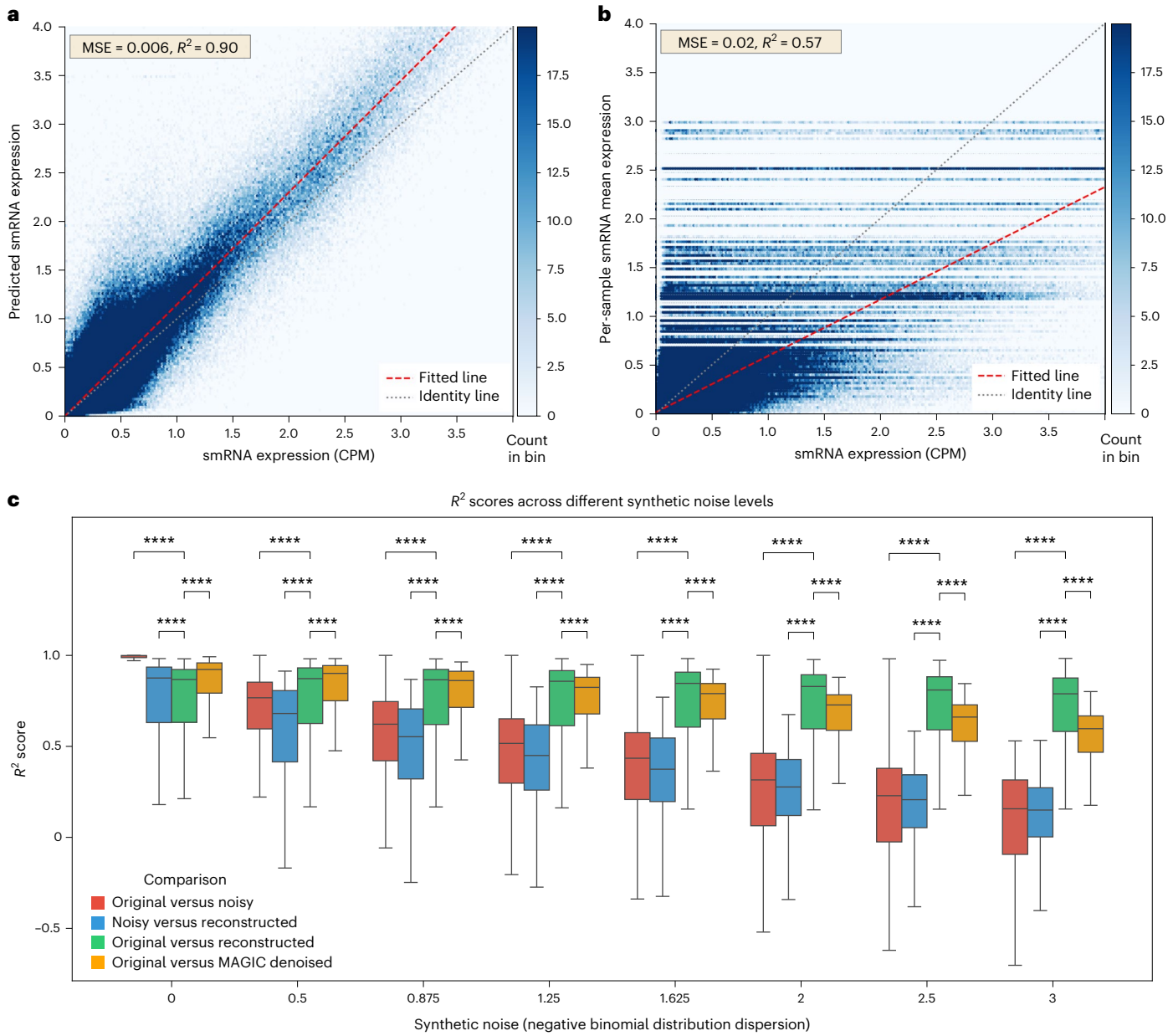


Fig. 2 | Exai-1's performance in reconstructing and denoising cfRNA abundance data. a, Density scatter plot showing Exai-1's reconstruction of masked cfRNA features in the test set. The *x* axis indicates the per-sample smRNA expression levels, whereas the *y* axis represents the reconstructed expression levels. A red linear regression line illustrates the correlation. MSE, mean squared error. **b**, Density scatter plot comparing the original cfRNA abundances to their mean values across samples (baseline). **c**, Box plots show R^2 between the different pairs of datasets, such as original versus noisy (red), noisy versus reconstructed (blue), original versus reconstructed (green) and original versus MAGIC-denoised datasets. The horizontal axis indicates eight simulated 'noisy' datasets, each generated with increasing dispersion from a negative binomial distribution. Box-plot data points: 2,076 samples within the validation set. Horizontal line of the box plot, median. Box range, interquartile range (IQR). Whisker, most extreme value within the quartile ± 1.5IQR. Paired *t*-tests were conducted to compare the model performance across experimental conditions. Statistical significance is indicated using the following thresholds: $P \leq 0.0001$ (****), $0.0001 < P \leq 0.001$ (***), $0.001 < P \leq 0.01$ (**), $0.01 < P \leq 0.05$ (*) and $P > 0.05$ (ns, not significant).

generalizable classifiers to predict the tissue of origin using this latent space (Extended Data Fig. 6). Together, these results support that the variational autoencoder component contributes meaningfully to representation learning.

We hypothesized that the reconstruction capabilities of Exai-1 should also allow us to increase the signal-to-noise ratio of liquid biopsy datasets. To assess our hypothesis, we introduced controlled noise into the original cfRNA expression profiles. Specifically, we modelled the observed cfRNA abundances using a negative binomial distribution and systematically increased the dispersion parameter to generate eight datasets representing varying levels of noise in data. For each noise setting, we compared how closely both noisy inputs and Exai-1's reconstructions matched the original, noise-free abundance profiles. For comparison, we also used Markov-affinity-based graph imputation of cells (MAGIC)[34]. Exai-1 consistently produced reconstructions that were substantially closer to the original data than the noisy inputs themselves (Fig. 2c). Although MAGIC performed better than Exai-1 for smaller levels of noise, Exai-1 performance remained consistent as we increased the level of noise, but as expected from an affinity graph denoising method, MAGIC's performance dropped. Furthermore, although the R^2 value between noisy and original data declined monotonically with increasing noise, Exai-1's reconstructions exhibited only

https://doi.org/10.1038/s42256-025-01148-x

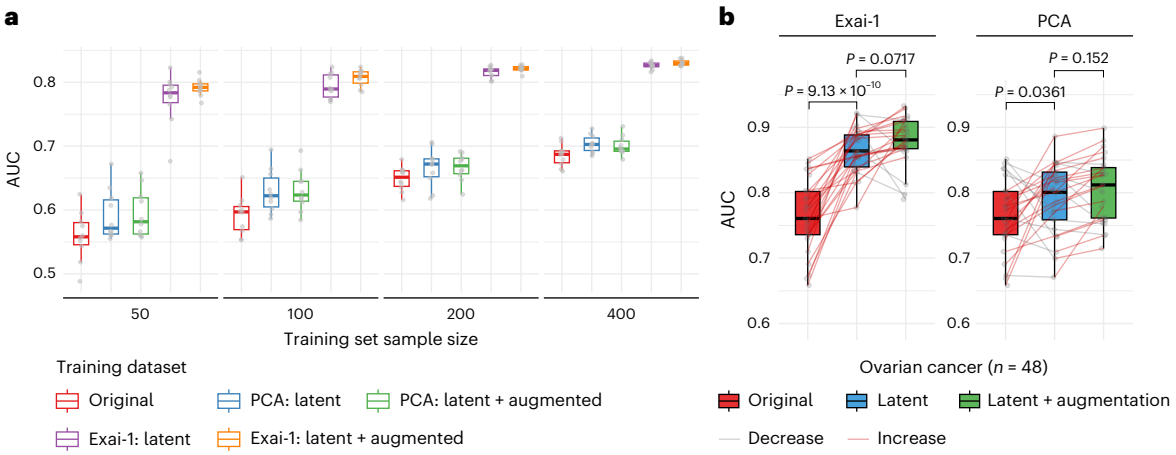


Fig. 3 | Improved cancer detection using Exai-1 latent space and generative sampling. a, AUC comparison of cancer detection classifiers trained on cfRNA samples in the original feature space (red), latent space (PCA, blue; Exai-1, purple) and latent space with augmentation (PCA, green; Exai-1, orange). Box plots summarize the performance across training sizes (50–400 samples) with ten random seeds per size. Each point corresponds to one experiment with a distinct random seed. **b**, Same data as **a**, but using 25 seeds instead of 10, where XGBoost models were trained on a dataset of ovarian cancer samples ($N = 48$). *P* values are from a one-sided *t*-test. Original versus Exai-1 latent comparison yielded a $t = -7.72$, 95% CI $[-\infty, -0.079]$, $P < 0.001$. Exai-1 latent versus Exai-1 latent + augmentation comparison yielded $t = -1.49$, 95% CI $[-\infty, 0.002]$, $P = 0.072$. Original versus PCA latent comparison yielded a $t = -1.84$, 95% CI $[-\infty, -0.002]$, $P = 0.036$. PCA latent versus PCA latent + augmentation comparison yielded a $t = -1.04$, 95% CI $[-\infty, 0.009]$, $P = 0.152$. Horizontal line of box plot, median; box range, IQR; whisker, most extreme value within the quartile ± 1.5IQR; individual points, outliers beyond a whisker.

a minimal decrease in R^2. These findings demonstrate Exai-1's ability not only to accurately reconstruct the cfRNA profiles but also effectively denoise them, preserving essential biological signals even in the presence of substantial noise.

Building on Exai-1's denoising capability, we hypothesized that the reconstructed features could be used to generate synthetic cfRNA profiles, thereby augmenting training datasets and enhancing cancer detection performance using standard classifiers. We had previously shown that such synthetic profiles can improve model training and generalization[10]. To evaluate this, we subsampled our training set to smaller datasets ranging from 50 to 400 samples (doubling the sample size at each step), each repeated ten times (Extended Data Fig. 7). For each subset, we trained an XGBoost classifier to distinguish cancer from the control samples, comparing two approaches: training exclusively on samples with original cfRNA profiles (original-only) versus training on datasets augmented with synthetic cfRNA profiles generated by Exai-1's reconstruction (original+reconstructed), effectively doubling the training sample size. For comparison, we also used principal component analysis (PCA)-based data augmentation (Methods) as a baseline for other types of data augmentation. We conducted analogous experiments in the latent domain, using either the deterministic latent representations from Exai-1 or PCA. For augmentation, we used generative sampling with 10% masked tokens in Exai-1 and Gaussian noise injection in PCA. We found that augmentation with synthetic cfRNA profiles improved the model performance, both when using denoised profiles in the original feature space and, more markedly, when using Exai-1's latent space augmented through generative sampling (Fig. 3a).

Our results demonstrate that augmentation in both feature space and latent space consistently improved the cancer detection performance. Although gains in the original feature space were most evident at smaller training sizes (up to 400 samples), the substantial improvements achieved in the Exai-1 latent space persisted even at larger sample sizes. We found that these improvements were substantially greater with Exai-1 than with PCA, both in the original feature space and in the latent representations (Fig. 3a and Extended Data Fig. 7). For completeness, we also evaluated 'reconstructed-only' training arms in which classifiers were fit exclusively to Exai-1 or PCA reconstructions. Although these models achieved strong performance when evaluated on reconstructed test data, they underperformed on the original test data, reflecting distributional shifts introduced by generative sampling in Exai-1 or Gaussian noise injection in PCA.

Furthermore, a linear mixed-effects model estimated an average area under the curve (AUC) increase of 0.03 (95% CI, 0.01–0.05, ($P = 0.002$)) when Exai-1 synthetic cfRNA profiles were included compared with using original samples alone. Within the latent space, this gain was significantly higher (0.22, 95% CI, 0.2–0.24, $P < 10^{-22}$). Augmentations through generative sampling in the latent space provided further improvements (0.02, 95% CI, 0.0–0.03, $P = 0.015$).

Such foundation-model-driven augmentation is particularly beneficial in clinical contexts in which sample availability is often limited. To assess the practical utility of synthetic cfRNA augmentation, we pretrained Exai-1 on a corpus that excluded ovarian cancer. For downstream evaluation, we trained XGBoost classifiers on 48 ovarian cancer samples and matched controls, using either raw cfRNA features or the Exai-1 latent space, with and without synthetic reconstructions for augmentation. Across 25 random seeds, using the Exai-1 latent space increased the AUC by 0.1 (95% CI, 0.08–0.12) relative to raw cfRNA; augmenting the training set with synthetic profiles provided a further 0.01 improvement (95% CI, 0–0.02; Fig. 3b).

We note that occasional drops in AUC for the augmented arm (Extended Data Fig. 7) occur mainly in two situations: (1) when reconstructions are generated from very small or biased training sets, which can exaggerate random patterns; (2) when the baseline sample size is already large, the added variability from reconstruction reduces the signal-to-noise ratio. Although such effects were sometimes seen in the original feature space, models trained in the latent space consistently outperformed those trained in the original space. Together, these findings demonstrate the utility of Exai-1's data denoising and synthetic cfRNA generation for enhancing cancer detection, especially in rare diseases for which samples are limited.

Exai-1's smRNA and sample embeddings capture key biological and technical factors

Transcription, processing and secretion of smRNAs are controlled by a suite of regulatory programs and pathways[35–37]. Therefore, similar to their common long RNA counterparts, smRNAs show patterns of co-expression across samples. As Exai-1 learns an embedding for each cfRNA species, we anticipate that various dimensions of this embedding

will capture such co-expression blocks. To explore this possibility, we first clustered cfRNAs based on their abundance profile within our dataset. For each cluster, we then generated a representative embedding by averaging the embeddings of individual cfRNAs across each dimension. We partitioned the features into ten abundance-based clusters of smRNAs and then assessed whether each cfRNA embedding dimension was significantly associated with any of these clusters by randomly shuffling the cluster memberships and recalculating the cluster embeddings. For each dimension, we compared these randomized values against the real values to construct a null distribution. This analysis revealed that smRNA co-expression clusters were significantly associated with distinct cfRNA embedding dimensions (FDR < 1%; Fig. 4a). Each embedding dimension was enriched in an average of 2.4 (±1.3 standard deviation) co-expression clusters, and conversely, each cluster was associated with an average of 8 (±5.2 standard deviation) embedding dimensions. Overall, these findings indicate that Exai-1 successfully learns sparse representations that are associated with the underlying biological characteristics of co-expressed smRNAs.

To further enhance the biological interpretability and address the contribution of individual cfRNAs to prediction tasks, we computed gradient-based Shapley additive values[38] for the cancer classification token in the held-out test set (Extended Data Fig. 8). Across all biotypes, 73% of oncRNAs exhibited positive contributions towards cancer predictions, compared with 60.4%–69.7% for other smRNA classes. Mapping the top-ranked cfRNAs to nearby genes revealed multiple cancer-associated loci, including *MYT1L*[39], *CTBP1* (ref. 40) and miR-186 (ref. 41). Gene ontology enrichment analysis of these features highlighted processes such as cell-cycle regulation, apoptosis and cell motility—pathways consistent with cancer aetiology. These results indicate that Exai-1 not only delivers strong predictive performance but also prioritizes biologically meaningful cfRNAs, underscoring its potential for early detection and minimal residual disease monitoring.

Next, to visualize the 32-dimensional sample embeddings learned by Exai-1, we used uniform manifold approximation and projection (UMAP) to project the embedding representations onto the held-out test set. Although UMAP is designed for summary visualization and is not suitable for making inferences about the underlying data, our UMAP plots (Fig. 4b–d) still reflect the observed patterns in the embedding space and highlight distinct separations between cancer and control samples, as well as differentiation based on technical variables such as assay version and biofluid type used to collect the blood samples. These visualizations indicate that Exai-1 embeddings effectively capture both biological differences and technical variations among samples.

We next assessed the impact of batch effects and the conservation of biological variance (presence of cancer) within Exai-1 embeddings, comparing them to embeddings obtained via PCA, PCA-based batch correction method Harmony[42], Scanorama[43], scVI[44] and scANVI[45] (Fig. 4e). In all of these scenarios, we used the held-out test set of the dataset in which the Exai-1 model was not trained on. Methods other than PCA used the same batch label (assay version) for correction in this dataset.

Following a strategy similar to that in ref. 46, we used batch removal based on the (average) silhouette and principal component regression comparison[47] to evaluate batch effect removal based on the assay version.

To assess the conservation of biological variance, we used label conservation metrics, including cell-type local inverse Simpson index[42], cancer-type silhouette width and 'isolated label scores'[46].

During training, Exai-1 leverages triplet margin loss to minimize the distance of the samples from the Gaussian latent distribution, which share a common tissue of origin and are distant (Methods). Due to this label-agnostic batch correction, Exai-1's embeddings significantly outperformed both PCA and Harmony across all metrics related to batch effect removal and the conservation of biological variance. This result can be attributed to the biologically informed contrastive loss in shaping the Exai-1's sample and feature embeddings, leading to the learning of a biologically rich embedding space. Collectively, this result demonstrates that Exai-1 effectively captures and reflects the biological patterns in the learned embedding spaces, facilitating numerous downstream applications, especially in disease detection and diagnostics, even in limited-sample scenarios leveraging few-shot learning.

Leveraging pretrained Exai-1 to enhance generalizability across sample source

Although serum is more widely available in many clinical settings, plasma samples typically provide higher-quality cfRNA data with lower background noise and are, therefore, a more favourable biofluid for the development of liquid biopsy assays. This disparity in both availability and quality poses a considerable challenge in developing robust liquid biopsy classifiers, as classifiers trained solely on one biofluid type often fail to generalize to another. We, therefore, investigated whether the foundational capabilities of Exai-1 could allow the training of more generalizable classifiers for tasks such as cancer detection.

To evaluate this hypothesis, we used paired plasma and serum samples from the same patients, which were not used during Exai-1 training. First, we used the original 7,349 smRNA log1p ($\log1p(x) = \log(x+1)$) of the counts per million (CPM) values to train a cancer detection XGBoost model using 391 plasma samples from individuals with cancer and 485 from individuals without cancer. The cross-validated performance of this model on plasma samples achieved an area under the receiver operating characteristic curve (AUROC) of 0.74 (95% CI, 0.7–0.77), whereas on serum samples, this model only had an AUROC of 0.56 (95% CI, 0.51–0.61; Fig. 5a). In a similar setup, however, when we used the latent space of Exai-1 for training a plasma-specific XGBoost model, the model had a higher and comparable performance on both plasma and serum (Fig. 5b). Other methods such as *k*-NN, ElasticNet, SVM Classifier, latent representations from other models such as Orion[10], or an ablated Exai-1 variant with zero-weight classification tokens did not match the performance achieved with Exai-1 (Extended Data Fig. 9).

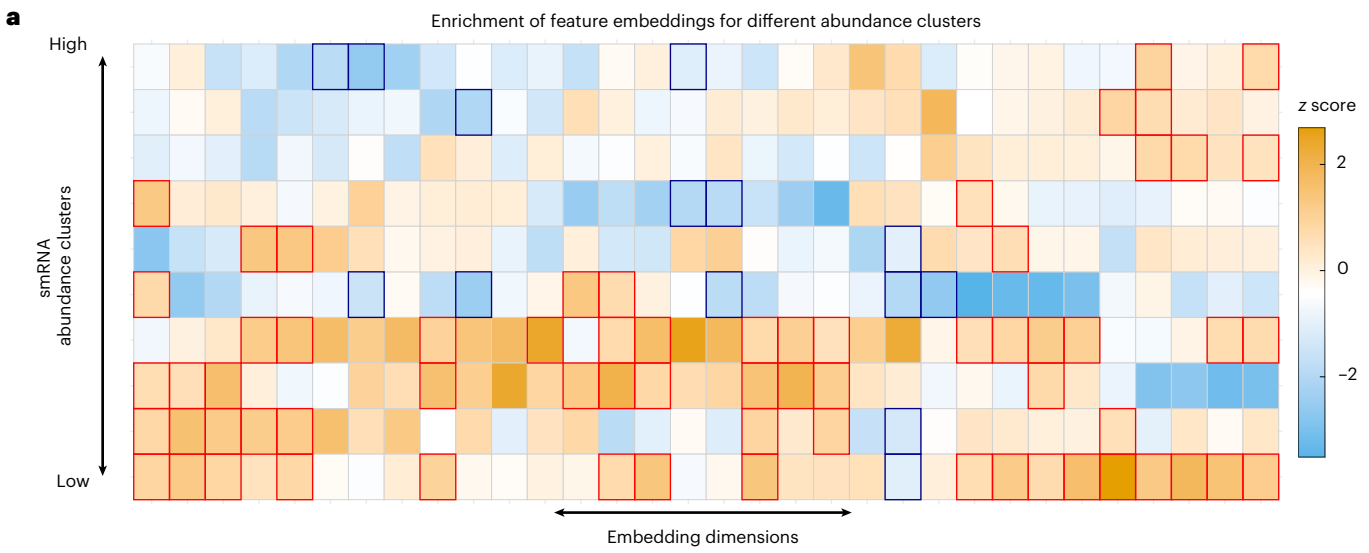
These results highlight Exai-1's potential to effectively address data-type biases inherent in liquid biopsies, making biofluid-specific classifiers broadly applicable to out-of-distribution datasets and, thus, enhancing the robustness and practical utility of future predictive tools in liquid-biopsy-based cancer diagnostics.

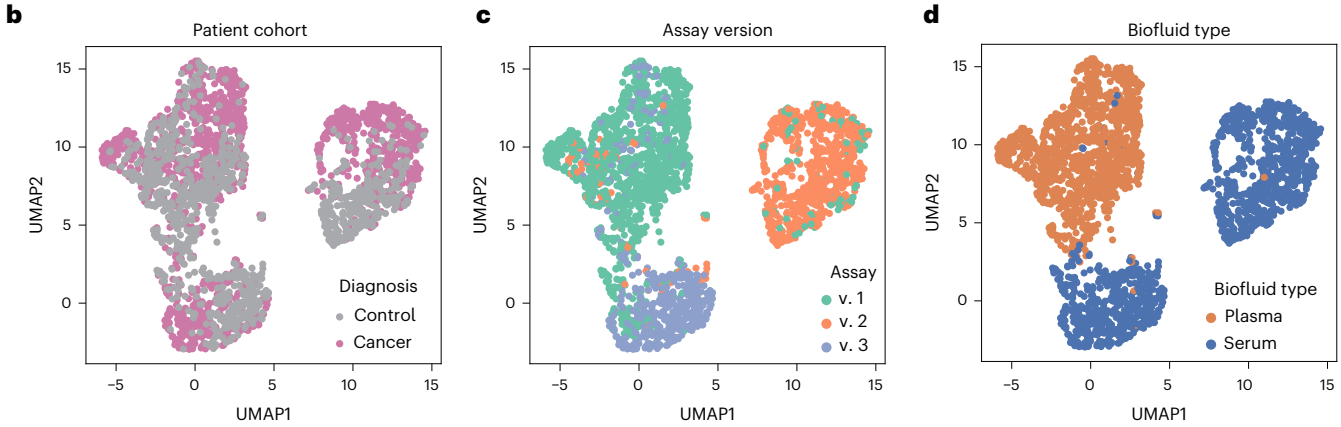
Discussion

In this study, we introduced Exai-1, a multimodal generative model pretrained on a large dataset of cell-free smRNA profiles derived from over 13,000 plasma and serum samples. By integrating RNA-FM-based sequence embeddings with cfRNA abundance data, Exai-1 learns rich representations that capture both structural properties of smRNAs and patterns in their abundance. This multimodal approach allows Exai-1 to bridge the gap between RNA sequence context and cell-type-specific expression dynamics, enabling more biologically meaningful learning.

Exai-1 demonstrates robust performance across clinically relevant tasks, highlighting its potential as a transformative platform for liquid biopsy applications. Its generative capabilities allow for the accurate reconstruction of masked cfRNA profiles, capturing both biological variation and technical artefacts present in cfRNA data. A key enabler for Exai-1's capabilities is the sheer scale of the cfRNA data on which it was pretrained—over 306 billion tokens from 8,339 samples. This massive dataset facilitates the robust pretraining of its more complex architecture. Exai-1 differs from our previous task-specific generative artificial intelligence model, Orion, in several key aspects: it leverages sequence representations, incorporates a self-attention mechanism and uses a deterministic decoder capable of reconstructing cfRNA profiles. These architectural choices, powered by large-scale pretraining, offer clear advantages for downstream tasks. Although Orion may be more suitable for smaller datasets where extensive pretraining is

https://doi.org/10.1038/s42256-025-01148-x





e

Method	Bioconservation: Isolated labels	Leiden NMI	Leiden ARI	*k*-means NMI	*k*-means ARI	Silhouette label	cLISI	Batch correction: BRAS	PCR comparison	Aggregate score: Batch correction	Bio conservation	Total
Exai-1	0.53	0.06	0.01	0.14	0.18	0.53	0.36	0.58	0.36	0.47	0.26	0.34
Harmony	0.50	0.02	0.00	0.00	0.01	0.51	0.17	0.50	0.37	0.43	0.17	0.28
Scanorama	0.50	0.02	0.01	0.00	-0.00	0.50	0.21	0.55	0.29	0.42	0.18	0.27
scANVI	0.51	0.04	0.02	0.00	-0.00	0.51	0.55	0.52	0.09	0.30	0.23	0.26
scVI	0.50	0.02	0.01	0.00	-0.00	0.50	0.19	0.49	0.00	0.24	0.17	0.20
PCA	0.50	0.02	0.00	0.00	0.01	0.51	0.17	0.44	0.00	0.22	0.17	0.19

Fig. 4 | Representation of smRNA characteristics and cancer biology on feature and sample embeddings learned by Exai-1. a, Heat map depicting the standardized values of the learned feature embeddings across various abundance clusters. The significantly enriched embedding dimensions are highlighted with bold red boxes for positive values and bold blue boxes for negative values. **b**, UMAP representation of Exai-1's sample embeddings from the test set, colour coded by cancer and control diagnosis. **c**, Exai-1's sample embeddings from the test set, colour coded by assay versions. **d**, Exai-1's sample embeddings from the test set, colour coded by biofluid type. **e**, Batch effect removal and biological variation conservation benchmark for Exai-1, Scanorama, scVI, scANVI, Harmony and PCA. Scores range from 0 to 1, with higher values indicating better performance. Exai-1 and PCA were applied without access to any batch or biological labels in this dataset. By contrast, the other methods were provided with batch labels (assay version), and scANVI additionally leveraged biological labels (cancer status). cLISI, cell-type local inverse Simpson index; PCR, principal component regression; BRAS, batch removal-adapted silhouette.

not feasible, the differences enabled by the large dataset make Exai-1 a stronger foundation model for cfRNA analysis when sufficient data are available (Extended Data Fig. 4). Furthermore, the Exai-1 latent space, a product of this extensive pretraining, supports better generalizability across biofluids, and its deterministic reconstruction outperforms Orion's probabilistic approach, which models cfRNA using a zero-inflated negative binomial distribution (Extended Data Fig. 9). This property facilitates effective data denoising and enables the generation of synthetic cfRNA profiles, offering a practical solution for augmenting training datasets, particularly in settings where clinical sample availability is limited. Furthermore, Exai-1 embeddings inherently reflect known biological co-expression patterns, preserving clinically relevant structure in a low-dimensional latent space. By capturing biologically meaningful variance and minimizing technical batch effects, Exai-1

https://doi.org/10.1038/s42256-025-01148-x

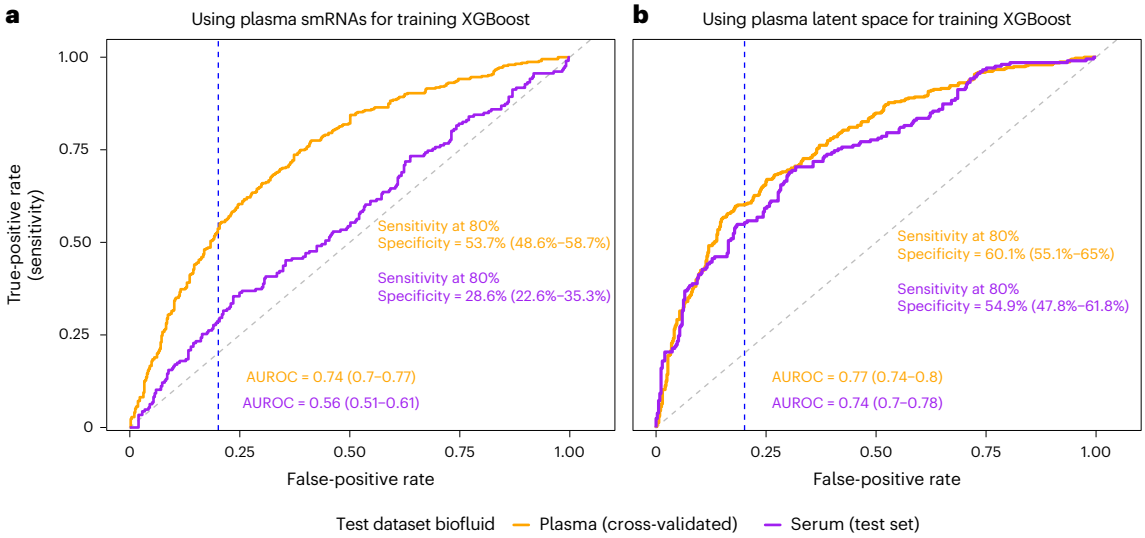


Fig. 5 | Exai-1 allows generalizability for out-of-distribution datasets. **a**, Receiver operating characteristic plot shows the performance of an XGBoost model trained using fivefold cross-validation on plasma smRNAs. Orange shows the cross-validated performance of the model on plasma samples, whereas purple shows the performance of the model on serum samples. **b**, Similar data as **a**, but with the key difference of using Exai-1's plasma sample latent space for training and evaluating the XGBoost models.

emerges as a robust foundation model suitable for a range of clinical diagnostic applications.

One of the major challenges addressed by Exai-1 is the variation introduced by different biofluid types, such as serum versus plasma samples, which often limits the generalizability of conventional classifiers. Serum and plasma each have their own advantages and disadvantages with respect to biobank availability and biomarker signal-to-noise ratio. Exai-1 decouples the biological signal from technical variations within the latent space, making it possible for a plasma-trained model to generalize to out-of-distribution serum data, a task that was not possible in the original feature space (Fig. 5a). This was largely due to the approach we used for minimizing unknown technical variations through triplet margin loss, as well as the sparsity constraints within Exai-1's variational inference objectives. This feature is particularly important for improving the diagnostic accuracy and expanding the clinical applicability of cfRNA-based models across diverse patient populations and clinical contexts.

Previously, we showed the power of variational inference for modelling cancer-specific oncRNAs in which we outperformed conventional methods such as XGBoost[10]. Enriching for a cancer-specific feature space, as we showed, allows for training more accurate cancer-specific classifiers. In line with previous observations, we still observe that when using identical samples, the latent space learned through Exai-1 allows for training better classifiers than when using the original feature space (Fig. 5). This further confirms that Exai-1's representation learning enhances biological signals and disentangling them from confounders, leading to more generalizable classifiers.

Unlike models trained on textual data such as ChatGPT, which benefit from vast amounts of labelled data, Exai-1 operates in a domain where ground-truth samples are inherently scarce and expensive to generate. Although natural language processing foundation models improve as their datasets expand[48,49], cfRNA-based models must navigate data constraints that limit the direct scalability of their training sets. However, as larger and more diverse cfRNA datasets become available, we anticipate continued improvements in Exai-1's performance and clinical utility.

The ability of Exai-1 to improve cross-biofluid generalization and maintain performance in data-limited regimes suggests potential utility in settings where sample availability is constrained. Although Exai-1 was trained and evaluated exclusively on cfRNA data generated using a single platform and processing pipeline, the dataset itself is highly diverse, encompassing both serum and plasma biofluids, three distinct biofluid collection tube types and samples collected from hundreds of international sites. Clinical validation on independent datasets will be an important next step to confirm generalizability and assess performance in broader clinical and research contexts. Future development efforts will focus on enhancing Exai-1 by expanding its training datasets, incorporating additional multimodal inputs and refining its contextual special tokens to further improve disease classification and biomarker discovery. Such improvements are necessary to position Exai-1 as a versatile foundational model, capable of substantially advancing non-invasive cancer diagnostics.

Methods

Dataset

Here we utilized an in-house dataset of serum and plasma collected from 13,014 samples sourced from 11 different suppliers and more than 128 different collection sites. These included 7,265 samples from individuals with cancer and 5,749 samples from individuals without any known history of cancer. We used RNA isolated from 0.5–1 ml of serum or plasma from each donor to generate and sequence cfRNA libraries for each sample, achieving an average depth of 46 ± 26 million single-end reads. Tumours in patients with cancer originated from breast (2,956), lung (1,932), colorectum (753), kidney (313), stomach (307), pancreas (213), bladder (169), prostate (156), liver (145), uterus (95), oesophagus (91), ovary (71) and cervix (64). We used 8,339 samples for training, 2,079 for hyperparameter tuning and model selection (validation set) and 2,596 held-out test set samples for reporting the performance metrics throughout the manuscript.

Data processing

For the processing of raw sequencing files, we used bclconvert (v. 4.0.3), Cutadapt[50] (v. 4.1), FastQC (v. 0.12.1), UMICollapse[51] (v. 1.0.0), Bowtie-2 (ref. 52; v. 2.4.5), SAMtools[53] (v. 1.16.1), pysam[54] (v. 0.20.0) and BEDTools[55] (v. 2.30.0).

Feature selection

We curated a feature set comprising 799,289 smRNAs. Our selection process focused on tRNA, snoRNA, miRNA, yRNA, lncRNA and oncRNAs. The library of oncRNAs was discovered through de novo loci

forming[8,10] and association analyses comparing smRNA sequencing between tumour and adjacent normal tissue samples in The Cancer Genome Atlas. We first excluded any feature detected in fewer than 2% of the samples, thereby removing exceedingly rare cfRNAs. We next imposed a coefficient-of-variation threshold of 1, retaining only those features exhibiting sufficient variability across the data (Extended Data Fig. 1a–e).

This step resulted in 394,175 oncRNAs, 716 miRNAs, 704 tRNAs, 761 yRNAs and 610 snoRNAs. To refine the oncRNA subset further based on sequence and structure information, we used RNA-FM[31] embeddings, extracting the 32 PCs that collectively captured over 91% of the variation. On the basis of these PCs, we applied an iterative *k*-means clustering strategy, that is, starting with 1,000 random clusters. Each cluster with more than ten members was iteratively divided accordingly to result in ten or fewer oncRNAs. The oncRNA closest to the centroid was chosen as a single representative per subcluster. This process yielded 4,558 oncRNAs for inclusion in subsequent analyses (Extended Data Fig. 1e,f).

Architecture of Exai-1

Exai-1 is inspired by the success of variational autoencoders in modelling sparse genomic datasets[44], masked language modelling applications in genomics[24,56], power of leveraging other foundation models for token embedding[25], and capability of combined transformer and variational Bayes architectures in capturing high-level context[57]. Exai-1 facilitates the handling of large-scale cfRNA data without overparameterizing and overfitting to the training data.

Let $\boldsymbol{x}_i \in \mathbb{Z}_+^d$ denote log1p of the CPM-mapped reads for $(d-4)$ smRNAs of the ith sample. We appended four additional rows, $\mathbf{x}_{i,d-3}\ldots\mathbf{x}_{i,d}$ as placeholders for classification tokens (cancer detection and tissue of origin) as well as sample information tokens (collection tube type and assay version). Although classification tokens are always masked by setting them to −1, sample information tokens are always provided. Since we have four classification and sample information tokens, each with t_b different targets, $b = 1\ldots4$, these tokens can be represented as $\boldsymbol{y}_i^b \in \{0 \ldots t_b - 1\}$.

For encoding each smRNA token, we leverage the sequence and structure-aware embeddings of the RNA-FM model by initializing the weights of the embedding layer with a min–max-scaled version of the PCs of RNA-FM embeddings, ensuring the range of weights to be within [−0.1, 0.1] for numeric stability.

Exai-1 encoder works as follows:

1. Token embedding step transforms smRNAs ($\mathcal{X}$) into embedding space $\mathcal{G}$ ($f_t : \mathcal{X} \to \mathcal{G}$). As a result, vector $\boldsymbol{x}_i \in \mathbb{Z}_+^d$ will be projected into the embedding matrix $G_i \in \mathbb{R}^{d\times\delta}$ with the embedding dimension $\delta = 32$.
2. The signal embedding step performs element-wise multiplication of smRNA embeddings with log1p CPM level of expression, resulting in $f_g : \mathcal{G} \to \mathcal{R}$.
3. A dimensionality reduction layer with two-dimensional layer normalization to prevent any numeric instability and a multilayer perceptron to reduce the dimensions from d to 128: $f_h : \mathcal{R} \to \mathcal{H}$.
4. A self-attention transformer layer: $f_{\text{transformer}} : \mathcal{H} \to \tilde{\mathcal{H}}$.
5. A second dimensionality reduction layer generating parameters of a diagonal Gaussian distribution for variational inference, $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$.

The self-attention layer computes the attention scores over the reduced dimension of the hidden layer and embedding dimension $\delta = 32$ through the generation of

$$\begin{aligned} Q \in \mathbb{R}^{128\times32} &= H_i W_Q \\ K \in \mathbb{R}^{128\times32} &= H_i W_K \\ V \in \mathbb{R}^{128\times32} &= H_i W_V \end{aligned}.$$

In Exai-1, the attention output $\tilde{H}_i$, therefore, is a result of two layers of

$$\text{softmax}\left(\frac{QK^T}{\sqrt{\delta}}\right)V,$$

each followed by a two-layer dense network with dimensions of 128, Gaussian error linear unit activation layer[58], layer normalization and a residual connection involving bilinear interpolation of R_i. This step allows each feature representation within H_i to attend to other variables within the same reduced feature space, thereby enhancing the model's ability to capture intricate, context-aware representations of the latent variables.

Exai-1 attempts to learn a probabilistic low-dimensional latent space $\mathbf{z}_i$ from $\tilde{H}_i$, ensuring the preservation of biological information. To achieve this, Exai-1 flattens $\tilde{H}_i$ into $\tilde{\boldsymbol{h}}_i = \text{vec}(\tilde{H}_i)$ and uses a dense layer to learn the parameters of the diagonal Gaussian distribution $\boldsymbol{z}_i \approx \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i I)$.

Exai-1's decoder, therefore, samples from $\mathbf{z}_i$ during training to reconstruct the original input ($\hat{\boldsymbol{x}}_i$). Similar to the encoder, Exai-1's decoder involves the following:

1. A dense layer to increase the dimension from zi to H_i'
2. A self-attention decoder $f'_{\text{transformer}} : \mathcal{H}' \to \tilde{\mathcal{H}}'$ with identical architecture as ftransformer but with independent parameters
3. A dense layer to map the dimensions to the original embedding space: $f_g' : \tilde{\mathcal{H}}' \to \mathcal{R}'$
4. Multilayer perceptrons with 20% dropout, a hidden layer of dimension 16 and a Leaky rectified linear unit activation function. These multilayer perceptrons are responsible for smRNA expression reconstruction as well as token classification.

To emphasize the preservation of biological information during the dimension reduction of Exai-1, we developed a new approach for the selection of positive and negative anchors for computing the triplet margin loss. Previously, we showed that through the explicit definition of known technical variables (for example, batch effects), we can solely minimize the technical differences among the samples through triplet margin loss[10]. Here we show that without any knowledge of technical sources of variation, we can remove batch effects and preserve biological information. We achieved this through the definition of known biological information of the samples during training (for example, tissue of origin for cancer samples).

For each sample, we identify all possible 'positive' anchors $\mathcal{J}$, $i \notin \mathcal{J}$, such that they share the same classification label $\mathbf{y}_i = \mathbf{y}_j$ for all $j \in \mathcal{J}$. We compute the Euclidean distance (d) among all pairs of $\mathbf{z}_i$ and $\mathbf{z}_j$ and sample positive anchors according to the weights set through

$$\boldsymbol{p}_{i,j} = \frac{e^{d(\boldsymbol{z}_i,\boldsymbol{z}_j)}}{\sum_{\mathcal{J}} e^{d(\boldsymbol{z}_i,\boldsymbol{z}_j)}}.$$

Samples of the same biological label that are more distant in the embedding space, therefore, will become closer. For sampling negative anchors, we randomly sample from the pool of all samples with a different biological label.

1. **Reconstruction through masked tokens**. During training, we randomly mask 25% of smRNAs of $\mathbf{x}_{i,d-4}$ at each mini-batch among all the samples, including non-expressed smRNAs, generating $\boldsymbol{x}_{\mathcal{M}}$, akin to the masked language modelling used in BERT[59] and other similar models. The reconstruction Smooth L1 loss[60], therefore, can be represented as

$$\ell_{\text{SL1}} = \begin{cases} \frac{(x-\hat{x})^2}{2}, & \text{if } |x-\hat{x}| < 1 \\ |x-\hat{x}| - 0.5, & \text{otherwise} \end{cases}$$

We compute this loss once for x and once for $\boldsymbol{x}_{\mathcal{M}}$.

2. **KL divergence**: We minimize

$$\ell_{\mathrm{KL}} = D_{\mathrm{KL}}\left(q_{\boldsymbol{z}}(\boldsymbol{z}|\tilde{\boldsymbol{h}})||p(\boldsymbol{z})\right), \quad (1)$$

where D_{KL} is the KL divergence[61] and $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, I)$ is the prior distribution for $\mathbf{z}$, a standard Gaussian.

3. **Triplet margin loss**. For each sample i, we sample ω triplets as follows:
   (a) Randomly pick a 'positive' anchor $j \neq i$ such that they share the same classification label $\mathbf{y}_i = \mathbf{y}_j$ and the sampling probability is assigned according to the distance of the two samples within $\mathbf{z}_i$.
   (b) Randomly pick a 'negative' anchor $j' \neq i$ such that they do not share the same classification label $\boldsymbol{y}_i \neq \boldsymbol{y}_{j'}$.

During training, we add a cost function that minimizes the differences among the samples that share the same biological label (for example, cancers of the same tissue), whereas moving samples with different labels (for example, cancer samples from non-cancer samples) further apart:

$$\ell_{\mathrm{TML}} = \frac{1}{w \times c}\sum_{i}\sum_{(i,j,j')\in\mathcal{T}_i}\max(||\boldsymbol{z}_i - \boldsymbol{z}_j||_2^2 - ||\boldsymbol{z}_i - \boldsymbol{z}_{j'}||_2^2 + \alpha, 0), \quad (2)$$

where α is a hyperparameter that enforces what should be the minimum difference of distances between a sample and its positive and negative anchors in the latent space, and it is set to $\alpha = 0.1$.

4. **Cancer detection**. We require token $d-1$, which is always initialized as missing value indicator −1, to predict the presence of cancer by minimizing the cross-entropy loss:

$$\ell_{\mathrm{CD}} = -\sum_{i=1}^{n} y_i^{\mathrm{CD}} \log[\hat{x}_{i,d-1}].$$

5. **Tissue of origin**. We also require token d, which is always initialized as missing value indicator −1, to predict the tissue of origin in the presence of cancer through the convergence of an additional cross-entropy loss:

$$\ell_{\mathrm{TOO}} = -\sum_{i=1}^{n} y_i^{\mathrm{TOO}} \log[\hat{x}_{i,d}].$$

6. **Assay version**. We require token $d-2$, which is always according to the assay version of the sample of interest, to preserve this information in the decoded output through the convergence of an additional cross-entropy loss:

$$\ell_{\mathrm{AV}} = -\sum_{i=1}^{n} y_i^{\mathrm{AV}} \log[\hat{x}_{i,d-2}].$$

7. **Biofluid type**. We also require token $d-3$, which is always according to the biofluid (serum or plasma), to preserve this information in the decoded output through the convergence of an additional cross-entropy loss:

$$\ell_{\mathrm{BT}} = -\sum_{i=1}^{n} y_i^{\mathrm{BT}} \log[\hat{x}_{i,d-3}].$$

Therefore, Exai-1 attempts to converge the following loss function:

$$\ell_{\mathrm{Exai-1}} = \lambda_1\ell_{\mathrm{SL1}_x} + \lambda_2\ell_{\mathrm{SL1}_{(x_{\mathcal{M}})}} + \lambda_3\ell_{\mathrm{KL}} + \lambda_4\ell_{\mathrm{TML}} + \lambda_5\ell_{\mathrm{CD}} + \lambda_6\ell_{\mathrm{TOO}} + \lambda_7\ell_{\mathrm{AV}} + \lambda_8\ell_{\mathrm{BT}}.$$

Hyperparameters for training Exai-1 were set through grid search and the selected combinations included $\lambda_1, \lambda_2 = 100, \lambda_3, \lambda_5 = 10, \lambda_4, \lambda_6, \lambda_7, \lambda_8 = 1$, learning rate of 0.0001, weight decay of 0.00001, gradient clipping value of 0.1, mini-batch size of 256, an training for 5,000 epochs. The efficient architecture of Exai-1 allowed us to train the model over 1 NVIDIA T4 GPU for 60 h.

PCA-based data augmentation. We applied PCA with 32 components—matching the latent dimensionality of Exai-1—to the z-scored log1p CPM matrix of the training dataset. Gaussian noise (mean, 0; standard deviation, 0.1) was then added to the resulting embedding space, and the perturbed embeddings were projected back to the original feature space to generate the reconstructed profiles. These reconstructed profiles were used for benchmarking the data augmentation performance.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The pretrained model and cfRNA expression matrices are available via Zenodo at https://doi.org/10.5281/zenodo.16943244 (ref. 62). Other data are available from the corresponding authors upon request provided criteria for academic research use are met and data transfer agreements are in place. Please contact info@exai.bio for all data access requests.

Code availability

The Exai-1 code repository is available via GitHub at https://github.com/exai-oss/exai-1 as well as via Zenodo at https://doi.org/10.5281/zenodo.17401411 (ref. 63). The repository includes the code for reproducing all the key results.

References

1. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nat. Rev. Clin. Oncol.* **10**, 472–484 (2013).
2. Corcoran, R. B. & Chabner, B. A. Application of cell-free DNA analysis to cancer treatment. *N. Engl. J. Med.* **379**, 1754–1765 (2018).
3. Millner, L. M., Linder, M. W. & Valdes, R. Circulating tumor cells: a review of present methods and the need to identify heterogeneous phenotypes. *Ann. Clin. Lab. Sci.* **43**, 295–304 (2013).
4. Roskams-Hieter, B. et al. Plasma cell-free RNA profiling distinguishes cancers from pre-malignant conditions in solid and hematologic malignancies. *npj Precis. Oncol.* **6**, 28 (2022).
5. Sundling, K. E. & Lowe, A. C. Circulating tumor cells: overview and opportunities in cytology. *Adv. Anat. Pathol.* **26**, 56–63 (2019).
6. Cescon, D. W., Bratman, S. V., Chan, S. M. & Siu, L. L. Circulating tumor DNA and liquid biopsy in oncology. *Nat. Cancer* **1**, 276–290 (2020).
7. Ignatiadis, M., Sledge, G. W. & Jeffrey, S. S. Liquid biopsy enters the clinic—implementation issues and future challenges. *Nat. Rev. Clin. Oncol.* **18**, 297–312 (2021).
8. Fish, L. et al. Cancer cells exploit an orphan RNA to drive metastatic progression. *Nat. Med.* **24**, 1743–1751 (2018).
9. Wang, J. et al. Systematic annotation of orphan RNAs reveals blood-accessible molecular barcodes of cancer identity and cancer-emergent oncogenic drivers. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.19.585748 (2024).
10. Karimzadeh, M. et al. Deep generative AI models analyzing circulating orphan non-coding RNAs enable detection of early-stage non-small cell lung cancer. *Nat. Commun.* **15**, 10090 (2024).

11. OpenAI. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
12. Consens, M. E. et al. Transformers and genome language models. *Nat. Mach. Intell.* **7**, 346–362 (2025).
13. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
14. Linder, J., Srivastava, D., Yuan, H. et al. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat. Genet.* **57**, 949–961 (2025); https://doi.org/10.1038/s41588-024-02053-6
15. Fu, X., Mo, S., Buendia, A. et al. A foundation model of transcription across human cell types. *Nature* **637**, 965–973 (2025); https://doi.org/10.1038/s41586-024-08391-z
16. Sanabria, M., Hirsch, J., Joubert, P. M. et al. DNA language model GROVER learns sequence context in the human genome. *Nat. Mach. Intell.* **6**, 911–923 (2024); https://doi.org/10.1038/s42256-024-00872-0
17. Brixi, G. et al. Genome modeling and design across all domains of life with Evo 2. Preprint at *bioRxiv* https://doi.org/10.1101/2025.02.10.634996 (2025).
18. Celaj, A. et al. An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.20.558508 (2023).
19. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
20. Connell, W., Khan, U. & Keiser, M. J. A single-cell gene expression language model. Preprint at https://arxiv.org/abs/2210.14330 (2022).
21. Yang, X., Liu, G., Feng, G. et al. GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res.* **34**, 830–845 (2024); https://doi.org/10.1038/s41422-024-01034-y
22. Hao, M., Gong, J., Zeng, X. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024); https://doi.org/10.1038/s41592-024-02305-7
23. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
24. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
25. Rosen, Y. et al. Universal cell embeddings: a foundation model for cell biology. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.28.568918 (2023).
26. Kedzierska, K. Z., Crawford, L., Amini, A. P. & Lu, A. X. Assessing the limits of zero-shot foundation models in single-cell biology. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.19.563263 (2023).
27. Boiarsky, R., Singh, N. M., Buendia, A., Getz, G. & Sontag, D. A deep dive into single-cell RNA sequencing foundation models. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.16.562636 (2023).
28. Liu, T., Li, K., Wang, Y., Li, H. & Zhao, H. Evaluating the utilities of large language models in single-cell data analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.25.559297 (2023).
29. Larson, M. H. et al. A comprehensive characterization of the cell-free transcriptome reveals tissue-and subtype-specific biomarkers for cancer detection. *Nat. Commun.* **12**, 2357 (2021).
30. Zirak, B. et al. Revealing the grammar of small RNA secretion using interpretable machine learning. *Cell Genom.* **4**, 100522 (2024).
31. Shen, T. et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nat. Methods* **21**, 1414–1425 (2024).
32. Chen, J. et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. Preprint at https://arxiv.org/abs/2204.00300 (2022).
33. Schreiber, J., Singh, R. & Bilmes, J. A pitfall for machine learning methods aiming to predict across cell types. *Genome Biol.* **21**, 282 (2020).
34. Van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
35. Zirak, B., Mohsen N., Ali S. et al. Revealing the grammar of small RNA secretion using interpretable machine learning. *Cell Genom.* **4**, 100522 (2024); https://doi.org/10.1016/j.xgen.2024.100522
36. Garcia-Martin, R. et al. MicroRNA sequence codes for small extracellular vesicle release and cellular retention. *Nature* **601**, 446–451 (2022).
37. Tosar, J. P. et al. Assessment of small RNA sorting into different extracellular fractions revealed by high-throughput sequencing of breast cell lines. *Nucleic Acids Res.* **43**, 5601–5616 (2015).
38. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
39. Melhuish, T. A. et al. Myt1 and Myt1l transcription factors limit proliferation in GBM cells by repressing YAP1 expression. *Biochim. Biophys. Acta Gene Regul. Mech.* **1861**, 983–995 (2018).
40. Blevins, M. A., Huang, M. & Zhao, R. The role of CtBP1 in oncogenic processes and its potential as a therapeutic target. *Mol. Cancer Ther.* **16**, 981–990 (2017).
41. Xiang, Y., Tian, Q., Guan, L. & Niu, S.-S. The dual role of miR-186 in cancers: oncomir battling with tumor suppressor miRNA. *Front. Oncol.* **10**, 233 (2020).
42. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
43. Hie, B. L., Kim, S., Rando, T. A., Bryson, B. & Berger, B. Scanorama: integrating large and diverse single-cell transcriptomic datasets. *Nat. Protoc.* **19**, 2283–2297 (2024).
44. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
45. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
46. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
47. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
48. Zhou, C., Li, Q., Li, C. et al. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. *Int. J. Mach. Learn. & Cyber.* https://doi.org/10.1007/s13042-024-02443-6 (2024).
49. Jaegle, A. et al. Perceiver IO: a general architecture for structured inputs & outputs. Preprint at https://arxiv.org/abs/2107.14795 (2021).
50. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
51. Liu, D. Algorithms for efficiently collapsing reads with unique molecular identifiers. *PeerJ* **7**, e8275 (2019).
52. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421–432 (2019).
53. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. Bonfield, J. K. et al. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**, giab007 (2021).

55. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
56. de Lima Camillo, L. P. et al. CpGPT: a foundation model for DNA methylation. Preprint at *bioRxiv* https://doi.org/10.1101/2024.10.24.619766 (2024).
57. Wang, T. & Wan, X. T-CVAE: transformer-based conditioned variational autoencoder for story completion. In *Proc. International Joint Conference on Artificial Intelligence* 5233–5239 (IJCAI Organization, 2019).
58. Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs). Preprint at https://arxiv.org/abs/1606.08415 (2016).
59. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805 (2019).
60. Girshick, R. Fast R-CNN. In *Proc. IEEE International Conference on Computer Vision* 1440–1448 (IEEE, 2015).
61. Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
62. Karimzadeh, M. et al. Datasets accompanying A multi-modal cell-free RNA language model for liquid biopsy applications. *Zenodo* https://doi.org/10.5281/zenodo.16943244 (2025).
63. Karimzadeh, M. exai-oss/exai-1: v1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.17401411 (2025).

Acknowledgements

The results shown here are in part based on the data generated by the The Cancer Genome Atlas Research Network (https://www.cancer.gov/tcga). H.G. is a core investigator at Arc Institute. This study is funded by Exai Bio Inc.

Author contributions

Authors are listed in alphabetical order for all contributions. B.A., H.G., F.H., M.K. and A.M.S. designed and planned the study. B.A., T.B.C., N.-C.C., L.F., H.G. and F.H. defined the smRNA identities. B.A., H.G., F.H. and M.K. developed the Exai-1 algorithm. M.K. and A.M.S. performed the analyses. N.-C.C., S.S. and J.W. processed and analysed the datasets. A.M.-R., A.M.S., B.B. and M.K. prepared the final version of the released code and datasets. L.F., A. Hartwig, A. Huang, R.H., S.C., T.L. and D.N. accessioned and processed the samples in the laboratory. B.A., H.G., F.H., A.M.S. and M.K. wrote the manuscript and incorporated feedback from other authors. B.B., N.-C.C. and A.M.-R. provided feedback on the final version of the manuscript.

Competing interests

B.A. is a cofounder and shareholder of Exai Bio Inc. H.G. is a cofounder, shareholder, and board member of Exai Bio Inc. F.H. is a former employee and current advisor to the company. All other authors declare no competing interests.

Additional information

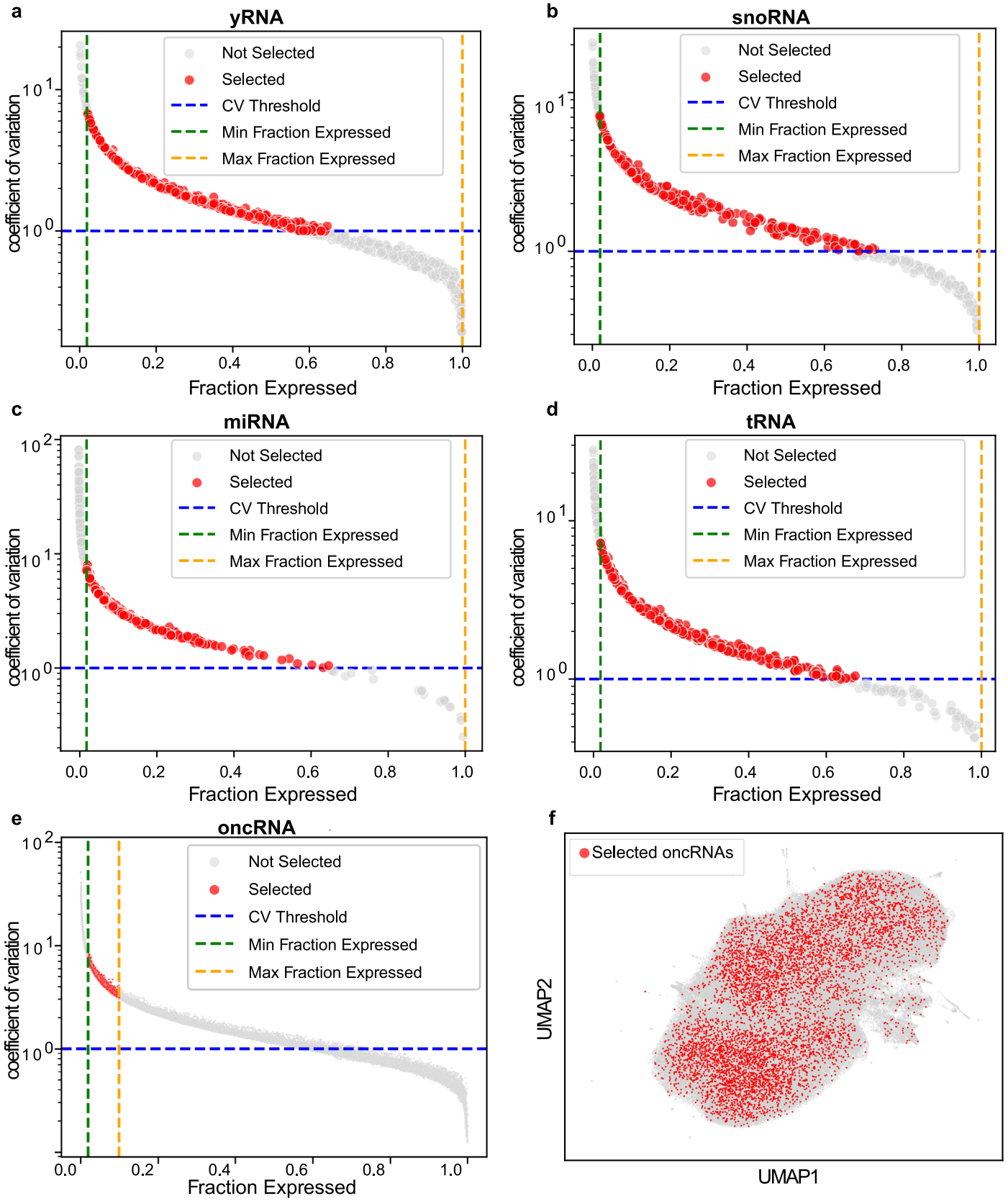
Extended data is available for this paper at https://doi.org/10.1038/s42256-025-01148-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-025-01148-x.

Correspondence and requests for materials should be addressed to Fereydoun Hormozdiari, Babak Alipanahi or Hani Goodarzi.

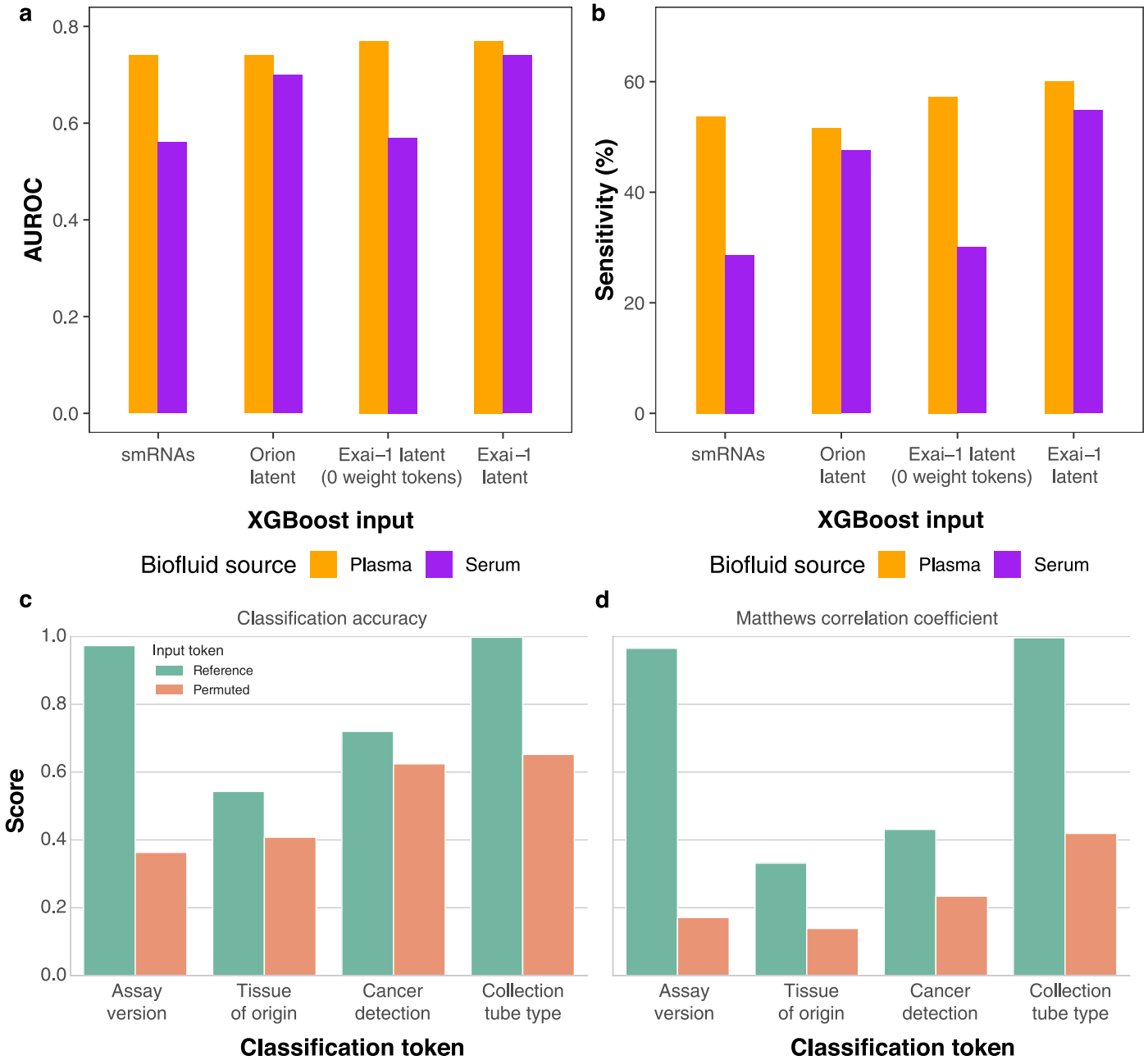
Peer review information *Nature Machine Intelligence* thanks Yingying Cao, Eytan Ruppin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-025-01148-x



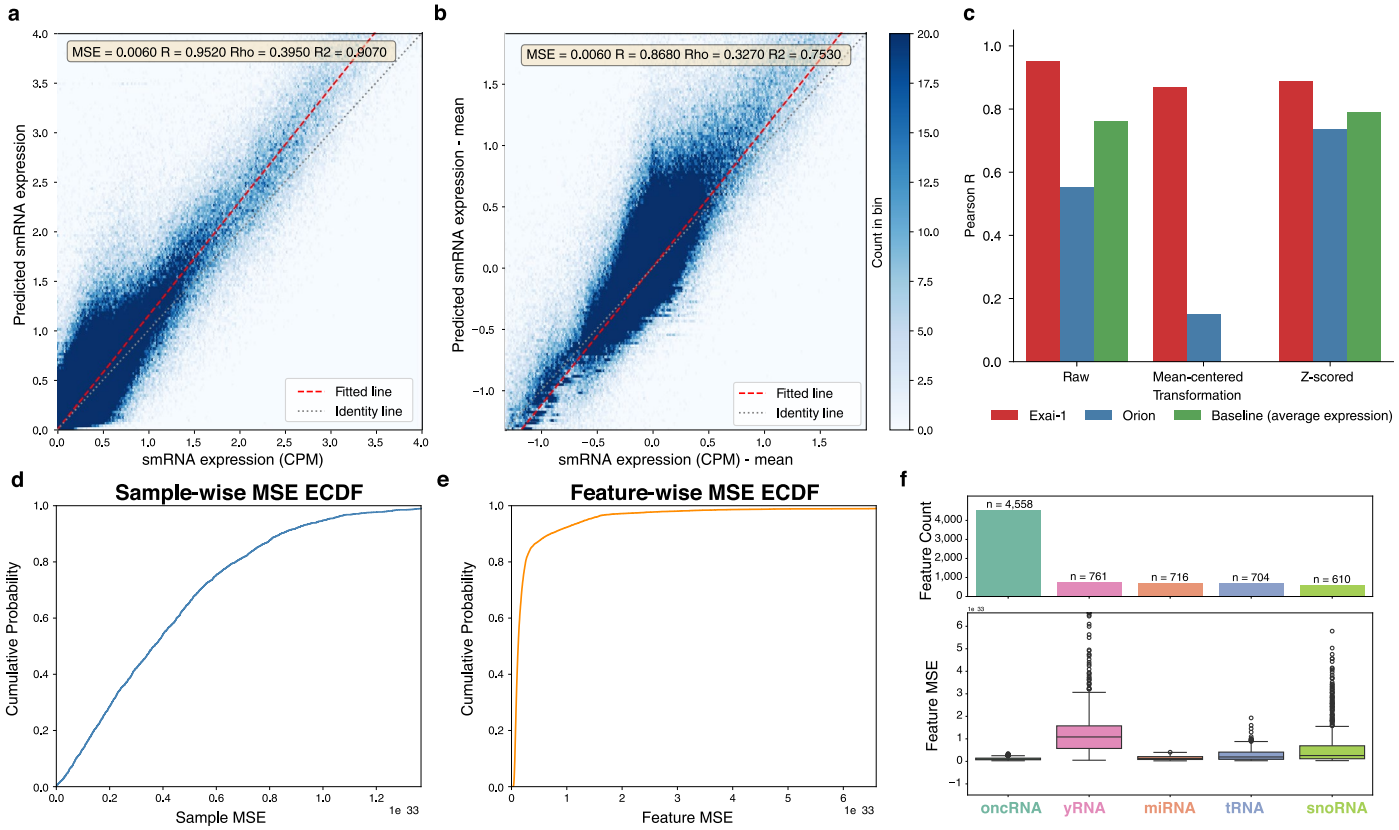
Extended Data Fig. 1 | Feature selection workflow for Exai-1. (**a**) X-axis shows the fraction of samples expressing each yRNA and Y-axis shows the coefficient of variation (CV). We used the subset of samples that surpassed the CV threshold (blue dashed line) while falling within minimum and maximum fraction-expressed cutoffs (green and orange dashed lines, respectively). (**b**–**e**) similar to (**a**) for snoRNA, miRNA, tRNA, and oncRNA, respectively. (**f**) UMAP projection of RNA-FM embeddings for the initially retained cfRNAs, highlighting in red the 4,558 representative oncRNAs selected from an initial pool of 394,175.

https://doi.org/10.1038/s42256-025-01148-x



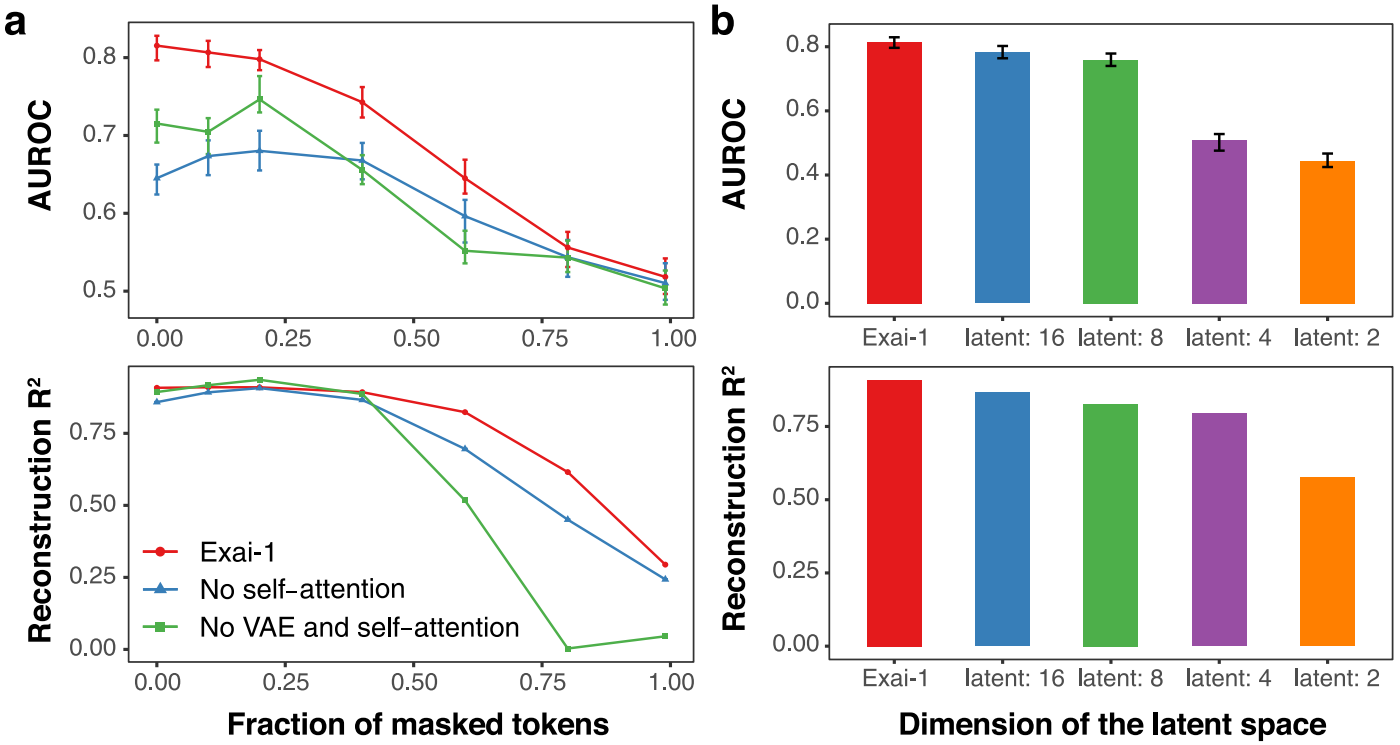
Extended Data Fig. 2 | Preservation of biological signals via classification tokens. (**a**) Cross-biofluid AUROC when models are trained on plasma samples and evaluated by cross-validation on plasma (orange, N=876) and on held-out paired serum samples (purple, N=559). We compare XGBoost trained on raw smRNA features, Orion latent space, Exai-1 latent space with classification-token loss weights set to w=0 (tokens ablated), and the default Exai-1 latent space. (**b**) Same as (**a**), showing sensitivity at 80% specificity. (**c**) Accuracy of Exai-1's classification tokens on the held-out test set (N=2,596) under the reference setting (green) versus after permuting token values across samples (orange). (**d**) Same as (**d**), showing Matthews correlation coefficient on the y-axis.

https://doi.org/10.1038/s42256-025-01148-x



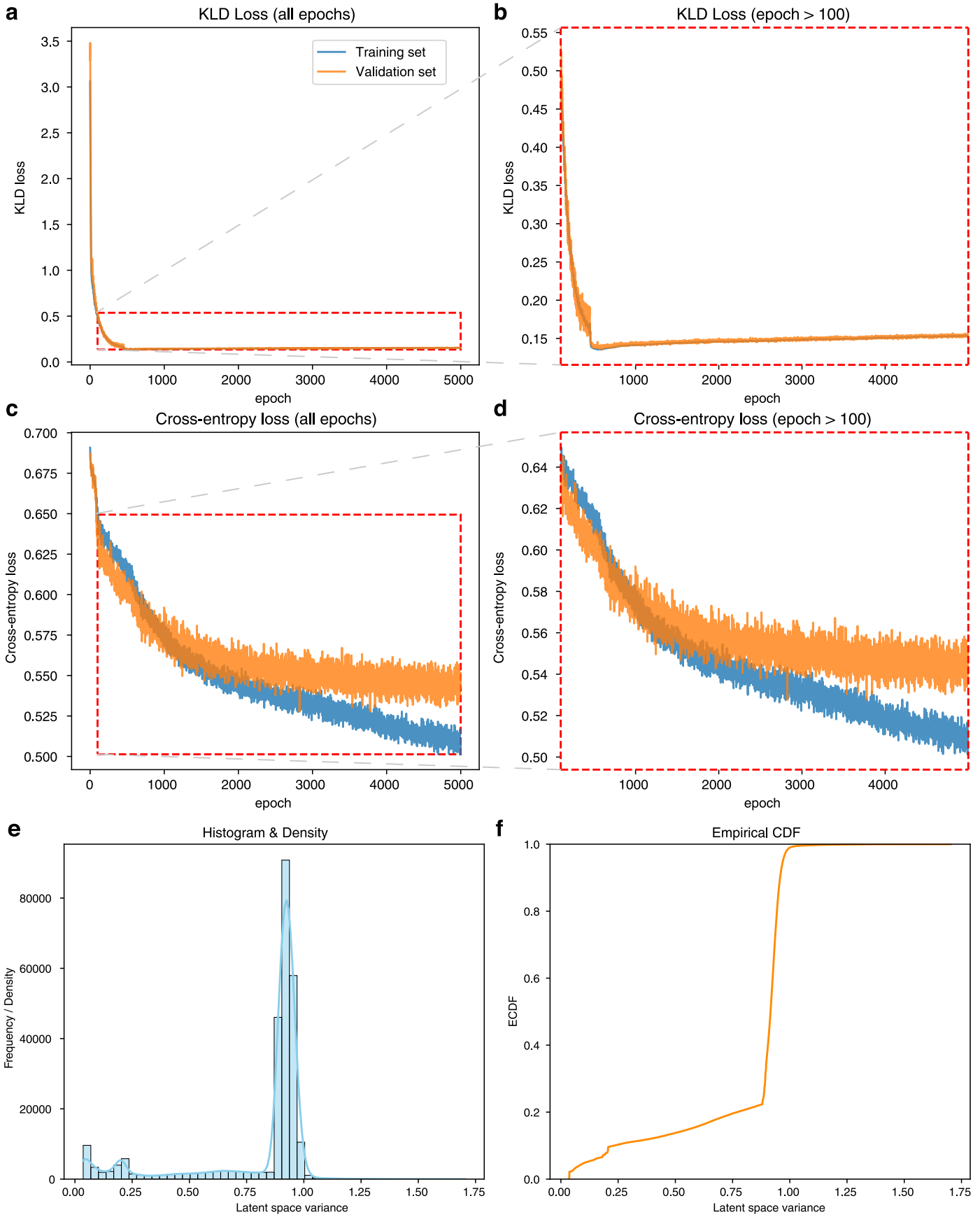
Extended Data Fig. 3 | Reconstruction performance. (**a**) Density scatterplot comparing observed smRNA log1p CPM (x-axis) with Exai-1 reconstructions (y-axis) for masked features in the held-out test set. (**b**) As in (**a**), after per-feature mean centering of log1p CPM (value of each feature minus its across-sample mean). (**c**) Pearson correlation (r) between observed and reconstructed values for Exai-1 (red), Orion (blue), and a baseline that uses each feature's across-sample mean (green). Left: raw log1p CPM; middle: mean-centered; right: per-feature z-scored. (**d**) Empirical cumulative distribution function (ECDF) of per-sample reconstruction MSE. (**e**) ECDF of per-feature reconstruction MSE. (**f**) Boxplots of per-feature reconstruction MSE stratified by small-RNA biotype; the bar plot above indicates the number of features per biotype.

https://doi.org/10.1038/s42256-025-01148-x



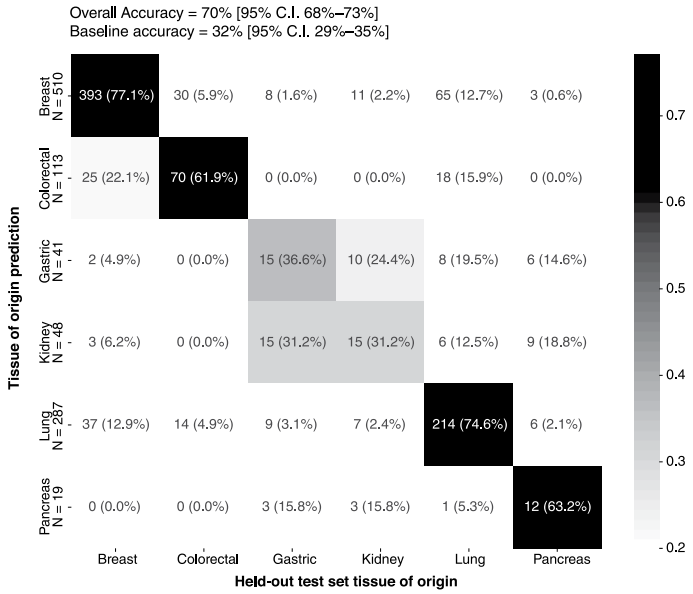
Extended Data Fig. 4 | Hyperparameter impacts on Exai-1 performance. (**a**) We examined the performance of Exai-1 (red) compared to an ablated Exai-1 model without the transformer components (blue) or without the entire transformer-VAE component (green). Top panel shows the AUROC for cancer detection task (y-axis) when masking different percent of tokens (x-axis). Similarly, bottom panel shows the reconstruction performance as measured by R^2 (y-axis). (**b**) We examined the impact of different latent dimensions on AUROC of the cancer detection token (top; error bars indicate 95% C.I. estimate of AUC over 2,596 samples of the held-out test set) as well as the reconstruction task R^2 (bottom; average over N = 2,596 samples and 7,349 features from the held-out test set). X-axis shows Exai-1 (red) with optimal latent size of 32 and other models using a smaller latent size (2–16).

https://doi.org/10.1038/s42256-025-01148-x



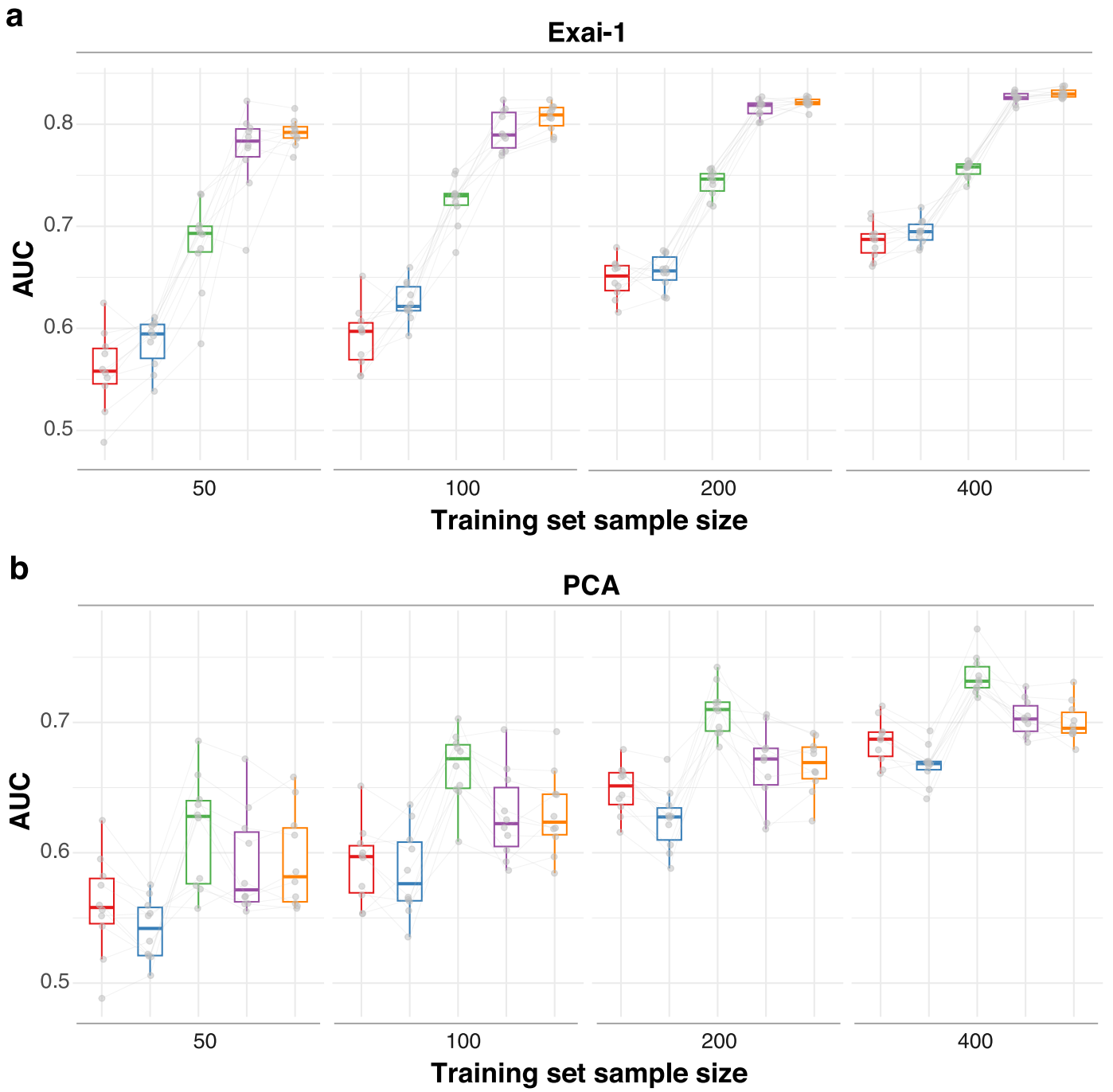
Extended Data Fig. 5 | Kullback-Leibler divergence. (**a**) Convergence of KLD loss for training (blue) and validation set (orange). (**b**) Same as (**a**), but starting at epoch 100. (**c**) Cross-entropy loss for cancer detection token. (**d**) Same as (**c**), but starting at epoch 100. (**e**) Histogram and density plot of flattened latent variance matrix. (**f**) Empirical cumulative density function of the flattened latent variance matrix.

https://doi.org/10.1038/s42256-025-01148-x



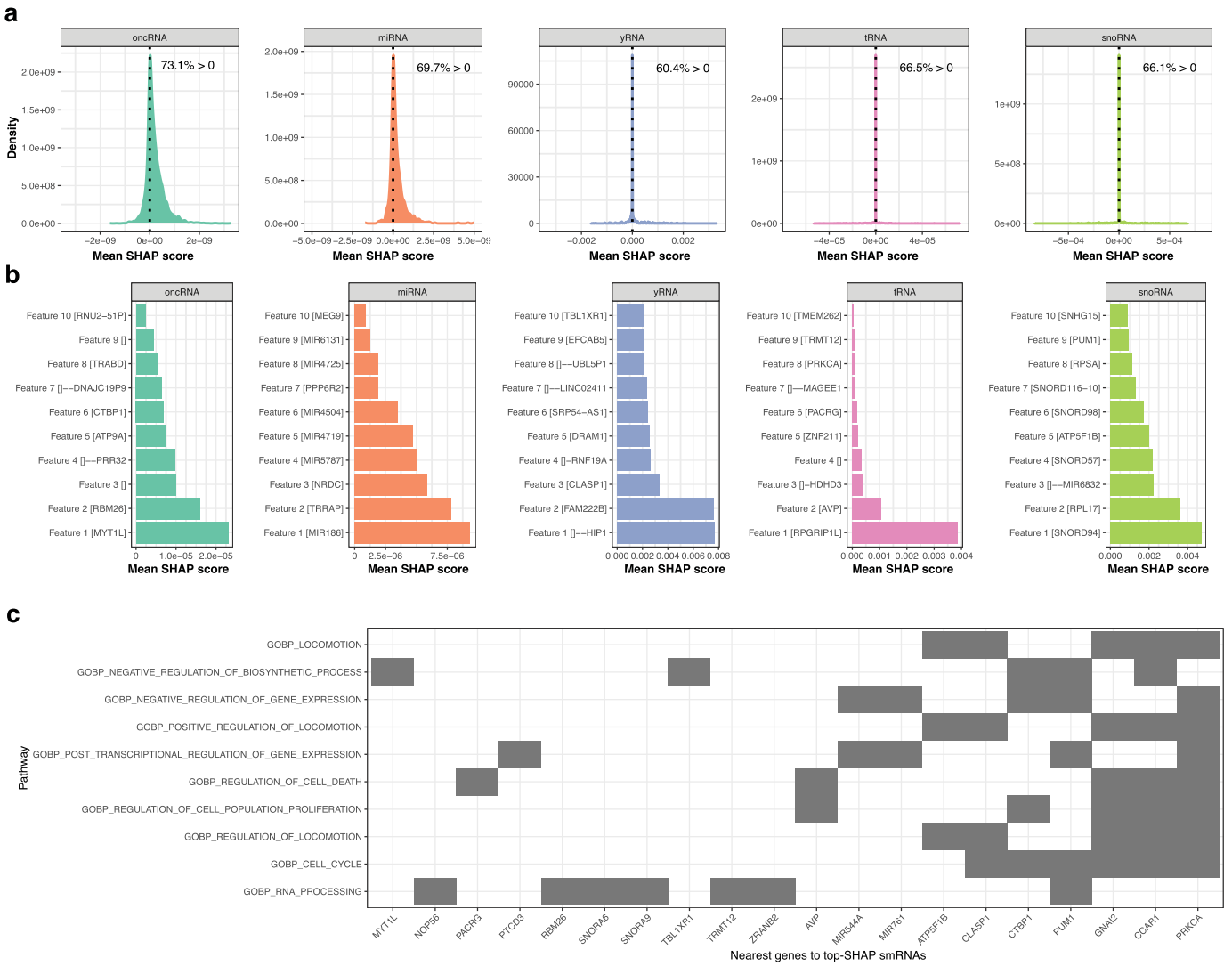
Extended Data Fig. 6 | Tissue of origin performance. Performance of a multi-class XGBoost model trained on latent space of the training set to predict cancer samples tissue of origin and applied on the latent space of the held-out test set. X-axis shows the true labels and y-axis shows the predicted labels.

https://doi.org/10.1038/s42256-025-01148-x



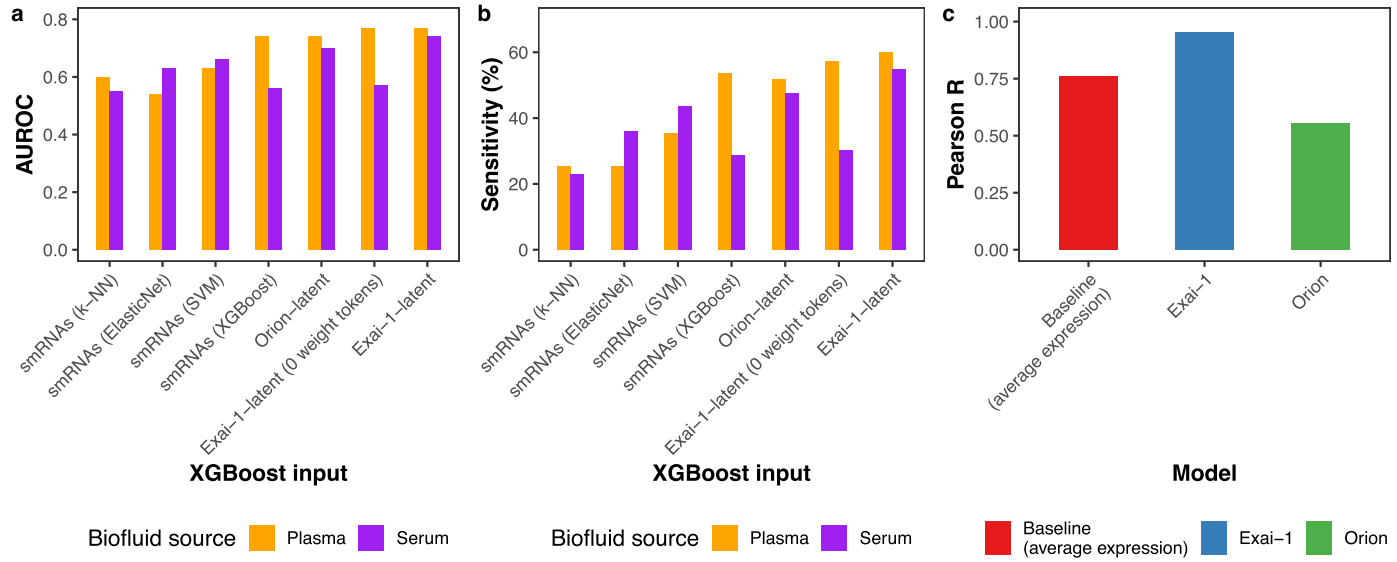
Extended Data Fig. 7 | Impact of using feature space and latent space for downstream applications. (**a**) Boxplots show performance of XGBoost model trained on original feature space (red), original and Exai-1-augmented data (blue), only Exai-1-augmented data (green), Exai-1 latent space data (purple), and Exai-1 latent space with augmentation (orange). The only-augmented set (green), marked with * in the legend, is trained on the reconstructed space of the training set. Augmented data in the original space are generated through the decoder of Exai-1. Latent space is generated through the encoder of Exai-1. Augmented data in latent space of Exai-1 are generated through 10% masking and generative sampling. Boxplots show sampling with 10 different random seeds per experiment. (**b**) Same as (**a**), but using PCA instead of Exai-1. We used PCA with the same latent dimension as Exai-1. Augmented data add Gaussian noise (mean=0, σ=1) to PCA latent space and use it either as is (latent + augmented) or reconstruct it back to the original dimension (only augmented and original + augmented).

https://doi.org/10.1038/s42256-025-01148-x



Extended Data Fig. 8 | Feature-level interpretability via SHAP for the cancer-detection head. (**a**) Kernel density estimate (KDE) of SHAP values by small-RNA biotype on the held-out test set. Labels report the fraction of features with SHAP > 0 (that is, pushing the prediction toward 'cancer'). (**b**) Top 10 features per biotype ranked by median SHAP (x-axis), annotated with the nearest gene. Brackets encode genomic proximity (feature midpoint to nearest gene body): [A] overlap with gene A; []A within 1 kilobase pair; [−]A within 10 kilobase pairs; [− −] A within 100 kilobase pairs; [] no gene within 1 megabase pair. (**c**) Gene Ontology (GO) Biological Process enrichment of genes from (**b**); filled cells indicate pathway–gene membership. Only terms passing FDR < 0.05 are shown.

https://doi.org/10.1038/s42256-025-01148-x



Extended Data Fig. 9 | Comparison of Exai-1 with other methods. (**a**) Barplot shows AUROC of different models fitted on plasma smRNAs, latent space of the Orion model, and latent space of the Exai-1 model (right). Orange: cross-validated performance on plasma samples, N=876. Purple: test-set performance of plasma-trained XGBoost models on serum samples of the same patients, N=559. Patients with multiple replicates are limited to the same folds during cross-validation. (**b**) Similar to (**a**), with y-axis representing sensitivity at 80% specificity. (**c**) Pearson R for correlation of expression of each smRNA in each sample compared to the average of 100 random samples from Orion's zero-inflated negative binomial model (left; green), Exai-1 decoder output (middle; blue), and average expression among all samples (right; red).

nature portfolio

Corresponding author(s): Fereydoun Hormozdiari
Babak Alipanahi
Hani Goodarzi

Last updated by author(s): Oct 20, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	Model architecture and example notebook for training and reproducing the key results is provided at https://github.com/exai-oss/exai-1/ and deposited to https://doi.org/10.5281/zenodo.17401411. Data analysis was done through custom code which is available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We have provided the pre-trained model and cfRNA expression matrices on Zenodo (DOI: 10.5281/zenodo.16943244). Other data will be made available upon request, subject to criteria for academic research use and data transfer agreements. Please contact info@exai.bio for all data access requests.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	13014. The number of datasets included in this study was determined based on the availability of in-house high-quality, well-annotated cfRNA profiles.
Data exclusions	During sample selection, we excluded samples of individuals matching any of the following criteria: Cancer patient is not treatment-naive at time of collection Age ≤ 18 Prior cancer diagnosis (with or without treatment) Any surgery within one month of collection Received non-cancer systemic immune modulation therapy within 60 days of collection (e.g. monoclonal antibodies) History of organ transplantation Infusion of any blood products within 30 days of collection Current active COVID-19 infection Any prior history of cancer therapy (e.g. surgical, radiation, or medical including neoadjuvant treatment) Current or prior pregnancy within 12 months of collection
Replication	N/A. The analyses were based on pre-existing cfRNA sequencing datasets rather than newly generated experimental measurements. All data used were obtained from independent patient cohorts collected under standardized protocols, and the study focused on computational modeling, cross-validation, and independent dataset testing rather than experimental replication. Model robustness and reproducibility were instead ensured through multiple validation strategies, including cross-cohort generalization, repeated subsampling, and performance benchmarking on held-out datasets. These approaches serve the same purpose as replication by confirming that the findings are consistent, reproducible, and not driven by any single dataset or random variation.
Randomization	Training, validation, and test set adjusted for sample source and cohort (cancer vs. control).
Blinding	Blinding was not applicable, as the study involved retrospective computational analyses of de-identified cfRNA datasets with no direct investigator interaction during data collection.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.