Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CoxFormer enables spatial omics inference with multimodal generative modeling

Jin Liu

liujinlab@cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen https://orcid.org/0000-0002-5707-2078

Yiyang Yang

The Chinese University of Hong Kong Shenzhen https://orcid.org/0009-0009-9746-9529

Xu Liao

Columbia University

Haoyu Zhang

The Chinese University of Hong Kong Shenzhen

Yida Wu

The Chinese University of Hong Kong Shenzhen

Xiaobo Sun

Emory University

Yao Wang

Xi'an Jiaotong University

Tianshu Yu

School of Data Science, The Chinese University of Hong Kong - Shenzhen https://orcid.org/0000-0002-6537-1924

Article

Keywords:

Posted Date: March 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9062650/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on August 20th, 2026. See the published version at https://doi.org/10.1038/s41467-026-76404-8.

CoxFormer enables spatial omics inference with multimodal generative modeling

Yiyang Yang[1,3†], Xu Liao[1†], Haoyu Zhang[1†], Yida Wu[1], Xiaobo Sun[2], Yao Wang[3*], Tianshu Yu[1*], and Jin Liu[1*]

[1]School of Data Science, The Chinese University of Hong Kong-Shenzhen, Shenzhen, China,

[2]Department of Human Genetics, Emory University School of Medicine, Atlanta, USA.

[3]School of Management, Xi'an Jiaotong University, Xi'an, China,

[†]Equal contributions.

[*]Corresponding author. Email: yao.s.wang@gmail.com, yutianshu@cuhk.edu.cn, liujinlab@cuhk.edu.cn

Abstract

Gene co-expression provides a principled basis for transcriptome-wide gene–gene relationships, yet the available high-quality estimates typically cover less than half of the genome. In the current biotechnological landscape, spatial omics faces complementary measurement constraints: targeted *in situ* platforms restrict gene coverage, whereas sequencing-based assays lack cellular resolution. When extended to the entire transcriptome, co-expression relationships may help to address these limitations by enabling the generation of unassayed gene expression data at a subcellular resolution. Here, we present CoxFormer (Co-expression pre-trained Transformer). This universal gene-embedding framework uniquely integrates literature-derived gene knowledge with data-driven co-expression networks from bulk tissues and large-scale single-cell atlases to model transcriptome-wide gene–gene relationships. CoxFormer learns robust 512-dimensional embeddings for 32,016 human genes, capturing their functional and regulatory structure. It also serves as a generative prior that is deployed through a flexible framework for transcriptome-wide spatial inference under multimodal contexts. We demonstrate that this generative framework unifies four key spatial applications beyond measured genes: histology-based imputation, gene activity prediction for chromatin accessibility, subcellular super-resolution inference for unassayed genes, and the detection of pathological tissue regions. Ultimately, CoxFormer provides a versatile foundation for multimodal spatial omics analysis, leading to whole-transcriptome insights beyond the constraints of limited gene panels.

Introduction

Gene co-expression reflects the coordinated regulation of genes across diverse cellular states and has long served as an important indicator for understanding gene functions and mechanisms [1, 2]. At the transcriptome-wide scale, such co-expression patterns help systematically outline gene-gene relationships and provide a principled basis for transcriptomic analysis. However, existing high-quality co-expression estimates typically rely on large curated databases and stringent expression-filtering strategies. These estimates often exclude lowly expressed genes and restrict analyses to a limited subset of the transcriptome, typically covering less than 50% of genes, thereby constraining the use of co-expression knowledge [3, 4].

A similar coverage issue arises in the field of spatial transcriptomics. Specifically, targeted spatial transcriptomics platforms such as Xenium [5], CosMx [6], and MERFISH [7, 8] provide high-sensitivity, subcellular spatial resolution, but they rely on predefined gene panels, precluding unbiased whole-transcriptome profiling. While traditional imputation methods can extend gene coverage, they typically require single-cell RNA sequencing (scRNA-seq) references for unassayed genes and are developed for a single modality, limiting their transfer capability to other modalities such as transcriptomics and epigenomics. Another challenge lies in sequencing-based spatial transcriptomics platforms, such as ST [9], 10x Visium, and Slide-seq [10]. These profile expression at spot-level resolution across the transcriptome, with each spot typically aggregating transcripts from multiple cells. The absence of transcriptome-wide gene–gene relationships means these existing resolution-enhancement methods are only useful for genes directly measured by the assay and lack generative ability.

A universal gene embedding that captures transcriptome-wide gene–gene relationships and enables the generation of unassayed gene expression may help address these limitations. Existing efforts to learn gene representations broadly fall into two categories. Transformer-based foundation models, including scBERT [11], Geneformer [12], and scGPT [13, 14], are extensively pretrained on large-scale single-cell datasets, demanding massive corpora and substantial computational resources [15, 16]. Alternative methods leverage existing large language models to extract gene knowledge from the biomedical literature, as exemplified by GenePT [17] and its subsequent applications, e.g., GenePert [18], Scouter [19], and scLAMBDA [20]. However, these text-derived embeddings often fail to incorporate domain-specific biological knowledge, such as experimentally validated and quantitatively curated co-expression relationships. Beyond these methodological challenges, the utility of existing methods is frequently limited by the narrow application scopes, which largely focus on the static representational power of gene embeddings for gene- and cell-level prediction tasks while ignoring their generative potential. As a result, the corresponding generative strategies remain poor at leveraging gene embeddings to impute expression beyond measured panels and handle multimodal datasets.

Taken together, an ideal gene-embedding framework should be capable of tackling the following three tasks: (1) Integrate complementary sources of biological information, such as literature-derived knowledge of gene functions or pathways, and quantitative relationship data, including co-expression networks or correlation patterns; (2) Provide a single set of robust and reusable gene embeddings that can be directly applied to diverse downstream multimodal applications at the gene, cell, and tissue levels across various technologies and platforms; (3) Capture the inherent complex relationships among genes to enable whole-transcriptome

inference and predict molecular profiles for previously unmeasured genes solely from the measured panel.

To address these challenges in molecular profiling and gene embedding, we introduce the CoxFormer (Co-expression pre-trained Transformer) package. Unlike existing co-expression estimation methods, which cover $\sim 50\%$ of genes, our framework comprehensively integrates biological prior knowledge with data-driven information from co-expression networks and transcriptome-wide correlation patterns, yielding co-expression relationships for $32,016$ human genes. These heterogeneous sources are unified to generate a representative and universal gene embedding that demonstrates robust performance in fundamental gene- and cell-level benchmark tasks. To extend its use beyond static embeddings, we further developed a flexible generative framework that adapts CoxFormer embedding to complex multimodal scenarios across technologies and platforms. This enables context-aware generation across multiple biological dimensions, including gene identity, spatial resolution, and biological conditions. Here, we demonstrate these generative capabilities through advanced applications such as super-resolution enhancement, the prediction of expression and activity score for genes not included in targeted assays, and the detection of pathologically relevant tissue regions. Unlike existing methods with limited applicability, CoxFormer provides a versatile and comprehensive foundation for analyzing multimodal data at the gene, cell, and tissue level.

Results

An overview of CoxFormer

To overcome the limitations of current gene embeddings and spatial omics techniques, we systematically apply complementary data-driven information and biological knowledge sources to obtain universal CoxFormer embeddings for $32,016$ human genes (Fig. 1a). We obtain additional data-driven quantitative information from two sources: the COXPRESdb co-expression matrix from bulk microarray data on $18,858$ genes [1], and transcriptome-wide correlation information for $32,101$ genes computed based on over 10 million cells from 123 tissue projects in the Human Cell Atlas [21]. To collate biological information, we use web crawlers to obtain descriptions of $43,661$ genes from biological literature in the NCBI [22], GeneCards [23], and UniProt [24] websites. We then convert these descriptions into 3072-dimensional numerical vectors using text-to-embedding API from OpenAI. To integrate these two complementary sources, we construct a human transcriptome-wide graph, where each node represents a gene and each edge represents the relationship between genes. We assign the text-derived embeddings as node features, and single-cell correlations as the initial edge weights. The goal of the graph neural network learning is to reconstruct the bulk co-expression associations in the observed partial gene set and propagate these patterns to the full gene set. Finally, we encode the human whole transcriptome co-expression profiles from the estimated completed graph into 512-dimensional dense representations using an autoencoder framework, yielding the final CoxFormer embeddings.

The CoxFormer embeddings provide fundamental innovations from multiple key perspectives. Fundamentally, CoxFormer incorporates descriptive biological knowledge relevant to genes from comprehensive databases and is empowered by OpenAI API [25]; this avoids the need

for the extensive and laborious pretraining stage used for traditional foundation models and captures rich semantic information on gene functions. Building upon this semantic foundation, CoxFormer integrates information on quantitative gene relationships based on empirical expression data, including large-scale co-expression patterns from bulk tissues and single-cell atlases for millions of cells. By injecting these empirical quantitative insights into large language models, CoxFormer exploits the power of both general-purpose AI and domain-specific biological knowledge. Crucially, unlike existing gene embeddings, which focus on static applications, CoxFormer innovatively applies its generative ability to multimodal context-aware scenarios, thus addressing key limitations of current spatial transcriptomics technologies that rely on predefined gene panels or limited spatial resolutions. By capturing the universal relationships among genes, CoxFormer has the ability to generate across multiple biological dimensions. Generating along the gene identity dimension enables the prediction of genes not included in predefined panels and further completes whole-transcriptome inference from targeted assays; generating along the spatial dimension enables super-resolution accuracy in sequencing-based spatial transcriptomics platforms, while generating along pathobiological conditions enables the detection of diseased tissues. This generative capability is extensively applicable, from simple static gene/cell tasks to complex context-aware applications for multimodal datasets across technologies and platforms. CoxFormer's robustness and versatility facilitate downstream biologically interpretable tasks and advanced translational research.

We illustrate the benefits of CoxFormer through its applications in comprehensive gene-, cell-, and tissue-level tasks. At the gene level, CoxFormer shows strong and consistent performance in representative gene-property prediction tasks, including the prediction of transcription-factor (TF) attributes and promoter methylation states. At the cell level, it supports accurate cell-type annotation across multiple scRNA-seq datasets. At the tissue level, CoxFormer addresses several critical limitations of spatial omics technologies by incorporating a flexible Transformer-based spatial generative framework (Fig. 1b) that uniformly integrates CoxFormer embeddings with histology images and spatial coordinates to leverage multimodal information. This framework has a unique generative capability that enables four key applications, as demonstrated in Fig. 1c: (i) prediction of spatial transcriptomic profiles from histology; (ii) estimations of gene activity scores for chromatin accessibility; (iii) super-resolution imputation from spot-level to subcellular resolution; and (iv) detection of pathologically relevant tissue regions. We demonstrate these advanced applications for different technologies and platforms, including Visium, spatial ATAC-RNA-seq, and Xenium. We show that CoxFormer provides a comprehensive foundation for both fundamental single-cell omics analyses and multimodal spatial omics applications. The framework overcomes predefined gene panel constraints and low spatial resolution limitations, enabling whole-transcriptome inference across diverse biological contexts.

CoxFormer captures biologically meaningful manifolds

A biologically meaningful gene embedding is expected to organize genes into compact, well-separated functional modules while placing related programs in nearby regions of the manifold. Aligning with these principles, CoxFormer embeddings create a highly organized functional map when visualized through Uniform Manifold Approximation and Projection (UMAP) (Fig. 2a).

To validate this structure, we perform Gene Ontology (GO) enrichment analysis of the clustered genes. This reveals highly significant annotations that reflect the broad-ranging and well-known functional programs of human genes, including immune-proliferative programs, core cellular machinery, metabolic processes, and tissue-specialized functions. These programs are further illustrated by the enriched GO terms and representative genes (Fig. 2a and Supplementary Fig. S1a-d, S2a-c). For instance, the immune module is characterized by the enrichment of essential pathways such as *regulation of immune response* and *T cell activation*. The functionally and biologically meaningful nature of CoxFormer embeddings is further substantiated by the representative lineage genes, notably CD family members (e.g., *CD2*, *CD27*, *CD28*) and CCL chemokines (e.g., *CCL3*, *CCL4*, *CCL5*). Moreover, the manifold exhibits an informative geometric structure that reveals important biological insights. For instance, the immune and proliferation programs display clear co-localization; they occupy nearby regions and form a continuous immune–proliferation axis, rather than appearing at distant locations. Together, these results indicate that CoxFormer organizes genes into interpretable functional modules while preserving both the functional specificity and biological neighborhoods of related gene programs.

We further systematically evaluate CoxFormer in comparison to the representative gene embeddings learned from bulk transcriptomics (FRoGS [26], TCGA-Embedding [27], Gene2Vec [28]), single-cell transcriptomics (scGPT [13], GeneFormer [12]) and the biomedical literature (GenePT [17], BioVec [29], Mut2Vec [30]). We benchmark these embeddings on four gene-level binary classification tasks: identifying the dosage sensitivity of TFs, differentiating TFs by their regulatory range, and distinguishing bivalent promoters from non-methylated promoters and bivalent promoters from Lys4-only-methylated promoters (Fig. 2b). Performance is assessed using multiple complementary classification metrics and aggregated into a unified gene-task score (Methods). Overall, CoxFormer achieves the best gene-task score across the four tasks, outperforming embeddings that rely primarily on expression data as well as those derived mainly from biomedical text. This advantage is consistent with CoxFormer's design, which jointly leverages curated co-expression resources and single-cell correlation to capture transcriptome-wide gene–gene relationships and pairs them with gene-level textual descriptions that encode functional semantics. While scGPT-Human achieves a competitive aggregated score, it requires computationally intensive large-scale pretraining, whereas CoxFormer is trained efficiently on a single GPU. In each task, CoxFormer shows particularly strong performance in identifying TF dosage sensitivity and distinguishing bivalent from Lys4-only-methylated promoters. Detailed per-task results are provided in Supplementary Fig. S3a-b and S4a-b.

Transitioning from genetic functionality to cell biology, we evaluate all the methods using cell-level tasks in which cell embeddings are constructed as expression-weighted averages of gene embeddings [17]. We consider four scRNA-seq datasets: Diffuse Large B-cell Lymphoma FFPE (DLBL), Lung Cancer FFPE (LC), Breast Cancer FFPE (BC), and Human $CD34^+$ Bone Marrow (BM). Both unsupervised clustering and supervised cell-type classification are assessed using standard metrics and summarized into an aggregated cell-task score (Methods), where a higher value indicates better overall performance (Supplementary Fig. S5a-b and S6a-b). Notably, CoxFormer remains highly competitive in the cell-level tasks, while showing the best overall performance when aggregating gene- and cell-level results (Fig. 2b). We further visualize the learned CoxFormer cell embeddings with UMAP across all datasets (Fig. 2c). The UMAP

manifold shows distinct and clustered patterns of cell types; for example, BM cells trace a coherent hematopoietic continuum with clear separation of major lineages [31], whereas BC cells split cleanly into malignant versus immune/stromal compartments and the finer immune substructure is further resolved [32, 33].

In functional analysis and other applications, CoxFormer learns meaningful representations of human biology, facilitates downstream clustering and both gene- and cell-level classification tasks, and exhibits excellent and robust performance across diverse biological contexts.

CoxFormer accurately predicts the expression of unseen genes without external single-cell reference atlases

CoxFormer shows robust performance across fundamental gene- and cell-level tasks, with its learned embeddings capturing biologically meaningful representations and regulatory relationships. We therefore ask whether CoxFormer can move beyond static representation and serve as a generative prior in spatial assays, extending its applicability from a limited gene panel to the transcriptome-wide expression of unassayed genes. To that end, we develop a Transformer-based spatial generative framework to generate expression profiles for specified target genes. We integrate multimodal spatial context when available, including high-resolution histology images, coordinate-based positional encodings, and the observed expression (Methods). A key distinction from standard imputation pipelines is that our approach does not rely on external single-cell reference atlases; instead, it exploits the gene–gene relationships encoded in CoxFormer to generalize about unseen targets based on observed genes in a zero-shot manner. We apply this framework to spatial transcriptomics datasets and evaluate its ability to recover the spatial expression patterns of held-out genes.

To simulate the common scenario of running spatial assays with a limited gene panel, we train the model to infer unmeasured genes from a restricted set of measured transcripts from six human breast cancer (HBC) Visium datasets [34]. Specifically, the top 1,000 most highly variable genes (HVGs) are used to construct a pseudo-panel setting, where 900 HVGs are treated as observed inputs for model fitting and the remaining 100 HVGs are held out as unseen targets to evaluate prediction quality (Methods; Fig. 3a). We compare CoxFormer with representative scRNA-seq reference-based methods, including SpaOTsc [35], SpaGE [36], Tangram [37], gimVI [38], stPlus [39], novoSpaRc [40], Seurat [41], and LIGER [42], all provided with a matched reference data of $95,092$ cells. Without requiring this reference information, CoxFormer achieves the highest prediction accuracy across all datasets, as reflected by the averaged Pearson's correlation coefficient (Fig. 3b); similar conclusions can be drawn from the other metrics, as reported in Supplementary Fig. S7. We additionally implement ablation studies to investigate the contribution of histology images and spatial coordinates (Supplementary Fig. S7). Incorporating a comprehensive spatial context yields the best performance, indicating that the framework benefits from multimodal information. Importantly, when spatial context is removed and only gene embeddings are used, CoxFormer still outperforms all scRNA-seq reference-based baselines (Supplementary Fig. S7).

We further visualize the predicted expression of a representative held-out gene, *APOC1*, which is upregulated in breast cancer and closely linked to lipid metabolism and tumor-associated macrophage programs [43]. Across the six HBC datasets, some of the competing

methods collapse to show near-uniform low expression and fail to generate the spatial pattern (e.g., novoSpaRc and LIGER), whereas others show clear dataset-dependent failures. For example, gimVI produces a reasonable pattern for HBC4 but largely misses the signals in other datasets, while SpaGE captures the overall spatial pattern in several datasets but breaks down for HBC5. In contrast, CoxFormer consistently recovers the spatial pattern of *APOC1* expression across all six datasets, closely matching the ground truth (Fig. 3c).

CoxFormer extends accurate imputation to spatial epigenomics data

We further extend the application of CoxFormer to the spatial assay of transposase-accessible chromatin, highlighting its cross-omics applicability beyond transcriptomics. The task is to predict gene activity scores (GAS), a gene-level accessibility measure obtained by aggregating chromatin fragments within gene loci and weighting their distances to transcription start sites (Methods). We evaluate this setting in a human hippocampus spatial ATAC–RNA-seq dataset with paired spatial epigenomics and spatial transcriptomics measurements. To emulate a practical setting, where only a subset of genes is observed for model fitting, we construct a pseudo-panel over the top $1,000$ HVGs identified from the paired transcriptomics data. Specifically, 900 genes are used as observed inputs to train the predictor, and the remaining 100 genes are held out and used to assess the method's ability to predict GAS for unseen genes (Methods). As with the gene expression profile prediction setting, CoxFormer does not require an external reference to predict the GAS of unseen genes; instead, it leverages the inherent relationships among genes captured in gene embeddings.

To benchmark the GAS prediction performance, we compare CoxFormer with three ablated variants that isolate individual information sources, namely gene co-expression (the COXPRESdb-derived co-expression matrix), gene correlation (transcriptome-wide gene–gene correlations estimated from the Human Cell Atlas), and gene description (text-derived gene representations obtained via the OpenAI embedding API). Notably, we do not include scRNA-seq reference-based spatial methods (e.g., Tangram and SpaGE) because these approaches are solely designed for expression imputation and are not applicable to cross-omics GAS prediction. CoxFormer outperforms these variants across evaluation metrics computed between predicted and ground-truth GAS, as shown in the radar plot in Fig. 3d. This demonstrates that CoxFormer can successfully integrate data-driven and biological literature information from diverse sources.

We next visualize the spatial patterns of predicted and ground-truth GAS to assess whether the models capture biologically interpretable accessibility programs (Fig. 3e). Five representative genes (*SMIM27*, *BIN2*, *ASAH1*, *NPAS4*, and *SMDT1*) are selected to cover hippocampus-relevant neuronal programs and fundamental biological processes. The results show that CoxFormer reliably captures the ground-truth spatial GAS patterns, while the co-expression variant model only partially captures correct signals (Fig. 3e, Supplementary Fig. S8a).

We then focus on two well-established hippocampal layers, the dentate gyrus granule cell layer (DG GCL) and the CA3 pyramidal layer (CA3 Pyr) [44]. Using the top differentially expressed genes (DEGs) between these two regions, we visualize region-by-gene GAS heatmaps (Fig. 3f, Supplementary Fig. S8b). The ground truth shows a clear block structure, where many genes display smooth and gradual expression patterns between DG GCL and CA3

Pyr, reflecting coordinated layer-specific accessibility programs. CoxFormer largely preserves this block structure and the associated directional changes for the top DEGs, whereas the co-expression variant model recovers limited gene patterns. Overall, this spatial epigenomics benchmark shows that, beyond transcriptomics, CoxFormer embeddings support robust unseen-gene generalization for ATAC-derived GAS, demonstrating its cross-omics applicability to epigenomics.

CoxFormer uniquely enables expression imputation and spatial resolution enhancement for unassayed genes

We have shown that CoxFormer can use a limited gene panel to generalize on the transcriptome-wide expression of unseen genes, without relying on external scRNA-seq references. However, a major limitation of many spatial transcriptomics platforms is that expression is typically measured at the spot resolution, which masks within-spot heterogeneity and limits access to cellular-scale patterns. Coupling transcriptome-wide inference with enhanced spatial resolution would therefore yield both broader molecular coverage and finer spatial detail from such measurements. Using our Transformer-based spatial generative framework, we enable subcellular-scale prediction within each spot while predicting the expression profiles of genes absent from the panel (Methods; Fig. 4a).

To demonstrate this capability in a cross-sectional setting, we study two high-quality Xenium *in situ* human skin sections. Section A is profiled using the standard Human Skin Gene Expression Panel (378 genes), whereas Section B is measured with a custom panel limited to 280 genes. To create spot-level supervision for super-resolution, the Xenium cellular transcripts are aggregated into coarse pseudo-spots, each with a diameter of $55\mu m$. We further partition the 280 overlapping genes into three sets (Fig. 4b): (1) fully observed training genes ($n = 168$), which are expressed in both sections and used as model input; (2) partially observed test genes ($n = 56$), which are expressed in Section A but artificially masked in Section B and are used to benchmark conventional imputation; and (3) unobserved zero-shot genes ($n = 56$), which are artificially masked in both sections and used to assess zero-shot performance on unseen genes.

With this design, we quantify the cross-section imputation accuracy for the 56 partially observed test genes. Evaluation is performed at the spot level (Methods), and CoxFormer consistently outperforms iStar [45], a representative super-resolution baseline, across all metrics (Fig. 4c). At an enhanced resolution, CoxFormer produces smoother and more spatially coherent gene maps, whereas iStar often yields fragmented patterns (Supplementary Fig. S9). To assess whether these super-resolved predictions preserve tissue structure, clustering is performed on the predicted expression in Section B (Fig. 4d). CoxFormer recovers the ground-truth structures and clearly delineates the epidermis, immune infiltration sites, and melanoma tumor regions (Fig. 4d-e), whereas the results from iStar show evident mixing of the epidermal and immune-infiltrated domains.

The same framework is further evaluated in a challenging zero-shot setting, where target genes are entirely absent from the measured panel. Using the 56 genes unobserved in both sections, CoxFormer produces subcellular-scale, spatially structured predictions for marker genes that delineate distinct tissue regions (Fig. 4f) [46]. For example, *CLDN1*, an epithelial marker, shows elevated expression localized to the epidermal region, consistent with the region

boundaries in Fig. 4d and the Xenium ground truth. Likewise, *CD3D*, a T-cell marker, is selectively enriched in the immune-infiltrated area, matching the ground-truth pattern. Beyond these regional markers, CoxFormer also captures intricate extracellular-matrix–associated programs. The expression patterns of *LUM* and *COL6A3* exhibit pronounced ring-like stromal structures that align with the Xenium measurements, demonstrating that the model can recover the delicate spatial architectures of completely unassayed genes. Multi-scale visualizations across four resolutions (4×, 16×, 64×, and 256×) are provided in Supplementary Fig. S10 and S11, and full-section predictions are shown in Supplementary Fig. S12.

Finally, using Section B, we investigate whether transcriptome-wide subcellular imputation with enhanced spatial resolution can make region-resolved cell–cell communication (CCC) analysis more informative (Methods). In an analysis of the original 280-gene panel, CCC inference is severely underpowered and yields only four robust ligand–receptor pairs (Supplementary Fig. S13a-b). Upon expanding the 280 measured genes of Section B to 32,018 genes with CoxFormer, the inferred CCC landscape becomes substantially richer (Fig. 4g), revealing additional plausible region-associated signaling modules that remain undetected in the limited panel. For example, extracellular-matrix and adhesion signaling (e.g., *COL1A1*→*SDC1*) become prominent, consistent with matrix-rich microenvironments [47]. Epidermal junction programs are recovered with strong epidermis-associated specificity (e.g., *DSG1*→*DSC3*) [48]. In parallel, immune-recognition signals emerge in the immune-infiltrated compartment (e.g., *SFTPD*→*SIRPA*) [49]. These interactions also exhibit spatial patterns consistent with the region annotations in Fig. 4d (further visualized in Supplementary Fig. S14), supporting more interpretable and region-resolved CCC discovery beyond the scope of the limited gene panel.

CoxFormer identifies pathological regions without reference

In addition to enabling whole-transcriptome prediction and super-resolution, CoxFormer embeddings serve as an intrinsic "healthy reference", as illustrated by their use in the delineation of pathological regions using an unsupervised, reference-free strategy. There are three steps to the strategy (Fig. 5a). First, we train the Transformer-based spatial generative framework on only spatially stable housekeeping genes, allowing it to learn mapping from gene to gene expression in the "healthy" distribution. Second, we generate transcriptome-wide expression profiles with the trained model as a "pseudo-healthy" reference. Third, we quantify the residuals between the real and generated gene expression of non-housekeeping genes, then apply clustering to automatically segment the tissue into normal and abnormal regions (Methods). This framework is mainly based on a hypothesis that the divergence between the real pathological expression and the generated healthy baseline would reveal disease-specific alterations.

We benchmark the pathological annotation task using four human colorectal cancer liver metastasis (CRLM) sections with annotated abnormal regions, with colorectal cancer (CRC) and liver metastasis (LM) tissues for both untreated and therapy-perturbed (PR) conditions (Fig. 5b). We compare CoxFormer with representative reference-required imputation methods (SpaGE, SpaOTsc, Tangram, gimVI, Seurat, stPlus, novoSpaRc, and LIGER). Across all evaluation criteria, CoxFormer achieves the most accurate abnormal-region annotations (Fig. 5c), with per-section performance detailed in Supplementary Fig. S15a–S18a. Importantly, even without leveraging external scRNA-seq references, CoxFormer still yields an area under the

receiver operating characteristic curve (AUROC) of 0.91 averaged across four sections (Fig. 5c, left panel) and a recall score over 0.98 (Fig. 5c, right panel). As a representative example, we focus on the untreated CRC section, with the results for the other sections shown in the Supplementary Fig. S15-S18. CoxFormer's quantitative advantage is visually demonstrated in the segmentation results in the spatial maps (Fig. 5d, Supplementary Fig. S15b-S18b) and the corresponding UMAP visualizations with tumor segmentation (Supplementary Fig. S15c–S18c), where it detects abnormal regions precisely delineating the tumor boundaries, whereas the competing methods produce noisy or incorrect detections.

We further demonstrate CoxFormer's power to detect abnormal regions via its ability to: (1) reconstruct the housekeeping gene profiles, and (2) differentiate between the real pathological expression and the generated "pseudo-healthy" pattern. To validate the framework's reconstruction capability, we compare the true and reconstructed gene expression heatmaps (Fig. 5e, Supplementary Fig. S15d-S18d). There are 10 representative housekeeping genes, including *PRCC* and *PSMD6* [50]. CoxFormer reconstructs these gene profiles in close alignment with the ground truth, demonstrating that it can provide a stable reference against pathological alterations. Next, we contrast the measured expression with the "pseudo-healthy" reference for CRC-associated markers (Fig. 5f). Consistent with prior reports that *GUCA2A* and *CEACAM7* are downregulated in colorectal tumors due to dedifferentiation [51], CoxFormer reveals their suppressed expression within the abnormal compartment relative to the pseudo-healthy baseline. This finding suggests that the housekeeping-conditioned model captures gene–gene dependencies and reveals the systematic deviations accompanying malignant transformation.

Finally, we visualize the expression patterns of the top 200 DEGs between the normal and abnormal regions segmented by CoxFormer across all four sections (Supplementary Fig. S19a-b and S20a-b). We then perform GO biological process enrichment using these DEGs, as shown for the untreated CRC section in Fig. 5g. The enriched terms are primarily related to extracellular remodeling and motility-related programs, including *extracellular matrix organization*, *extracellular structure organization*, and *regulation of cell migration/positive regulation of cell motility*. These functions are consistent with stromal reorganization and invasion-associated processes in colorectal cancer liver metastasis [47], evidence that the CoxFormer-identified abnormal domains capture biologically meaningful pathological tissue states.

Discussion

As a result of our work, we propose the use of CoxFormer as a universal gene embedding framework that comprehensively integrates biological knowledge and data-driven co-expression relationships to enable whole-transcriptome inference across diverse contexts and technologies. Unlike existing methods that either require intensive pretraining on massive single-cell corpora or rely solely on general-purpose language models, CoxFormer systematically combines qualitative biological descriptions from literature databases and quantitative co-expression patterns from both bulk tissues and large-scale single-cell atlases. Moreover, CoxFormer employs a graph neural network to reconstruct and complete transcriptome-wide co-expression networks, then utilizes autoencoder architecture to compress these relationships into low-dimensional gene embeddings. The whole framework generates robust, reusable, and universal CoxFormer embeddings that capture the intrinsic complex relationships among genes. Finally, we have

designed a flexible generative framework for applying CoxFormer embeddings to diverse downstream applications, ranging from gene-level to complex spatial inference tasks.

Conventionally, predicting the expression of unmeasured genes in spatial omics data requires external reference datasets, which are often unavailable or difficult to obtain. CoxFormer successfully enables the accurate prediction of spatial gene expression, gene activity scores, and super-resolution enhancement without requiring external references. This ability is due to its exploitation of the universal nature of CoxFormer embeddings and our flexible generative framework, which can be adapted to diverse technologies and platforms.

Through the comprehensive applications of CoxFormer to diverse datasets and downstream tasks, we demonstrate that the framework addresses several critical limitations of multiple spatial omics technologies. Via its spatial pattern prediction of unmeasured genes, CoxFormer accurately predicts the expression profiles in Visium spatial transcriptomics of human breast cancer data with the highest prediction accuracy outperforming competing methods, which typically require matched external reference datasets. Beyond the generative ability for transcriptomics, CoxFormer also applies to epigenomics by generating gene activity scores for chromatin accessibility in spatial ATAC-RNA-seq data from the human hippocampus. In a super-resolution enhancement task, CoxFormer enhances the detection accuracy from spot-level expression to a subcellular resolution, as validated on Xenium platforms. Transitioning from spatial prediction to whole-transcriptome generation, it further extends the targeted assays from 280 to 32,018 genes, identifying CCC programs undetectable by the original panel, such as extracellular-matrix and adhesion signaling (e.g., *COL1A1*→*SDC1*). Lastly, CoxFormer shows strong potential for pathological detection and translational research, as demonstrated by its precise tumor region detection in colorectal cancer liver metastasis samples across different treatment conditions. These comprehensive and vital applications highlight CoxFormer's versatility and robustness in overcoming predefined gene panel constraints, enabling whole-transcriptome inference, and further facilitating translational research.

Although CoxFormer provides a strong foundation for the rapidly evolving landscape of spatial omics technologies, several limitations remain. First, while CoxFormer already integrates spatial coordinates and histology images, future iterations could include more sophisticated spatial modeling to account for tissue context information; for example, an advanced model should be able to characterize complex organs and developmental processes when handling 3D spatial profiling technologies [52–54]. Second, the flexible generative framework of CoxFormer could be extended to the integration of additional modalities beyond transcriptomics and chromatin accessibility, including proteomics [55–57] and metabolomics data [58, 59]. Finally, CoxFormer currently leverages static co-expression networks, but incorporating dynamic context-specific gene relationships [60, 61] has the potential to further enhance the performance of downstream tasks for certain tissues or disease states.

Methods

Construction of CoxFormer embeddings

Data curation and multimodal gene representation.

To construct a unified representation of $32,016$ human genes, we utilize multimodal data from three complementary sources. First, we obtain high-confidence co-expression profiles from the COXPRESdb database (*Microarray_m.c7* version) to set up reliable gene-gene relationships. Specifically, we aggregate normalized bulk microarray data for 18,858 genes, resulting in a dense co-expression matrix $\mathbf{C}_{cox}$ $(18,858 \times 18,858)$.

Next, we integrate single-cell RNA sequencing (scRNA-seq) data from the Human Cell Atlas (HCA) to extend gene coverage beyond microarray limitations. This dataset contains 123 healthy tissue projects comprising 10.4×10^6 cells sequenced using 10x Genomics $3'$ V2 chemistry. Following standard quality control and project-specific Pearson correlation analysis, we construct a global correlation matrix $\mathbf{C}_{cor}$ $(32,101 \times 32,101)$ using a cell-count weighted aggregation approach. For each gene pair (i, j):

$$\left(\mathbf{C}_{cor}\right)_{ij} = \frac{\sum_k n_k \left(\mathbf{C}_{cor}^{(k)}\right)_{ij} \mathbb{I}_{ij}^{(k)}}{\sum_k n_k \, \mathbb{I}_{ij}^{(k)}}, \tag{1}$$

where n_k denotes the cell count of project k, and $\mathbb{I}_{ij}^{(k)} = 1$ if $\left(\mathbf{C}_{cor}^{(k)}\right)_{ij}$ is defined and 0 otherwise.

To incorporate functional semantics beyond transcriptional patterns, we also collect $82,718$ textual descriptions from the NCBI Gene, GeneCards, and UniProt databases. We group these descriptions by gene and encoded them using the OpenAI `text-embedding-3-large` model, which creates dense $3,072$-dimensional embedding vectors for all $43,661$ genes. We then define the target gene universe as the overlap between the text-annotated genes and the genes in $\mathbf{C}_{cor}$, resulting in $32,016$ genes for the construction of a unified representation.

Together, these multimodal sources provide complementary statistical, structural, and semantic views of gene relationships, forming the foundation of the CoxFormer embeddings.

Transcriptome-wide co-expression graph completion

We formulate transcriptome-wide co-expression estimation as a graph completion problem that integrates high-confidence bulk co-expression with transcriptome-wide but noisy scRNA-seq covariation. Let $\mathcal{G}_{\text{obs}} = (\mathcal{V}, \mathcal{E}_{\text{obs}})$ denotes the partially observed co-expression graph, where $\mathcal{E}_{\text{obs}}$ contains gene pairs with measured co-expression values $c_{i,j}$. Our objective is to learn a parametric model that predicts co-expression for any gene pair in the gene set $\mathcal{V}$, thereby extending the reliable patterns supported by curated edges to previously unmeasured pairs.

As a global structural prior, we construct a weighted correlation graph from the scRNA-seq correlation matrix $\mathbf{C}_{\text{cor}}$ by assigning each gene pair (i, j) an edge weight $w_{i,j} = (\mathbf{C}_{\text{cor}})_{i,j}$. Importantly, because transcriptome-wide correlations are substantially noisier than curated bulk co-expression measurements, we use $\mathbf{C}_{\text{cor}}$ to guide information propagation and neighborhood construction, rather than treating correlations as direct co-expression estimates. Considering that a fully connected graph over $|\mathcal{V}| > 30{,}000$ genes is computationally prohibitive, we sparsify

the correlation graph by retaining the top-K most correlated neighbors for each gene, which yields a scalable dependency structure while preserving local covariation neighborhoods. On the resulting sparse graph, we perform L layers of message passing with mean aggregation [62]. We initialize each gene node with its text-derived embedding as the node representations, $\mathbf{h}_i^{(0)} = \mathbf{z}_i$. At each layer $l \in \{1, 2, \ldots, L\}$, we aggregate messages from the top-K neighborhood $\mathcal{N}_K(i)$ as:

$$\mathbf{m}_i^{(l)} = \frac{1}{K} \sum_{j \in \mathcal{N}_K(i)} \mathbf{h}_j^{(l)}, \tag{2}$$

and update node representations by:

$$\mathbf{h}_i^{(l+1)} = \sigma\Big(\mathbf{W}^{(l)} \big[\text{concat}(\mathbf{h}_i^{(l)}, \mathbf{m}_i^{(l)})\big]\Big). \tag{3}$$

After L layers, the refined representation $\mathbf{h}_i = \mathbf{h}_i^{(L)}$ encodes gene-level functional semantics while being regularized by the correlation-defined neighborhood context, enabling co-expression to be predicted from paired representations (Supplementary Fig. S21a). Specifically, for a gene pair (i, j), we predict co-expression using an Multi-Layer Perceptron (MLP) edge predictor:

$$\hat{c}_{i,j} = f_{\text{cox}}\Big(\text{concat}(\mathbf{h}_i, \mathbf{h}_j)\Big). \tag{4}$$

We train the model by minimizing the mean squared error loss $\mathcal{L}_{\text{cox}}$ over observed co-expression edges $(i, j) \in \mathcal{E}_{\text{obs}}$ between measured values $c_{i,j}$ and predictions $\hat{c}_{i,j}$:

$$\mathcal{L}_{\text{cox}} = \frac{1}{|\mathcal{E}_{\text{obs}}|} \sum_{(i,j) \in \mathcal{E}_{\text{obs}}} \left(c_{i,j} - \hat{c}_{i,j}\right)^2. \tag{5}$$

After training, we apply the learned predictor to all gene pairs to obtain a completed co-expression matrix $\mathbf{C}_{\text{cox}}$ with entries $(\mathbf{C}_{\text{cox}})_{i,j} = \hat{c}_{i,j}$. This provides transcriptome-wide co-expression estimates consistent with curated bulk measurements yet informed by global scRNA-seq covariation (Supplementary Fig. S21b). Further details on the implementation are provided in the Supplementary Notes.

Latent space compression via Transformer-based autoencoding.

Although each gene can be represented by a $32,016$-dimensional co-expression vector, using these vectors directly is computationally expensive in downstream applications. To that end, we reduce the dimensionality of these co-expression vectors using a deep autoencoder. We implement both the encoder f_{enc} and the decoder f_{dec} using stacked Transformer blocks in a symmetric architecture [63]. Attention mechanism is applied throughout encoding and reconstruction, allowing the model to effectively capture biological dependencies and long-range interactions inherent in the gene regulatory network.

Formally, we specify the training objective as follows. For each gene i, we use the corresponding row vector $(\mathbf{C}_{cox})_{i,:} \in \mathbb{R}^{32,016}$ as the input co-expression feature vector. The encoder f_{enc} maps this high-dimensional vector to a 512-dimensional latent vector $\mathbf{e}_i$, and the decoder f_{dec} reconstructs the original input from $\mathbf{e}_i$:

$$\mathbf{e}_i = f_{\text{enc}}\big((\mathbf{C}_{cox})_{i,:}\big) \in \mathbb{R}^{512}, \quad \big(\widehat{\mathbf{C}}_{cox}\big)_{i,:} = f_{\text{dec}}(\mathbf{e}_i) \in \mathbb{R}^{32,016}. \tag{6}$$

We train the model by minimizing the mean squared reconstruction loss $\mathcal{L}_{\mathrm{AE}}$:

$$\mathcal{L}_{\mathrm{AE}} = \sum_i \|(\mathbf{C}_{cox})_{i,:} - (\widehat{\mathbf{C}}_{cox})_{i,:}\|_2^2, \tag{7}$$

which encourages the latent vector to preserve the co-expression structure encoded in $\mathbf{C}_{cox}$. We therefore use the encoder outputs $\{\mathbf{e}_i\}$ as the CoxFormer gene embeddings. Further details on the implementation are provided in the Supplementary Notes.

A multi-modal generative framework for predicting spatial epigenomics and transcriptomics using CoxFormer embeddings

To enhance the applicability of CoxFormer embeddings to advanced genomic analyses, we develop a multi-modal generative framework that integrates gene embeddings with available spatial cues, such as H&E histology and spot coordinates, to predict epigenomic and transcriptomic profiles. The framework can fuse gene embeddings with coordinate encodings and image-derived features when available, or operate on gene embeddings alone, enabling seamless use across sequencing technologies and heterogeneous data formats.

Multi-modal spatial information extraction and fusion

Spatial transcriptomics datasets usually provided two types of spatial information: spot coordinates and, in some assays, a matched histology image (e.g., H&E in 10x Genomics Visium). When histology is available, we extract morphology from image patches using a Vision Transformer (ViT), yielding patch features of shape $n_{\mathrm{patch}} \times d_{\mathrm{ViT}}$ [64]. To obtain spot-level image features, we average the patch features within each spot and form an image feature $\mathbf{X}_{\mathrm{img}} \in \mathbb{R}^{n_{\mathrm{spot}} \times d_{\mathrm{ViT}}}$, which is then projected by a two-layer MLP with ReLU activations to $\mathbb{R}^{n_{\mathrm{spot}} \times d_{\mathrm{img}}}$.

For spot coordinates, we encode the 2D locations $\mathbf{X}_{\mathrm{loc}} = [x, y] \in \mathbb{R}^{n_{\mathrm{spot}} \times 2}$ with a Fourier feature encoder. We use exponentially spaced frequencies $\omega_f = 2^{f-1}\pi$ for $f = 1, \ldots, F$ and compute sinusoidal features for each coordinate dimension [65]. For a spot i with $\mathbf{X}_{\mathrm{loc},i} = (x_i, y_i)$, the positional encoding is defined as:

$$\begin{aligned} \mathrm{PE}(\mathbf{X}_{\mathrm{loc},i}) = \big[\ & \sin(\omega_1 x_i), \cos(\omega_1 x_i), \sin(\omega_1 y_i), \cos(\omega_1 y_i), \ldots, \\ & \sin(\omega_F x_i), \cos(\omega_F x_i), \sin(\omega_F y_i), \cos(\omega_F y_i)\ \big]. \end{aligned} \tag{8}$$

Stacking all spots yielded $\mathrm{PE}(\mathbf{X}_{\mathrm{loc}}) \in \mathbb{R}^{n_{\mathrm{spot}} \times 4F}$. We then project them to $\mathbb{R}^{n_{\mathrm{spot}} \times d_{\mathrm{loc}}}$ using a two-layer MLP with ReLU activations.

When neither coordinates nor histology are provided, we introduce learnable latent vectors $\mathbf{X}_{\mathrm{learn}} \in \mathbb{R}^{n_{\mathrm{spot}} \times d_{\mathrm{learn}}}$ as the spatial context, which are initialized from a standard normal distribution and optimized during training so that they can be learned directly from the data, allowing the capture of dataset-specific spatial structure.

When both coordinates and histology are present, we construct a unified spatial context $\mathbf{X}_{\mathrm{spa}} \in \mathbb{R}^{n_{\mathrm{spot}} \times d_{\mathrm{model}}}$ by concatenating the image, coordinate, and learnable embeddings after projecting them onto dimensions that are summed to d_{model}. When only one spatial input is present, we project it to d_{model} dimensions and used it as the spatial context.

Integrating spatial context with CoxFormer embeddings for transcriptome-wide prediction

We formulate spatial prediction as gene-conditioned regression over the spatial context defined above. Because spatially proximal spots tend to exhibit correlated molecular states, we treat the spot-wise spatial context as a sequence of n_{spot} tokens and use an attention-based aggregation mechanism to capture dependencies across the entire tissue section.

To allow each spot to adaptively pool information from other spots, we first compute three learned linear projections of $\mathbf{X}_{\text{spa}}$: a *query* representation that specifies what a spot is looking for, a *key* representation that specifies what information a spot can offer, and a *value* representation that carries the content to be aggregated. Concretely,

$$\tilde{\mathbf{X}}_{\text{spa}} = \text{softmax}\left(\frac{(\mathbf{X}_{\text{spa}}\mathbf{W}_{query})(\mathbf{X}_{\text{spa}}\mathbf{W}_{key})^\top}{\sqrt{d_{key}}}\right)(\mathbf{X}_{\text{spa}}\mathbf{W}_{value}), \tag{9}$$

where $\mathbf{W}_{query}, \mathbf{W}_{key} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ and $\mathbf{W}_{value} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are learnable projection matrices, and softmax is applied row-wise to normalize attention weights over spots. This operation allows each spot to form a weighted sum of the value vectors from all spots, producing a context-aware spatial representation.

We then incorporate gene-specific information using the CoxFormer embedding of the target gene $\mathbf{e} \in \mathbb{R}^{512}$. Intuitively, we use $\mathbf{e}$ to modulate a shared spatial context so that the resulting spot-wise representations become specific to the queried gene and can be directly read out as its spatial expression profile. Concretely, we first project $\mathbf{e}$ into the model space via a learnable linear layer $\mathbf{W}_g$ to match the dimensionality of $\tilde{\mathbf{X}}_{\text{spa}}$. We then broadcast the projected gene vector to all spots, add it to the context-aware spatial representation as conditioning, and finally apply an MLP head to output one scalar per spot:

$$\tilde{\mathbf{e}} = \mathbf{e}\mathbf{W}_g + \mathbf{b}_g, \qquad \hat{\mathbf{y}} = f_{\text{mlp}}(\tilde{\mathbf{X}}_{\text{spa}} + \mathbf{1} \otimes \tilde{\mathbf{e}}). \tag{10}$$

Here $\mathbf{W}_g \in \mathbb{R}^{512 \times d_{\text{model}}}$ and $\mathbf{b}_g \in \mathbb{R}^{d_{\text{model}}}$ are learnable parameters, and $\mathbf{1} \in \mathbb{R}^{n_{\text{spot}}}$ is an all-ones vector, and $\otimes$ denotes the Kronecker product that broadcasts the projected gene embedding $\tilde{\mathbf{e}} \in \mathbb{R}^{d_{\text{model}}}$ to all spots. The $f_{\text{mlp}} : \mathbb{R}^{n_{\text{spot}} \times d_{\text{model}}} \to \mathbb{R}^{n_{\text{spot}} \times 1}$ outputs the predicted expression (or activity score) $\hat{\mathbf{y}} \in \mathbb{R}^{n_{\text{spot}}}$ for gene g across all spots. Further details on the implementation are provided in the Supplementary Notes.

Notably, our framework follows the spirit of the Transformer decoder block design [63], while adapting it to spatial molecular prediction. Specifically, we predict the spatial profile of a specified target gene across all spots in a single pass, rather than applying the autoregressive next-token prediction that is a common use of Transformer decoders. Training therefore does not require causal masking. Moreover, in the absence of CoxFormer embeddings, a Transformer-style transcriptome-wide formulation would typically represent genes as additional tokens, which naively leads to a gene axis on the order of 3×10^4 for the human genome and substantially increases the computational cost. By using the CoxFormer embedding as an external conditioning input, our framework preserves the Transformer's ability to model long-range spot-to-spot dependencies, while avoiding the need to present the whole genome as gene tokens.

Sparsity-aware objective and model optimization

To handle the sparsity and zero inflation of spatially resolved molecular profiles, we train the model with a loss that combines a weighted Huber term with an ℓ_1 penalty on the predictions. The weighting reduces the contribution of zero entries, the Huber term limites the influence of large residuals, and the ℓ_1 penalty helps suppress low-level positive background predictions. For the weighted Huber term, we first partition entries into positive sets ($\Omega_{nz} = \{i : y_i > 0\}$) and zero sets ($\Omega_z = \{i : y_i = 0\}$). And then we define the weighted Huber term as follows:

$$\mathcal{L}_{\text{Huber}-\text{w}} = \lambda_{nz} \frac{1}{|\Omega_{nz}|} \sum_{i \in \Omega_{nz}} \mathcal{L}_{\text{Huber}}(y_i, \hat{y}_i) + \lambda_z \frac{1}{|\Omega_z|} \sum_{i \in \Omega_z} \mathcal{L}_{\text{Huber}}(y_i, \hat{y}_i), \tag{11}$$

where λ_{nz} and λ_z control the relative contribution of positive and zero entries (default $\lambda_{nz} = 0.5$ and $\lambda_z = 0.5$). Specifically, the Huber term is:

$$\mathcal{L}_{\text{Huber}}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & |y - \hat{y}| \le \delta, \\ \delta |y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise}, \end{cases} \tag{12}$$

with an adaptive threshold δ computed per mini-batch as the standard deviation of residuals $(\hat{\mathbf{y}} - \mathbf{y})$ across spots for each sample. We finally optimize:

$$\mathcal{L} = \mathcal{L}_{\text{Huber}-\text{w}} + \lambda_{\ell_1} \|\hat{\mathbf{y}}\|_1, \tag{13}$$

where ℓ_1 penalty regularization λ_{ℓ_1}. Further details on the implementation are provided in the Supplementary Notes.

Downstream analysis

Gene property prediction

To assess whether the CoxFormer embeddings are able to capture intrinsic regulatory semantics, we benchmark their performance on four distinct binary classification tasks that discriminate between distinct gene properties [12]: (1) dosage-sensitive versus dosage-insensitive transcription factors (TFs); (2) long-range versus short-range transcription TFs; (3) bivalent versus non-methylated promoters and (4) bivalent versus Lys4-only-methylated promoters.

For each task, we train supervised models to predict gene properties from the gene embedding z_i. To ensure that our evaluation is robust to the choice of downstream model, we train two representative classifiers: Logistic Regression (LR), and Support Vector Machine (SVM). We then evaluate the generalization performance using five-fold stratified cross-validation, and report the mean performance across folds using the metrics, including AUROC, Accuracy, F1-score, Precision, and Recall.

Cell-types annotation

To evaluate the utility of CoxFormer embeddings in cell-level tasks, we derive a cell representation by aggregating gene embeddings according to each cell's normalized expression profile, then use these cell embeddings for cell-type clustering and classification across tissues. Given a gene

expression matrix $\mathbf{S} \in \mathbb{R}^{n_{\text{cell}} \times n_{\text{gene}}}$ for n_{cell} cells and n_{gene} genes, we formed the cell embedding $\mathbf{u}_i \in \mathbb{R}^{d_{\text{gene}}}$ for cell i by taking a normalized weighted sum:

$$\mathbf{u}_i = \frac{\sum_{g=1}^{n_{\text{gene}}} \mathbf{S}_{i,g}\, \mathbf{e}_g}{\sum_{g=1}^{n_{\text{gene}}} \mathbf{S}_{i,g} + \epsilon}, \tag{14}$$

where $\mathbf{S}_{i,g}$ denotes the expression of gene g in cell i, $\mathbf{e}_g \in \mathbb{R}^{d_{\text{gene}}}$ is the embedding of gene g, and $\epsilon = 10^{-12}$ is a small constant for numerical stability. Building on the cell embeddings defined above, we treat the provided annotations as reference labels. For an unsupervised evaluation, we cluster the cell embeddings with K-means and quantify the clustering quality using the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Adjusted Mutual Information (AMI). In a supervised evaluation, we employ a kNN classifier with stratified five-fold cross-validation and assessed its performance using AUROC, Accuracy, F1-score, Precision, and Recall.

Super-resolution gene expression prediction

Building on the prediction framework above, we extend the histology input from spot-aggregated image features to pixel-level image features for super-resolution inference. Within the spot-level setting, ViT patch features are averaged within each spot to obtain spot-level image features, which support spot-level prediction. While for super-resolution prediction, we avoid this within-spot averaging and instead predict gene expression directly from pixel features. In the implementation, we treat the n_{pix} pixel within a spot as a sequence and form a pixel-feature matrix $\mathbf{X}_{\text{pix}} \in \mathbb{R}^{n_{\text{pix}} \times d_{\text{ViT}}}$.

We then adapt the spot-level prediction framework at patch resolution. Specifically, each patch feature is mapped to the model dimension with a two-layer MLP with ReLU activations, yielding $\mathbf{X}_{\text{pix}} \in \mathbb{R}^{n_{\text{pix}} \times d_{\text{model}}}$. We then use self-attention over the n_{pix} patches within a spot so that each patch can aggregate information from the other patches in the same spot. The target gene is identified through its CoxFormer embedding, and the model obtains pixel-level predictions $\hat{y}^g_{i,k}$ for gene g at pixel k in spot i.

Considering that most spatial transcriptomics data encode gene expression at the spot resolution, we train the model under weak supervision by aggregating patch-level predictions within each spot:

$$\hat{Y}_{i,g} = \sum_{k=1}^{K} \hat{y}^k_{i,g}. \tag{15}$$

We then minimize the sparsity-aware objective defined above between the aggregated prediction $\hat{Y}^g_i$ and the measured spot-level expression Y^g_i. Training uses the same optimization setting as in the spot-level task, except that we use a batch size of 256.

Cell-cell communication analysis

We perform region-resolved cell-cell communication (CCC) analysis using the *LIANA* (version 1.6.1) library in Python [66], grouping cells into tissue regions (source-target groups) for interaction inference. We first identify candidate ligand-receptor interactions using `rank_aggregate()` with the consensus resource (`resource_name="consensus"`) under standard quality control

settings (`expr_prop=0.1`). We incorporate spatial context by constructing a Gaussian-kernel neighbor graph (`bandwidth=100`, `cutoff=0.1`) from cell centroid coordinates, and refine interactions using `bivariate()` with cosine similarity (local) and Moran's I (global). Significance is assessed via permutation testing (`n_perms=100`) with an additional sparsity filter (`nz_prop=0.2`).

Differential expression analysis

We identify differentially expressed genes (DEGs) between the target and reference expression profiles using the *Scanpy* (version 1.9.8) library in Python [67]. For each gene, we compare expression between the two groups using the Wilcoxon rank-sum test and adjusted P values with the Benjamini–Hochberg procedure. DEGs are defined as genes with an adjusted P value < 0.01 and an absolute log-fold change greater than 3.

GO enrichment analysis

We conduct GO enrichment analysis for each gene set selected from spatial transcriptomics predictions based on CoxFormer embeddings using the *GSEApy* (version 1.1.8) library in Python with the Enrichr database `GO_Biological_Process_2021` [68]. For each gene set, enrichment significance is assessed with Fisher's exact test and controlled for multiple comparisons using the Benjamini–Hochberg procedure. We then rank GO terms by their adjusted P values (`cutoff=0.05`), report significance as $-\log_{10}(\text{Adj. } P)$ and visualize the top terms within the Biological Process ontology to summarize the dominant functional signals.

Abnormal region detection

We develop an unsupervised method to delineate abnormal tissue regions. Specifically, we use our spatial transcriptomics prediction framework based on CoxFormer embeddings to infer transcriptome-wide expression from housekeeping genes, and consider these predictions as a pseudo-healthy reference. Accordingly, deviations between the predicted and measured expression are treated as signals of pathological alteration. In the implementation, we first select the top 200 disease-relevant genes using the differential expression analysis. For this gene set, we define the spot-wise residual as the difference between the predicted and measured expression:

$$R_{i,g} = Y_{i,g} - \hat{Y}_{i,g}, \tag{16}$$

where $Y_{i,g}$ and $\hat{Y}_{i,g}$ denote the predicted and measured expression of gene g at spot i, respectively. To delineate abnormal regions, we perform Principal Component Analysis (PCA) on the residual matrix R and apply Leiden clustering, adjusting the resolution to obtain two clusters. For each cluster c, we compute its mean absolute residual:

$$\bar{r}_c = \frac{1}{|c|\,|G|} \sum_{i \in c} \sum_{g \in G} |R_{i,g}|\,. \tag{17}$$

We label the cluster with the larger $\bar{r}_c$ as the abnormal region and evaluated detection performance against ground-truth pathological annotations using AUROC, Accuracy, F1 score, Precision, and Recall.

Comparisons of methods

We perform extensive comparisons between the performance of CoxFormer and existing methods in both simulated and real data analyses. In brief, for gene- and cell-level tasks, we compare CoxFormer embeddings with alternative gene embedding baselines. For spatial transcriptomics prediction and pathological region segmentation, we benchmark CoxFormer against representative scRNA-seq reference-based imputation methods. For the spatial epigenomics prediction task, we compare it against three single-source embedding variants using our spatial generative framework. For super-resolution enhancement, we compare CoxFormer with iStar.

Compared methods for gene- and cell-level tasks

We compare CoxFormer with gene embeddings learned using different information sources, including bulk transcriptomics, single-cell transcriptomics, and the biomedical literature, as provided at `https://zenodo.org/records/16764517`.

(1) scGPT [13]: Transformer-based gene embeddings are trained on single-cell gene expression, where scGPT-Human is trained on human single-cell transcriptomics and scGPT-PanCancer is trained on pan-cancer single-cell transcriptomics (dim = 512).

(2) GeneFormer [12]: Transformer-based gene embeddings are pretrained on single-cell gene expression data, including GeneFormer-6L30M (dim = 256), GeneFormer-12L30M (dim = 512), GeneFormer-12L95M (dim = 512), GeneFormer-20L95M (dim = 896), and the cancer-tuned GeneFormer-12L95M-Cancer (dim = 512). Here, "6L/12L/20L" indicates the number of Transformer layers and "30M/95M" indicates the pretraining scale in millions.

(3) GenePT [17]: text-derived gene embeddings are obtained by encoding NCBI Gene Summary with OpenAI embedding APIs, where GenePT-Ada uses `text-embedding-ada-002` (dim = 1536) and GenePT-Model3 uses `text-embedding-3-large` (dim = 3072).

(4) BioVec [29]: literature-derived gene embeddings are trained on biomedical corpora, including BioVec-CBOW (trained with the CBOW), BioVec-SkipGram (trained with the skip-gram), BioVec-FastText (trained with FastText), and BioVec-GloVe (trained with GloVe) (dim = 100).

(5) Mut2Vec [30]: multi-source gene embeddings integrate mutation profiles, the biomedical literature, and protein-protein interaction information (dim = 300).

(6) Gene2Vec [28]: bulk gene expression embeddings are learned from PCC-based gene–gene correlation patterns using the skip-gram (dim = 200).

(7) FRoGS [26]: bulk gene expression embeddings are trained on the ARCHS4 RNA-seq expression compendium (dim = 256).

(8) TCGA-Embedding [27]: bulk gene expression embeddings are trained on The Cancer Genome Atlas gene expression profiles (dim = 50).

Compared methods for spatial transcriptomics prediction and pathological region segmentation

We compare CoxFormer with existing scRNA-seq reference-based imputation methods that predict gene expression in spatial transcriptomics data. To ensure reproducibility, each baseline is run with the recommended settings, and key parameters are fixed as follows.

(1) gimVI [38]: implemented in the Python package *scvi-tools* and run following the official tutorial from `https://docs.scvi-tools.org/en/0.8.0/user_guide/notebooks/gimvi_tutorial.html`, where the spatial distribution of genes is obtained using the function `model.get_imputed_values()` with the argument `normalized = False`;

(2) SpaGE [36]: implemented from the GitHub repository *SpaGE* and run following the tutorial notebook at `https://github.com/tabdealal/SpaGE/blob/master/SpaGE_Tutorial.ipynb`, where we set the parameter `n_pv` = $N_{\text{gene}}/2$;

(3) Tangram [37]: implemented in the Python package *tangram* following the instructions at `https://github.com/broadinstitute/Tangram`, where we set the key parameters to `modes = "clusters"` and `density = "rna_count_based"`;

(4) Seurat [41]: implemented in the R package *Seurat* (version 5.3.0) following the online integration tutorial at `https://satijalab.org/seurat/articles/integration_mapping`, where we used `reduction = "cca"` and `k.filter = NA`; if $N_{\text{gene}} > 30$ we set `dims = 30`, otherwise we set `dims` = N_{gene}, and the predicted spatial distribution of genes is obtained using the function `TransferData()`;

(5) SpaOTsc [35]: implemented from the Python package *SpaOTsc* following the instructions at `https://github.com/zcang/SpaOTsc`, where the spatial distribution of genes is computed using the function `transport_plan()` with parameters `alpha = 0`, `rho = 1.0`, `epsilon = 0.1`, and `scaling = False`;

(6) novoSpaRc [40]: implemented in the Python package *novosparc* and run following the tutorial at `https://github.com/rajewsky-lab/novosparc/blob/master/reconstruct_drosophila_embryo_tutorial.ipynb`, where we set `alpha_linear = 0.5`, `loss_fun = "square_loss"`, and `epsilon` = 5×10^{-3};

(7) LIGER [42]: implemented from the GitHub repository `https://github.com/MacoskoLab/liger`, where the predicted spatial distribution of genes is obtained by k-nearest neighbors (kNN)-based imputation in the shared low-dimensional space (`knn = 30`);

(8) stPlus [39]: implemented from the GitHub repository `http://github.com/xy-chen16/stPlus` and following the recommended settings with `t_min = 5` and `n_neighbors = 50`.

Compared methods for spatial epigenomics prediction

We compare CoxFormer with three single-source embedding variants based on gene co-expression, gene correlation and gene description. The co-expression and correlation representations are constructed from an 18,858 $\times$ 18,858 COXPRESdb-derived matrix and a 32,101 $\times$ 32,101 transcriptome-wide correlation matrix estimated from the HCA, respectively. Because directly using these high-dimensional per-gene vectors is memory-prohibitive, we apply PCA to reduce them to 512 dimensions, consistent with CoxFormer. In contrast, the description-based representations comprised 43,661 $\times$ 3,072 text embeddings obtained via the OpenAI embedding API and were used in their original form. Across all variants, we keep the Transformer-based spatial generative framework, with the training protocol and all other settings unchanged.

Compared methods for super resolution enhancement

We compare CoxFormer with iStar [45], a widely used framework for enhancing the spatial resolution of Visium measurements by leveraging histology results to infer finer-grained expression

patterns. iStar is implemented from the GitHub repository `https://github.com/daviddaiweizhang/istar`, following the default training setting with `epochs = 400` and `lr = 1e-4`.

Evaluation metrics

We evaluate the performance of CoxFormer and competing methods using a set of complementary metrics. In brief, gene-level tasks and pathological region segmentation tasks are assessed as supervised classification problems using five standard metrics: AUROC, Accuracy, F1-score, Precision, and Recall (AUROC, ACC, F1, PRE, and REC, respectively). Cell-level tasks are evaluated with both supervised cell-type classification (the same five metrics) and unsupervised clustering (ARI, NMI, and AMI), and summarized into an aggregated cell-task score. Spatial prediction tasks are primarily evaluated using three complementary spatial inference metrics: Pearson correlation coefficient (PCC), structural similarity index (SSIM), and root mean squared error (RMSE).

Classification metrics

For supervised classification, we use metrics based on the confusion matrix and predicted probabilities. ACC, PRE, REC, and F1 quantify the point-estimate performance at a fixed decision threshold (0.5). Let TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. These metrics are defined as:

$$\begin{aligned} \mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \qquad \mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \\ \mathrm{REC} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad\qquad \mathrm{F1} = \frac{2\cdot\mathrm{PRE}\cdot\mathrm{REC}}{\mathrm{PRE}+\mathrm{REC}}. \end{aligned} \tag{18}$$

The AUROC evaluates the global discriminative ability across all decision thresholds. It can be expressed as the integral of a ROC curve that plots the true positive rate (TPR) against the false positive rate (FPR):

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(\tau)\, d\mathrm{FPR}(\tau). \tag{19}$$

Clustering metrics

For unsupervised clustering, we evaluate the agreement between a predicted partition U and ground-truth groups V using information-theoretic and pairwise-consistency metrics. Specifically, we report NMI, AMI, and ARI, where larger values indicate better alignment with the ground-truth structure.

NMI measures the mutual dependence between two partitions and is normalized by their entropies. With $I(U;V)$ denoting mutual information and $H(\cdot)$ denoting entropy:

$$\mathrm{NMI}(U,V) = \frac{2\, I(U;V)}{H(U)+H(V)}. \tag{20}$$

AMI corrects mutual information for chance agreement using the expected mutual information $\mathbb{E}[I(U;V)]$ under a standard permutation model:

$$\mathrm{AMI}(U,V) = \frac{I(U;V)-\mathbb{E}[I(U;V)]}{\frac{1}{2}(H(U)+H(V))-\mathbb{E}[I(U;V)]}. \tag{21}$$

ARI assesses pairwise agreement between partitions, corrected for chance. Let RI be the Rand index and $\mathbb{E}[\mathrm{RI}]$ be its expected value:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{1 - \mathbb{E}[\mathrm{RI}]}, \tag{22}$$

where the denominator uses $\max(\mathrm{RI}) = 1$.

Gene-task score

To provide a single summary number for gene-level benchmarks, we compute a gene-task score by averaging the five classification metrics over the four gene-level tasks. Let $\mathcal{T}_g$ denote the set of gene-level tasks and $\mathcal{M}_{\mathrm{cls}} = \{\mathrm{ACC}, \mathrm{PRE}, \mathrm{REC}, \mathrm{F1}, \mathrm{AUROC}\}$. The gene-task score is:

$$\text{gene task score} = \frac{1}{|\mathcal{T}_g|} \sum_{\text{task} \in \mathcal{T}_g} \left(\frac{1}{|\mathcal{M}_{\mathrm{cls}}|} \sum_{m \in \mathcal{M}_{\mathrm{cls}}} m(\text{task}) \right). \tag{23}$$

Cell-task score

Similarly, we compute a cell-task score by averaging classification and clustering metrics over the four scRNA-seq datasets. Let $\mathcal{T}_c$ denote the datasets and $\mathcal{M}_{\mathrm{clust}} = \{\mathrm{ARI}, \mathrm{NMI}, \mathrm{AMI}\}$. The cell-task score is:

$$\text{cell task score} = \frac{1}{|\mathcal{T}_c|} \sum_{\text{dataset} \in \mathcal{T}_c} \left(\frac{1}{|\mathcal{M}_{\mathrm{cls}} \cup \mathcal{M}_{\mathrm{clust}}|} \sum_{m \in \mathcal{M}_{\mathrm{cls}} \cup \mathcal{M}_{\mathrm{clust}}} m(\text{dataset}) \right). \tag{24}$$

Spatial inference metrics

To evaluate the consistency between the ground-truth and predicted spatial expression patterns, we use three complementary metrics: PCC, SSIM, and RMSE. For each gene g, let $\mathbf{y}_g = (y_{g,1}, \ldots, y_{g,n_{\mathrm{spot}}})^\top$ and $\hat{\mathbf{y}}_g = (\hat{y}_{g,1}, \ldots, \hat{y}_{g,n_{\mathrm{spot}}})^\top$ denote the ground-truth and predicted expression vectors across n_{spot} spots. Let $\mu(\cdot)$ and $\sigma(\cdot)$ denote the sample mean and sample standard deviation over spots, and let $\mathrm{cov}(\cdot, \cdot)$ denote the sample covariance. Higher values indicate better performance for PCC and SSIM, whereas lower values indicate better performance for RMSE.

PCC measures linear consistency between the two spatial profiles:

$$\mathrm{PCC}_g = \frac{\mathrm{cov}(\mathbf{y}_g, \hat{\mathbf{y}}_g)}{\sigma(\mathbf{y}_g)\, \sigma(\hat{\mathbf{y}}_g)}. \tag{25}$$

SSIM compares local spatial patterns by combining luminance, contrast, and structural terms. For each gene, we first rescale the spatial maps to $[0, 1]$:

$$y'_{g,j} = \frac{y_{g,j} - \min(\mathbf{y}_g)}{\max(\mathbf{y}_g) - \min(\mathbf{y}_g)}, \qquad \hat{y}'_{g,j} = \frac{\hat{y}_{g,j} - \min(\hat{\mathbf{y}}_g)}{\max(\hat{\mathbf{y}}_g) - \min(\hat{\mathbf{y}}_g)}, \qquad j = 1, \ldots, n_{\mathrm{spot}}, \tag{26}$$

and denote the rescaled vectors by $\mathbf{y}'_g$ and $\hat{\mathbf{y}}'_g$. With constants $C_1 = 0.01$ and $C_2 = 0.03$, SSIM is computed as:

$$\mathrm{SSIM}_g = \frac{\big(2\mu(\mathbf{y}'_g)\mu(\hat{\mathbf{y}}'_g) + C_1^2\big)\big(2\,\mathrm{cov}(\mathbf{y}'_g, \hat{\mathbf{y}}'_g) + C_2^2\big)}{\big(\mu(\mathbf{y}'_g)^2 + \mu(\hat{\mathbf{y}}'_g)^2 + C_1^2\big)\big(\sigma(\mathbf{y}'_g)^2 + \sigma(\hat{\mathbf{y}}'_g)^2 + C_2^2\big)}. \tag{27}$$

RMSE measures the absolute error between standardized profiles. We first apply z-score normalization across spots:

$$z_{g,j} = \frac{y_{g,j} - \mu(\mathbf{y}_g)}{\sigma(\mathbf{y}_g)}, \qquad \hat{z}_{g,j} = \frac{\hat{y}_{g,j} - \mu(\hat{\mathbf{y}}_g)}{\sigma(\hat{\mathbf{y}}_g)}, \tag{28}$$

and then compute:

$$\mathrm{RMSE}_g = \sqrt{\frac{1}{n_{\mathrm{spot}}} \sum_{j=1}^{n_{\mathrm{spot}}} \left(\hat{z}_{g,j} - z_{g,j}\right)^2}. \tag{29}$$

Datasets and preprocessing

Large-scale scRNA-seq data for correlation learning

In the computation of gene correlations, we curate a comprehensive scRNA-seq dataset from the HCA, comprising 10.4×10^6 cells across 123 healthy human tissue projects generated with 10x Genomics 3′ V2 chemistry. Raw count matrices are processed using the *Scanpy* (version 1.9.8) library [67], computing quality control (QC) metrics for mitochondrial (MT), ribosomal, and hemoglobin genes. We apply a rigorous outlier detection strategy based on median absolute deviations (MADs). Cells are excluded if their library size, detected gene count, or the percentage of counts in the top 20 genes deviated by more than 5 MADs from the population median. Furthermore, cells exhibiting $> 8\%$ mitochondrial counts or those identified as outliers (> 3 MADs) in mitochondrial proportion are removed to ensure high data quality.

ScRNA-seq data for cell-level task

We test cell-type annotation on four datasets, including Human $\mathrm{CD34}^+$ Bone Marrow (BM) ($n = 5,780$ cells; 10 cell types), Breast Cancer FFPE (BC) ($n = 10,689$ cells; 19 cell types), Diffuse Large B-cell Lymphoma FFPE (DLBL) ($n = 39,713$ cells; 17 cell types), and Lung Cancer FFPE (LC) ($n = 35,954$ cells; 19 cell types). For each scRNA-seq dataset, we normalize the raw counts by library size to a total sum of $10,000$, which is then followed by log-transformation ($\log(x+1)$).

Spatial transcriptomics data for gene expression imputation task

We benchmark CoxFormer's performance for spatial gene expression prediction using six human breast cancer (HBC) datasets profiled by the 10x Genomics Visium platform, paired with a matched scRNA-seq reference. For both spatial and scRNA-seq datasets, we first normalize the raw counts by library size to a total sum of $10,000$, which is then followed by log-transformation ($\log(x+1)$). Specifically, for the spatial transcriptomics data, we implement an additional quality control step, where genes expressed in fewer than 10% of spatial spots are filtered out to mitigate technical noise. We then identify highly variable genes (HVGs) using the *Scanpy* (version 1.9.8) with the `sc.pp.highly_variable_genes` function (`flavor="seurat_v3"`) [67], selecting the top 1,000 genes for spatial transcriptomics prediction. To assess whether CoxFormer can generate transcript profiles of unknown targets, we randomly partition 90% of HVGs for training and reserve the remaining 10% genes for evaluation.

Spatial epigenomics data for cross-modal applications

We demonstrate the utility of CoxFormer for cross-modality prediction with a human hippocampus dataset measured by spatial ATAC–RNA-seq. For the spatial transcriptomics layer, we apply library-size normalization to a total sum of $10,000$ and log-transformation to identify the top $1,000$ HVGs using *Scanpy* (v1.9.8) with `sc.pp.highly_variable_genes` (`flavor="seurat_v3"`) [67]. For spatial epigenomics data, the chromatin accessibility data are obtained as gene activity scores (GAS) from Ref. [44], which are calculated using the Gene Score model implemented in ArchR. To benchmark the prediction, we train the model on 90% of the identified HVG GAS profiles and reserve the remaining 10% for assessment.

Spatial subcelluar transcriptomics data for super-resolution enhancement

We utilize high-resolution human skin datasets profiled by the 10x Genomics Xenium *in situ* platform to benchmark CoxFormer's performance in super-resolution enhancement. We collect two sections, one that is measured with the standard Xenium Human Skin Gene Expression Panel comprising 280 genes, and another that is profiled with an augmented panel including custom targets, totaling 378 genes.

To evaluate super-resolution imputation, we construct pseudo-Visium datasets by spatially binning the high-resolution Xenium transcript coordinates into spot-level measurements. Specifically, we define capture locations within the tissue area at 100 μm intervals, constructing circular spots with a diameter of 55 μm to match the standard Visium spot geometry. We assign each Xenium transcript to the nearest spot center and aggregated transcript counts per gene within each spot to obtain a spot-by-gene expression matrix. Spots with zero assigned transcripts are removed, yielding two pseudo-Visium skin datasets at spot resolution.

In addition to spot-level expression, we derive cell-scale visual features from the paired histology images. We rescale the Xenium histology images to a resolution of 0.5 μm/pixel and extract a morphological feature vector for each non-overlapping 16×16 pixel block using a Vision Transformer (ViT) model pre-trained on large-scale histology images [64]. Notably, each block represents an 8 $\mu m \times 8$ μm tissue region, which approximately corresponds to the scale of a small mammalian cell scale ($10 - 20$ μm in diameter).

Spatial transcriptomics data for pathological regions segmentation

For abnormal region detection, we use four spatial transcriptomics sections from a colorectal cancer liver metastasis (CRLM) cohort, including two primary colorectal tumor (CRC) sections and two matched liver metastasis (LM) sections [69]. The slides span two clinical conditions, with two sections from an untreated case and two sections from a therapy-treated case showing partial response (PR). Concretely, CRC1 corresponds to an untreated CRC section, LM1 corresponds to an untreated LM section, CRC2 corresponds to a PR CRC section, and LM2 corresponds to a PR LM section. For each slide, we processe the raw spot counts using a standard workflow, first normalizing by library size to a total count of 10,000 per spot and then subjected to log-transformation using $\log(x + 1)$. We further perform an additional quality control step by filtering out genes expressed in fewer than 10% of spatial spots to reduce technical noise.

For gene splitting, we use a housekeeping (HK) gene-based strategy to predict a pseudo-healthy reference. Specifically, HK genes, defined by a curated list from the Ref. [50], serve as observed inputs for model training, whereas all non-HK genes are treated as prediction targets for abnormal region detection. For quantitative evaluation, we randomly hold out 100 HK genes from the input set and included them in the test targets, enabling a direct comparison between predicted and measured expression.

Data availability

All datasets utilized in this study are publicly accessible. (1) the COXPRESdb database (*Microarray_m.c7* version) is available at `https://coxpresdb.jp/download/`; (2) the Human Cell Atlas single-cell RNA sequencing (scRNA-seq) dataset is available at `https://explore.data.humancellatlas.org/projects`; (3) Gene classification datasets are available at `https://huggingface.co/datasets/ctheodoris/Genecorpus-30M/tree/main/example_input_files/gene_classification`. (4) scRNA-seq datasets for cell-level task: Diffuse Large B-cell Lymphoma FFPE (DLBL), Breast Cancer FFPE (BC), Lung Cancer FFPE (LC) are accessed from the CZ CELLxGENE collection `https://cellxgene.cziscience.com/collections/bd552f76-1f1b-43a3-b9ee-0aace57e90d6`, and Bone Marrow (BM) are downloaded via `scvelo.datasets.bonemarrow()`; (5) 10x Visium human breast cancer spatial transcriptomics data (HBC1–6) are available from Zenodo (`https://zenodo.org/record/4739739`), corresponding to samples `CID4535` (HBC1), `CID44971` (HBC2), `CID4465` (HBC3), `CID4290` (HBC4), `1160920F` (HBC5), and `CID3586` (HBC6); (6) 10x Chromium human breast cancer scRNA-seq data are available from GEO under accession `GSE176078` (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176078`); (7) spatial ATAC–RNA-seq human brain hippocampal spatial transcriptomics and epigenomics data are available at `https://cells.ucsc.edu/?ds=brain-spatial-omics`; (8) 10x Xenium human skin melanoma data (Section A) are available from 10x Genomics at `https://www.10xgenomics.com/datasets/human-skin-preview-data-xenium-human-skin-gene-expression-panel-add-on-1-standard`; (9) 10x Xenium human skin melanoma data (Section B) are available from 10x Genomics at `https://www.10xgenomics.com/datasets/human-skin-preview-data-xenium-human-skin-gene-expression-panel-1-standard`; (10) 10x Visium human colorectal cancer liver metastasis spatial transcriptomics data (CRC1–2 and LM1–2) are available from scCRLM (`http://www.cancerdiversity.asia/scCRLM`), corresponding to samples `ST-P1, colon` (CRC1), `ST-P4, colon` (CRC2), `ST-P2, liver` (LM1), and `ST-P4, liver` (LM2); (11) Illumina NextSeq 500 scRNA-seq data from normal human colon are available from GEO under accession `GSM7290762` (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7290762`); (12) Illumina NextSeq 500 scRNA-seq data from normal human liver are available from GEO under accession `GSM7290760` (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7290760`).

Code availability

CoxFormer is freely available as a Python package accessible at `https://github.com/yyyancy/CoxFormer` with detailed tutorials and documentation at `https://coxformer.readthedocs.io/en/latest/`. To facilitate replication of the figures and results presented in this paper, detailed workflows are provided as Jupyter notebooks within the GitHub repository and documentation.

References

[1] T. Obayashi, S. Kodate, H. Hibara, Y. Kagaya, K. Kinoshita, Coxpresdb v8: an animal gene coexpression database navigating from a global view to detailed investigations. Nucleic acids research **51**(D1), D80–D87 (2023)

[2] S. Van Dam, U. Vosa, A. van der Graaf, L. Franke, J.P. de Magalhaes, Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in bioinformatics **19**(4), 575–592 (2018)

[3] Z. Wang, H. Fang, N.L.S. Tang, M. Deng, Vcnet: vector-based gene co-expression network construction and its application to rna-seq data. Bioinformatics **33**(14), 2173–2181 (2017)

[4] S. Lu, S. Keleş, Debiased personalized gene coexpression networks for population-scale scrna-seq data. Genome Research **33**(6), 932–947 (2023)

[5] S. Marco Salas, L.B. Kuemmerle, C. Mattsson-Langseth, S. Tismeyer, C. Avenel, T. Hu, H. Rehman, M. Grillo, P. Czarnewski, S. Helgadottir, et al., Optimizing xenium in situ data utility by quality assessment and best-practice analysis workflows. Nature Methods pp. 1–11 (2025)

[6] S. He, R. Bhatt, C. Brown, E.A. Brown, D.L. Buhr, K. Chantranuvatana, P. Danaher, D. Dunaway, R.G. Garrison, G. Geiss, et al., High-plex imaging of rna and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nature biotechnology **40**(12), 1794–1806 (2022)

[7] K.H. Chen, A.N. Boettiger, J.R. Moffitt, S. Wang, X. Zhuang, Spatially resolved, highly multiplexed rna profiling in single cells. Science **348**(6233), aaa6090 (2015)

[8] J.R. Moffitt, D. Bambah-Mukku, S.W. Eichhorn, E. Vaughn, K. Shekhar, J.D. Perez, N.D. Rubinstein, J. Hao, A. Regev, C. Dulac, et al., Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science **362**(6416), eaau5324 (2018)

[9] P.L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J.F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J.O. Westholm, M. Huss, et al., Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)

[10] R.R. Stickels, E. Murray, P. Kumar, J. Li, J.L. Marshall, D.J. Di Bella, P. Arlotta, E.Z. Macosko, F. Chen, Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. Nature biotechnology **39**(3), 313–319 (2021)

[11] F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, J. Yao, scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. Nature Machine Intelligence **4**(10), 852–866 (2022)

[12] C.V. Theodoris, L. Xiao, A. Chopra, M.D. Chaffin, Z.R. Al Sayed, M.C. Hill, H. Mantineo, E.M. Brydon, Z. Zeng, X.S. Liu, et al., Transfer learning enables predictions in network biology. Nature **618**(7965), 616–624 (2023)

[13] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, B. Wang, scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature methods **21**(8), 1470–1480 (2024)

[14] S. Ding, J. Li, R. Luo, H. Cui, B. Wang, R. Chen, scgpt: end-to-end protocol for fine-tuned retinal cell type annotation. Nature Protocols pp. 1–21 (2025)

[15] E. Simon, K. Swanson, J. Zou, Language models for biological research: a primer. Nature Methods **21**(8), 1422–1429 (2024)

[16] M.E. Consens, C. Dufault, M. Wainberg, D. Forster, M. Karimzadeh, H. Goodarzi, F.J. Theis, A. Moses, B. Wang, Transformers and genome language models. Nature Machine Intelligence pp. 1–17 (2025)

[17] Y. Chen, J. Zou, Simple and effective embedding model for single-cell biology built from chatgpt. Nature biomedical engineering **9**(4), 483–493 (2025)

[18] Y. Chen, J. Zou, Genepert: Leveraging genept embeddings for gene perturbation prediction. bioRxiv pp. 2024–10 (2024)

[19] O. Zhu, J. Li, Scouter predicts transcriptional responses to genetic perturbations with large language model embeddings. Nature Computational Science pp. 1–8 (2025)

[20] G. Wang, T. Liu, J. Zhao, Y. Cheng, H. Zhao, Modeling and predicting single-cell multi-gene perturbation responses with sclambda. bioRxiv (2024)

[21] O. Rozenblatt-Rosen, M.J. Stubbington, A. Regev, S.A. Teichmann, The human cell atlas: from vision to reality. Nature **550**(7677), 451–453 (2017)

[22] G.R. Brown, V. Hem, K.S. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K.D. Pruitt, D.R. Maglott, et al., Gene: a gene-centered information resource at ncbi. Nucleic acids research **43**(D1), D36–D42 (2015)

[23] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T.I. Stein, R. Nudel, I. Lieder, Y. Mazor, et al., The genecards suite: from gene data mining to disease genome sequence analyses. Current protocols in bioinformatics **54**(1), 1–30 (2016)

[24] U. Consortium, et al., Uniprot: the universal protein knowledgebase in 2025. Nucleic Acids Research **53**(D1), D609 (2024)

[25] OpenAI. New embedding models and api updates. `https://openai.com/index/new-embedding-models-and-api-updates/` (2024). Accessed: 2026-01-16

[26] H. Chen, F.J. King, B. Zhou, Y. Wang, C.J. Canedy, J. Hayashi, Y. Zhong, M.W. Chang, L. Pache, J.L. Wong, et al., Drug target prediction through deep learning functional representation of gene signatures. Nature communications **15**(1), 1853 (2024)

[27] C.T. Choy, C.H. Wong, S.L. Chan, Infer related genes from large scale gene expression dataset with embedding. BioRxiv p. 362848 (2018)

[28] J. Du, P. Jia, Y. Dai, C. Tao, Z. Zhao, D. Zhi, Gene2vec: distributed representation of genes based on co-expression. BMC genomics **20**(Suppl 1), 82 (2019)

[29] Q. Chen, K. Lee, S. Yan, S. Kim, C.H. Wei, Z. Lu, Bioconceptvec: creating and evaluating literature-based biomedical concept embeddings on a large scale. PLoS computational biology **16**(4), e1007617 (2020)

[30] S. Kim, H. Lee, K. Kim, J. Kang, Mut2vec: distributed representation of cancerous mutations. BMC medical genomics **11**(Suppl 2), 33 (2018)

[31] J. Grootens, J.S. Ungerstedt, G. Nilsson, J.S. Dahlin, Deciphering the differentiation trajectory from hematopoietic stem cells to mast cells. Blood advances **2**(17), 2273–2281 (2018)

[32] W. Tang, X. Wang, B. Han, S.H. Jiang, H. Cao, Tumor-associated macrophages in cancer: from mechanisms to application. Molecular Biomedicine **6**(1), 145 (2025)

[33] M. Laviron, M. Petit, E. Weber-Delacroix, A.J. Combes, A.R. Arkal, S. Barthélémy, T. Courau, D.A. Hume, C. Combadière, M.F. Krummel, et al., Tumor-associated macrophage heterogeneity is driven by tissue territories in breast cancer. Cell reports **39**(8) (2022)

[34] S.Z. Wu, G. Al-Eryani, D.L. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J.R. Torpy, N. Bartonicek, et al., A single-cell and spatially resolved atlas of human breast cancers. Nature genetics **53**(9), 1334–1347 (2021)

[35] Z. Cang, Q. Nie, Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nature communications **11**(1), 2084 (2020)

[36] T. Abdelaal, S. Mourragui, A. Mahfouz, M.J. Reinders, Spage: spatial gene enhancement using scrna-seq. Nucleic acids research **48**(18), e107–e107 (2020)

[37] T. Biancalani, G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C.R. Vanderburg, Å. Segerstolpe, M. Zhang, et al., Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. Nature methods **18**(11), 1352–1362 (2021)

[38] R. Lopez, A. Nazaret, M. Langevin, J. Samaran, J. Regier, M.I. Jordan, N. Yosef, *A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements*, in *ICML Workshop on Computational Biology* (2019)

[39] C. Shengquan, Z. Boheng, C. Xiaoyang, Z. Xuegong, J. Rui, stplus: a reference-based method for the accurate enhancement of spatial transcriptomics. Bioinformatics **37**(Supplement_1), i299–i307 (2021)

[40] M. Nitzan, N. Karaiskos, N. Friedman, N. Rajewsky, Gene expression cartography. Nature **576**(7785), 132–137 (2019)

[41] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W.M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. cell **177**(7), 1888–1902 (2019)

[42] J.D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, E.Z. Macosko, Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell **177**(7), 1873–1887 (2019)

[43] H. Zhang, Y. Wang, C. Liu, W. Li, F. Zhou, X. Wang, J. Zheng, The apolipoprotein c1 is involved in breast cancer progression via emt and mapk/jnk pathway. Pathology-Research and Practice **229**, 153746 (2022)

[44] D. Zhang, Y. Deng, P. Kukanja, E. Agirre, M. Bartosovic, M. Dong, C. Ma, S. Ma, G. Su, S. Bao, et al., Spatial epigenome–transcriptome co-profiling of mammalian tissues. Nature **616**(7955), 113–122 (2023)

[45] D. Zhang, A. Schroeder, H. Yan, H. Yang, J. Hu, M.Y. Lee, K.S. Cho, K. Susztak, G.X. Xu, M.D. Feldman, et al., Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. Nature biotechnology **42**(9), 1372–1377 (2024)

[46] G. Reynolds, P. Vegh, J. Fletcher, E.F. Poyner, E. Stephenson, I. Goh, R.A. Botting, N. Huang, B. Olabi, A. Dubois, et al., Developmental cell programs are co-opted in inflammatory skin disease. Science **371**(6527), eaba6500 (2021)

[47] J. Winkler, A. Abisoye-Ogunniyan, K.J. Metcalf, Z. Werb, Concepts of extracellular matrix remodelling in tumour progression and metastasis. Nature communications **11**(1), 5120 (2020)

[48] O.J. Harrison, J. Brasch, G. Lasso, P.S. Katsamba, G. Ahlsen, B. Honig, L. Shapiro, Structural basis of adhesive binding by desmocollins and desmogleins. Proceedings of the National Academy of Sciences **113**(26), 7160–7165 (2016)

[49] S.J. Gardai, Y.Q. Xiao, M. Dickinson, J.A. Nick, D.R. Voelker, K.E. Greene, P.M. Henson, By binding sirpα or calreticulin/cd91, lung collectins act as dual function surveillance molecules to suppress or enhance inflammation. Cell **115**(1), 13–23 (2003)

[50] B.W. Hounkpe, F. Chenou, F. De Lima, E.V. De Paula, Hrt atlas v1. 0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive rna-seq datasets. Nucleic acids research **49**(D1), D947–D955 (2021)

[51] H. Xu, Y. Ma, J. Zhang, J. Gu, X. Jing, S. Lu, S. Fu, J. Huo, Identification and verification of core genes in colorectal cancer. BioMed Research International **2020**(1), 8082697 (2020)

[52] M. Schott, D. León-Periñán, E. Splendiani, L. Strenger, J.R. Licha, T.M. Pentimalli, S. Schallenberg, J. Alles, S.S. Tagliaferro, A. Boltengagen, et al., Open-st: High-resolution spatial transcriptomics in 3d. Cell **187**(15), 3953–3972 (2024)

[53] H. Zhao, X. Yin, S. Wang, Z. Geng, W. Yu, H. Yu, S. Ning, L. Wang, So3d: a comprehensive three-dimensional spatial omics resource for decoding tissue architecture in physiology and disease. Nucleic Acids Research p. gkaf998 (2025)

[54] P. Xie, J. Shen, Y. Yang, X. Wang, W. Liu, H. Cao, Y. Zheng, C. Wu, G. Mao, L. Chen, et al., Digital reconstruction of full embryos during early mouse organogenesis. Cell (2025)

[55] E. Karimi, N. Simo, N. Milet, W. TE, A. ALSH, N. QU, L. AIL, R. ABS, A. ALIND, N.M. GOODMA, et al., Method of the year 2024: spatial proteomics. Nat Methods **21**, 2195–2196 (2024)

[56] M. Wu, H. Tao, T. Xu, X. Zheng, C. Wen, G. Wang, Y. Peng, Y. Dai, Spatial proteomics: unveiling the multidimensional landscape of protein localization in human diseases. Proteome science **22**(1), 7 (2024)

[57] B. Hu, J. Zhu, F. Zhao, The evolving landscape of spatial proteomics technologies in the ai age. Fundamental Research (2024)

[58] K. Wheeler, C. Gosmanov, M.J. Sandoval, Z. Yang, L.I. McCall, Frontiers in mass spectrometry-based spatial metabolomics: Current applications and challenges in the context of biomedical research. TrAC Trends in Analytical Chemistry **175**, 117713 (2024)

[59] X. Min, Y. Zhao, M. Yu, W. Zhang, X. Jiang, K. Guo, X. Wang, J. Huang, T. Li, L. Sun, et al., Spatially resolved metabolomics: From metabolite mapping to function visualising. Clinical and Translational Medicine **14**(11), e70031 (2024)

[60] Q. Yuan, Z. Duren, Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. Nature Biotechnology **43**(2), 247–257 (2025)

[61] M.W. Funk, Y. Wang, L. Wang, Airqtl dissects cell state-specific causal gene regulatory networks with efficient single-cell eqtl mapping. Nature Communications (2025)

[62] W.L. Hamilton, Z. Ying, J. Leskovec, *Inductive Representation Learning on Large Graphs*, in *The Thirty-First Annual Conference on Neural Information Processing Systems* (2017), pp. 1024–1034

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Advances in neural information processing systems **30** (2017)

[64] R.J. Chen, C. Chen, Y. Li, T.Y. Chen, A.D. Trister, R.G. Krishnan, F. Mahmood, *Scaling vision transformers to gigapixel images via hierarchical self-supervised learning*, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 16144–16155

[65] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

[66] D. Dimitrov, P.S.L. Schäfer, E. Farr, P. Rodriguez-Mier, S. Lobentanzer, P. Badia-i Mompel, A. Dugourd, J. Tanevski, R.O. Ramirez Flores, J. Saez-Rodriguez, Liana+ provides an all-in-one framework for cell–cell communication inference. Nature Cell Biology **26**(9), 1613–1622 (2024)

[67] F.A. Wolf, P. Angerer, F.J. Theis, Scanpy: large-scale single-cell gene expression data analysis. Genome biology **19**(1), 15 (2018)

[68] Z. Fang, X. Liu, G. Peltz, Gseapy: a comprehensive package for performing gene set enrichment analysis in python. Bioinformatics **39**(1), btac757 (2023)

[69] Y. Wu, S. Yang, J. Ma, Z. Chen, G. Song, D. Rao, Y. Cheng, S. Huang, Y. Liu, S. Jiang, et al., Spatiotemporal immune landscape of colorectal cancer liver metastasis at single-cell level. Cancer discovery **12**(1), 134–153 (2022)

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (grant No. 12371283 to J.L.; No. 725B2030 to Y.Y.).

Author contributions

J.L. initiated and designed the study; Y.Y. developed and implemented the model and built the pipeline with assistance from X.L. and H.Z.; Y.Y., X.L., and H.Z. performed the benchmark evaluations and the downstream analyses; J.L., Y.Y., and X.L. wrote the manuscript; Y.Y., X.L., H.Z., Y.D.W., X.S., Y.W., T.Y., and J.L. edited and revised the manuscript.

Competing Interests

The authors declare no competing interests.

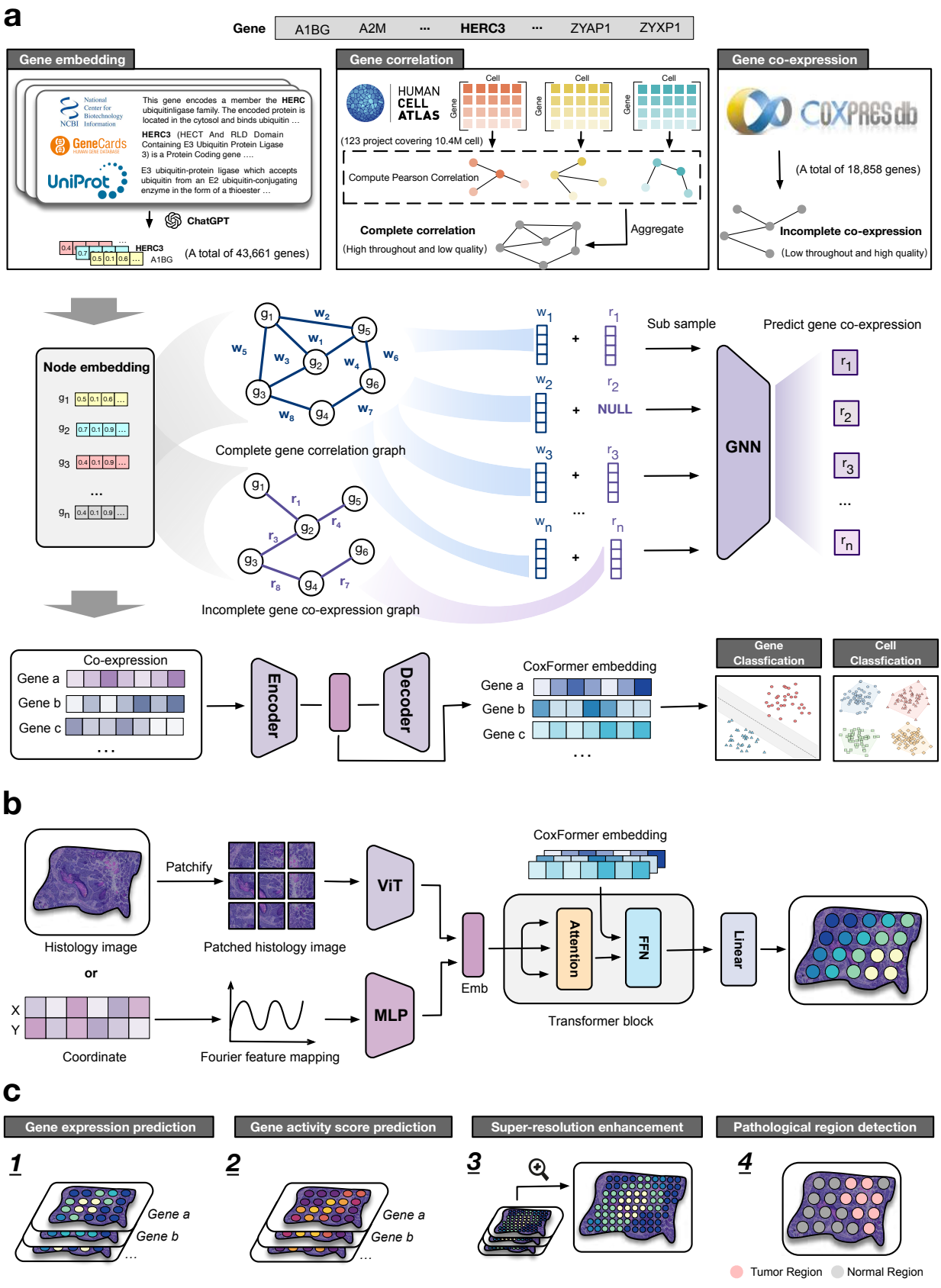

a
Gene
A1BG
A2M
···
HERC3
···
ZYAP1
ZYXP1
Gene embedding
This gene encodes a member the HERC ubiquitinligase family. The encoded protein is located in the cytosol and binds ubiquitin …
HERC3 (HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase 3) is a Protein Coding gene ….
E3 ubiquitin-protein ligase which accepts ubiquitin from an E2 ubiquitin-conjugating enzyme in the form of a thioester …
ChatGPT
HERC3
A1BG
(A total of 43,661 genes)
Gene correlation
HUMAN CELL ATLAS
Cell
Gene
(123 project covering 10.4M cell)
Compute Pearson Correlation
Complete correlation
(High throughout and low quality)
Aggregate
Gene co-expression
COXPRESdb
(A total of 18,858 genes)
Incomplete co-expression
(Low throughout and high quality)
Node embedding
Complete gene correlation graph
Incomplete gene co-expression graph
Sub sample
Predict gene co-expression
NULL
GNN
Co-expression
Gene a
Gene b
Gene c
Encoder
Decoder
CoxFormer embedding
Gene Classfication
Cell Classfication
b
Patchify
Histology image
Patched histology image
ViT
or
X
Y
Coordinate
Fourier feature mapping
MLP
Emb
CoxFormer embedding
Attention
FFN
Transformer block
Linear
c
Gene expression prediction
Gene activity score prediction
Super-resolution enhancement
Pathological region detection
1
2
3
4
Gene a
Gene b
Tumor Region
Normal Region

Figure 1: **Overview of CoxFormer universal gene embedding and cross-modal spatial inference.** CoxFormer consists of two sequential modules for (1) learning universal gene representations and (2) enabling versatile spatial omics prediction. **(a)** Universal gene-embedding learning. CoxFormer integrates complementary information from three sources: (i) curated gene descriptions collected from NCBI Gene, GeneCards, and UniProt and encoded as text-derived node features (43,661 genes); (ii) transcriptome-wide gene–gene correlations computed from >10 million HCA cells from across 123 projects to provide scalable correlation-based relational structures (32,101 genes); and (iii) high-quality but incomplete bulk-tissue co-expression associations from COXPRESdb (18,858 genes). A graph neural network (GNN) uses the text-derived node features together with correlation-informed message passing to propagate relational patterns across the transcriptome, reconstructing and completing bulk co-expression associations beyond the observed COXPRESdb subset. The completed whole-transcriptome co-expression profiles are then compressed into 512-dimensional CoxFormer embeddings using an autoencoder, yielding universal gene representations for downstream gene- and cell-level tasks. **(b)** Cross-modal spatial generative framework. A Transformer-based framework integrates CoxFormer embeddings with high-resolution histology images (via a visual Transformer [ViT] encoder) and spatial coordinates (via Fourier feature mapping and a Multi-Layer Perceptron), and models long-range spot-to-spot dependencies to capture spatial context for molecular prediction. **(c)** Downstream spatial applications enabled by the framework: (1) gene expression prediction, (2) gene activity score prediction, (3) super-resolution enhancement, and (4) pathological region detection.

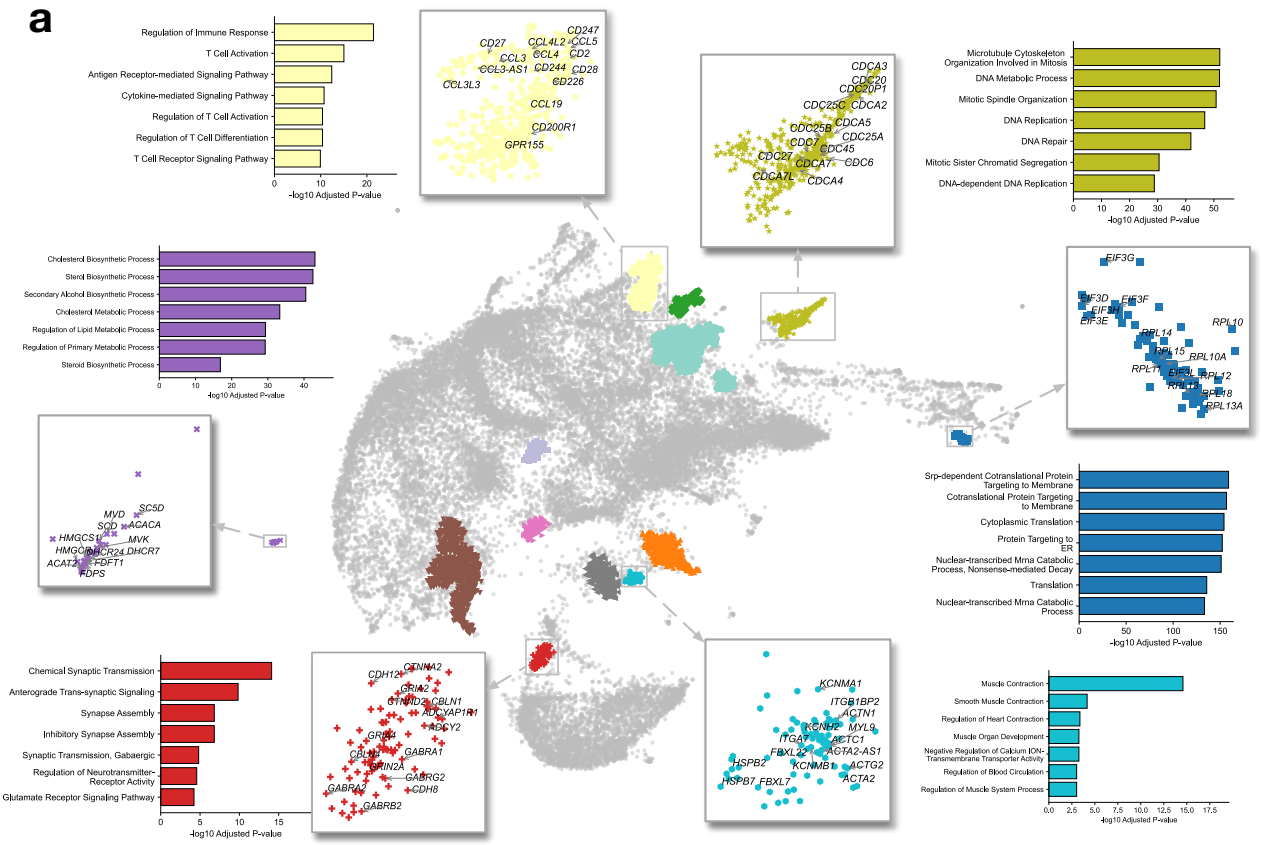

a
Regulation of Immune Response
T Cell Activation
Antigen Receptor-mediated Signaling Pathway
Cytokine-mediated Signaling Pathway
Regulation of T Cell Activation
Regulation of T Cell Differentiation
T Cell Receptor Signaling Pathway
-log10 Adjusted P-value
CD247
CD27
CCL4L2
CCL5
CCL3
CCL4
CD2
CCL3-AS1
CD244
CD28
CCL3L3
CD226
CCL19
CD200R1
GPR155
CDCA3
CDC20
CDC20P1
CDC25C
CDCA2
CDCA5
CDC25B
CDC25A
CDC7
CDC45
CDC27
CDCA7
CDC6
CDCA7L
CDCA4
Microtubule Cytoskeleton Organization Involved in Mitosis
DNA Metabolic Process
Mitotic Spindle Organization
DNA Replication
DNA Repair
Mitotic Sister Chromatid Segregation
DNA-dependent DNA Replication
-log10 Adjusted P-value
Cholesterol Biosynthetic Process
Sterol Biosynthetic Process
Secondary Alcohol Biosynthetic Process
Cholesterol Metabolic Process
Regulation of Lipid Metabolic Process
Regulation of Primary Metabolic Process
Steroid Biosynthetic Process
-log10 Adjusted P-value
EIF3G
EIF3D
EIF3F
EIF3H
EIF3E
RPL10
RPL14
RPL10A
RPL15
RPL7
RPL12
RPL18
RPL13A
MVD
SC5D
SQLE
ACACA
MVK
HMGCS1
HMGCR
DHCR24
DHCR7
ACAT2
FDFT1
FDPS
Srp-dependent Cotranslational Protein Targeting to Membrane
Cotranslational Protein Targeting to Membrane
Cytoplasmic Translation
Protein Targeting to ER
Nuclear-transcribed Mrna Catabolic Process, Nonsense-mediated Decay
Translation
Nuclear-transcribed Mrna Catabolic Process
-log10 Adjusted P-value
Chemical Synaptic Transmission
Anterograde Trans-synaptic Signaling
Synapse Assembly
Inhibitory Synapse Assembly
Synaptic Transmission, Gabaergic
Regulation of Neurotransmitter-Receptor Activity
Glutamate Receptor Signaling Pathway
-log10 Adjusted P-value
CTNNA2
CDH12
GRIA2
GRIN2A
CBLN1
ADCYAP1R1
ADCY2
GRIN1
CBLN2
GABRA1
GRIN2A
GABRG2
GABRA2
CDH8
GABRB2
KCNMA1
ITGB1BP2
ACTN1
KCNN2
MYL9
ITGA7
ACTC1
FBXL2
ACTA2-AS1
HSPB2
KCNMB1
ACTG2
HSPB7
FBXL7
ACTA2
Muscle Contraction
Smooth Muscle Contraction
Regulation of Heart Contraction
Muscle Organ Development
Negative Regulation of Calcium ION-Transmembrane Transporter Activity
Regulation of Blood Circulation
Regulation of Muscle System Process
-log10 Adjusted P-value
b
Gene Task Score
DS
BG
MC
TF
Cell Task Score
BM
DLBL
LC
BC
Overall Score
CoxFormer
ScGPT–Human
GeneFormer–20L95M
GeneFormer–6L30M
ScGPT–PanCancer
GF–12L95M–Cancer
GeneFormer–12L95M
GeneFormer–12L30M
Gene2Vec
FRoGS
GenePT–Model3
BioVec–FastText
BioVec–SkipGram
BioVec–CBOW
TCGA–Embedding
GenePT–Ada
Mut2Vec
BioVec–GloVe
Ranking
1
18
Score
0%
100%
c
DCBL
LC
BC
BM
Malignant
Epithelial/Gland
T cells
NK cells
Granulocytes
Mono/Macro
Mast cells
Endothelial
Fibroblasts
Pericytes
Smooth muscle
B/Plasma
Dendritic cells
Erythroid
Megakaryocytes
HSC
CLP
Precursors

Figure 2: **CoxFormer organization of genes into an interpretable functional manifold and evaluation across multiple gene- and cell-level tasks. (a)** UMAP of CoxFormer embeddings for 32,016 human genes. Genes form compact, well-separated modules, with related programs showing meaningful proximity. Colored clusters indicate representative functional groups; for each cluster, Gene Ontology enrichment analysis highlights coherent biological themes (bar plots), and representative marker genes are annotated in the insets. **(b)** Benchmark of CoxFormer and gene embedding baselines derived from bulk transcriptomics, single-cell transcriptomics, and biomedical literature. Bubble plot summarizing four gene-level binary classification tasks (dosage sensitivity of transcription factors, transcription-factor regulatory range, and promoter-state classification: bivalent versus non-methylated, and Lys4-only-methylated versus non-methylated) and four cell-level tasks across Diffuse Large B-cell Lymphoma FFPE (DLBL), Lung Cancer FFPE (LC), Breast Cancer FFPE (BC), and Human CD34+ Bone Marrow (BM) scRNA-seq datasets using cell embeddings computed as expression-weighted averages of gene embeddings. Bubble size denotes the normalized score (larger is better) and bubble color encodes the ranking within each metric (lighter indicates better performance). **(c)** UMAP visualization of CoxFormer-derived cell embeddings across the four datasets (DLBL, LC, BC, and BM) showing dataset-specific cell-state structure and separation of major populations.

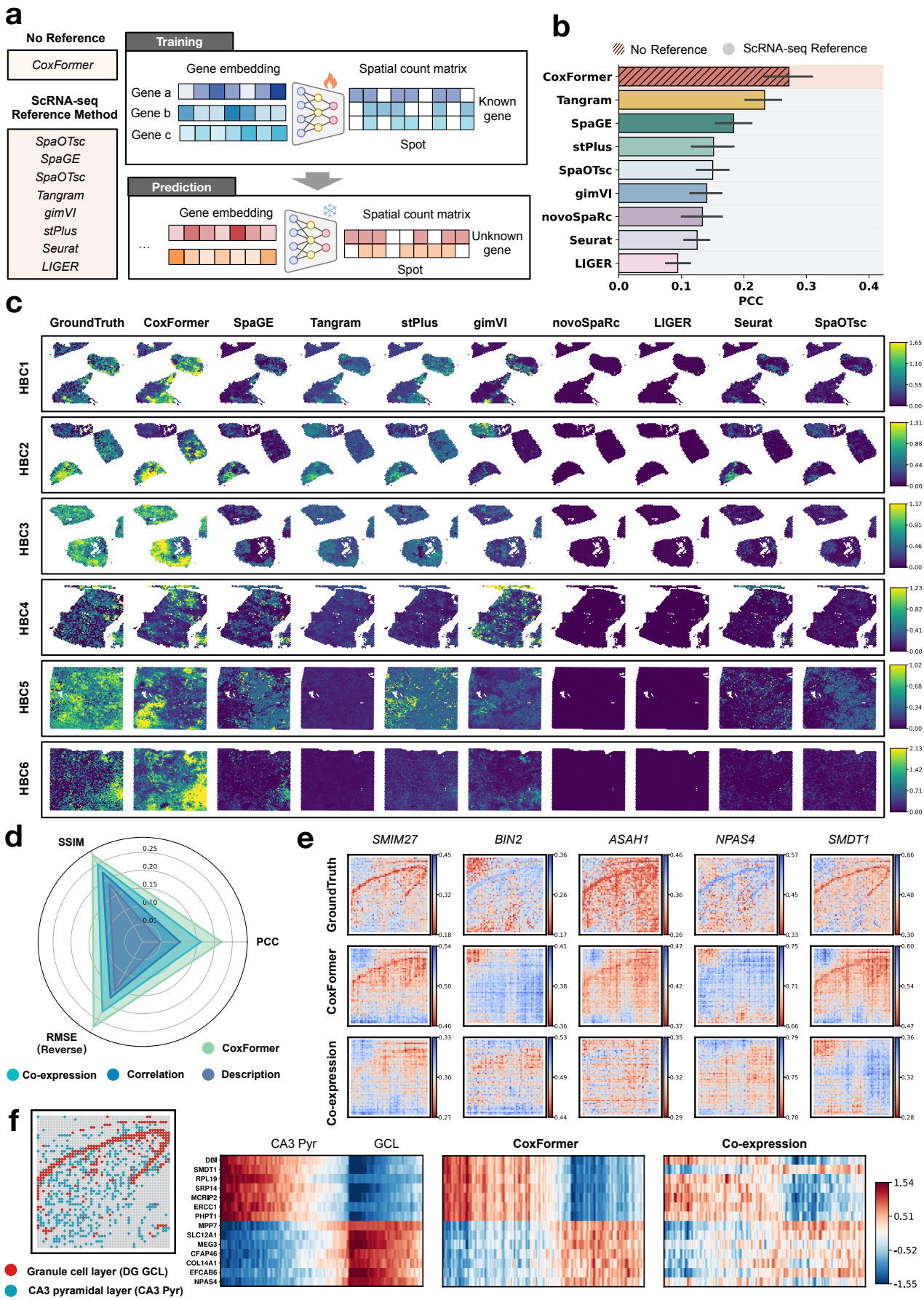

a
No Reference
CoxFormer
ScRNA-seq Reference Method
SpaOTsc
SpaGE
SpaOTsc
Tangram
gimVI
stPlus
Seurat
LIGER
Training
Gene embedding
Gene a
Gene b
Gene c
Spatial count matrix
Known gene
Spot
Prediction
Gene embedding
Spatial count matrix
Unknown gene
Spot
b
No Reference
ScRNA-seq Reference
CoxFormer
Tangram
SpaGE
stPlus
SpaOTsc
gimVI
novoSpaRc
Seurat
LIGER
0.0
0.1
0.2
0.3
0.4
PCC
c
GroundTruth
CoxFormer
SpaGE
Tangram
stPlus
gimVI
novoSpaRc
LIGER
Seurat
SpaOTsc
HBC1
HBC2
HBC3
HBC4
HBC5
HBC6
d
SSIM
PCC
RMSE (Reverse)
CoxFormer
Co-expression
Correlation
Description
e
SMIM27
BIN2
ASAH1
NPAS4
SMDT1
GroundTruth
CoxFormer
Co-expression
f
Granule cell layer (DG GCL)
CA3 pyramidal layer (CA3 Pyr)
CA3 Pyr
GCL
CoxFormer
Co-expression
DBI
SMDT1
RPL19
SRP14
MCRIP2
ERCC1
PHPT1
MPP7
SLC12A1
MEG3
CFAP46
COL14A1
EFCAB6
NPAS4
1.54
0.51
-0.52
-1.55

Figure 3: **CoxFormer enables spatial omics inference for unseen genes.** **(a)** Overview of the pseudo-panel benchmark for unseen-gene prediction in spatial transcriptomics. CoxFormer is trained on observed genes and predicts held-out targets in a zero-shot manner without external scRNA-seq reference atlases. **(b)** Mean prediction accuracy across six human breast cancer (HBC) Visium datasets, quantified by Pearson's correlation coefficient (PCC) on held-out genes. CoxFormer is compared with representative reference-based baselines (SpaOTsc, SpaGE, Tangram, gimVI, stPlus, novoSpaRc, Seurat and LIGER), all provided with a matched scRNA-seq atlas. **(c)** Representative spatial expression maps for a held-out gene (*APOC1*) across six HBC sections. CoxFormer consistently recovers the ground-truth spatial pattern, whereas competing methods show dataset-dependent failures or collapse to near-uniform predictions. **(d)** Cross-omics evaluation of a human hippocampus spatial ATAC-RNA-seq dataset via PCC, structural similarity index (SSIM), and root mean squared error (RMSE). Radar plot comparing CoxFormer with ablated variants that isolate individual information sources (co-expression, correlation, and description) for predicting gene activity scores (GAS) of held-out genes. **(e)** Predicted versus ground-truth GAS maps for five representative genes (*SMIM27*, *BIN2*, *ASAH1*, *NPAS4*, and *SMDT1*). CoxFormer better matches spatial contrast and localization than the ablated variants. **(f)** Region-by-gene GAS heatmaps for two hippocampal layers (dentate gyrus granule cell layer [DG GCL] and the CA3 pyramidal layer [CA3 Pyr]) using the top differentially expressed genes. CoxFormer largely preserves the layer-specific block structure and directional changes observed in the ground truth.

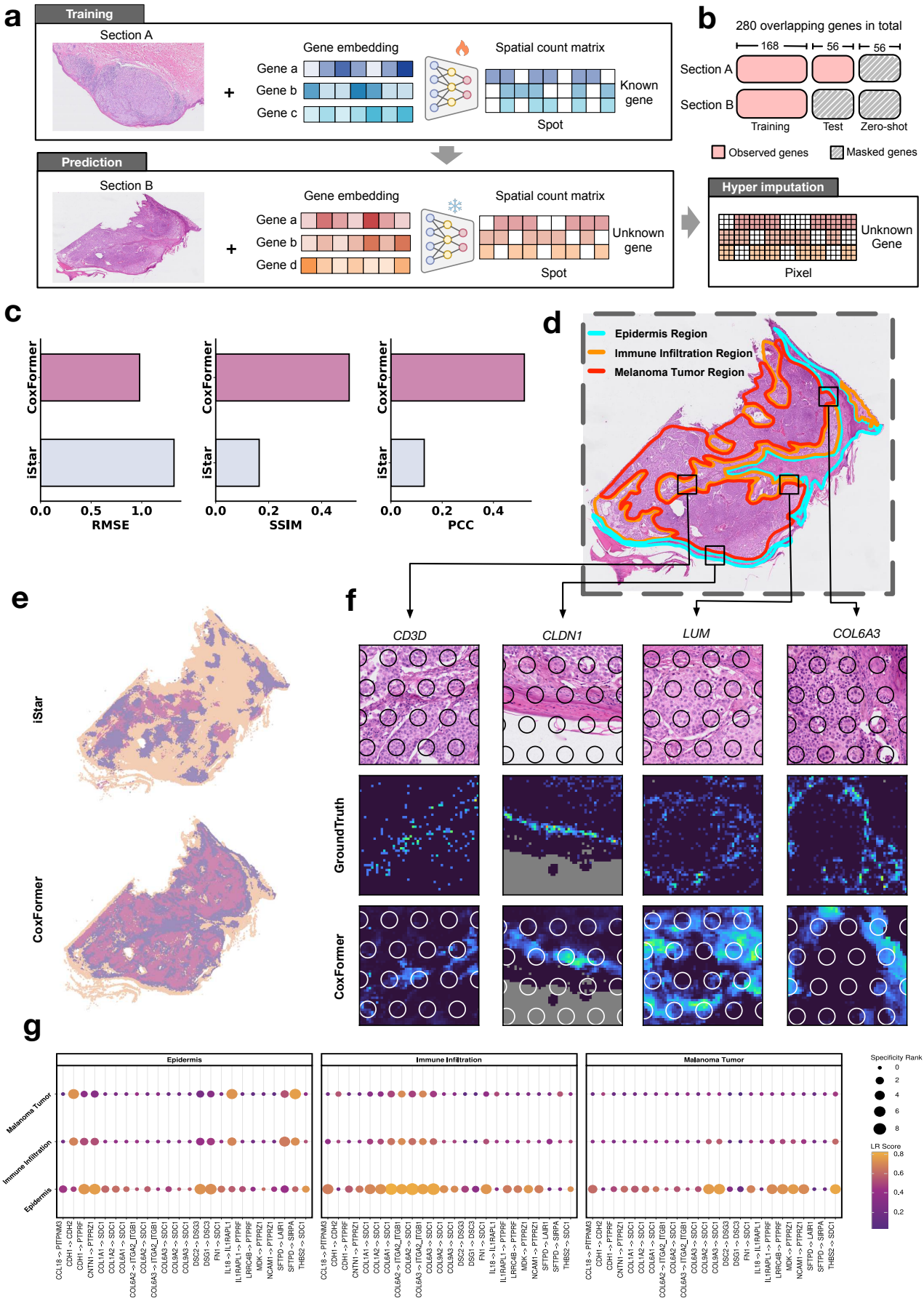

a
Training
Section A
Gene embedding
Gene a
Gene b
Gene c
Spatial count matrix
Known gene
Spot
Prediction
Section B
Gene embedding
Gene a
Gene b
Gene d
Spatial count matrix
Unknown gene
Spot
b
280 overlapping genes in total
168
56
56
Section A
Section B
Training
Test
Zero-shot
Observed genes
Masked genes
Hyper imputation
Unknown Gene
Pixel
c
CoxFormer
iStar
0.0
0.5
1.0
RMSE
0.0
0.2
0.4
SSIM
0.0
0.2
0.4
PCC
d
Epidermis Region
Immune Infiltration Region
Melanoma Tumor Region
e
iStar
CoxFormer
f
CD3D
CLDN1
LUM
COL6A3
GroundTruth
CoxFormer
g
Epidermis
Immune Infiltration
Malanoma Tumor
Malanoma Tumor
Immune Infiltration
Epidermis
Specificity Rank
0
2
4
6
8
LR Score
0.8
0.6
0.4
0.2

Figure 4: **CoxFormer enables super-resolution imputation and zero-shot prediction of unassayed genes in Xenium skin sections. (a)** Overview of the cross-section super-resolution setting. CoxFormer is trained on the observed genes across Sections A and B, then applied to Section B to generate spot-level and subcellular-scale predictions. **(b)** Gene partitioning for evaluation on two Xenium human skin sections (Section A: 378-gene panel; Section B: 280-gene panel). The 280 overlapping genes are split into training genes (fully observed; n=168), test genes (observed in Section A but masked in Section B; n=56) and zero-shot genes (masked in both sections; n=56). **(c)** Spot-level imputation accuracy for the 56 masked test genes in Section B. CoxFormer outperforms the super-resolution baseline iStar across RMSE, SSIM and PCC. **(d)** Predefined tissue regions in Section B (epidermis, immune infiltration and melanoma tumor) used for downstream evaluation. **(e)** Region separation obtained by clustering the predicted subcellular-scale expression in Section B for iStar and CoxFormer; the latter yields clearer separation of epidermal and immune-infiltrated domains. **(f)** Zero-shot subcellular-scale prediction for genes absent from the measured panel in Section B. Shown are representative marker genes (*CD3D*, *CLDN1*, *LUM* and *COL6A3*) with Xenium ground-truth and CoxFormer predictions; CoxFormer recovers compartment-specific patterns and fine-grained ring-like spatial structures. **(g)** Region-resolved cell-cell communication analysis in Section B after transcriptome-wide augmentation with CoxFormer-imputed genes. Dot plots show the top 10 ligand-receptor pairs per region ranked by specificity; dot size encodes specificity rank, and color indicates the ligand-receptor (LR) score.

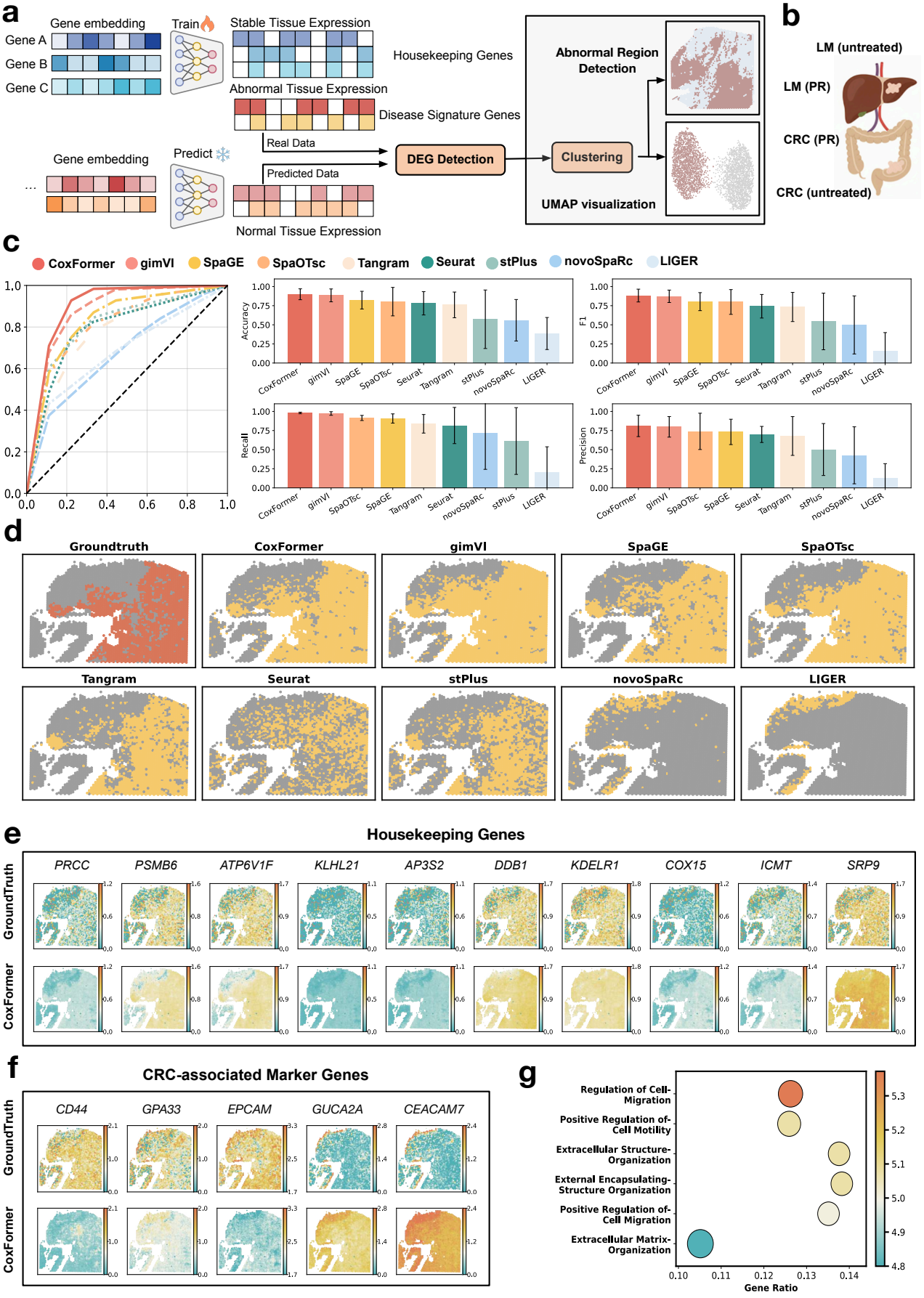

a
Gene embedding
Gene A
Gene B
Gene C
Train
Stable Tissue Expression
Housekeeping Genes
Abnormal Tissue Expression
Disease Signature Genes
Real Data
Predicted Data
Gene embedding
Predict
Normal Tissue Expression
DEG Detection
Clustering
Abnormal Region Detection
UMAP visualization
b
LM (untreated)
LM (PR)
CRC (PR)
CRC (untreated)
c
CoxFormer
gimVI
SpaGE
SpaOTsc
Tangram
Seurat
stPlus
novoSpaRc
LIGER
Accuracy
F1
Recall
Precision
d
Groundtruth
CoxFormer
gimVI
SpaGE
SpaOTsc
Tangram
Seurat
stPlus
novoSpaRc
LIGER
e
Housekeeping Genes
PRCC
PSMB6
ATP6V1F
KLHL21
AP3S2
DDB1
KDELR1
COX15
ICMT
SRP9
GroundTruth
CoxFormer
f
CRC-associated Marker Genes
CD44
GPA33
EPCAM
GUCA2A
CEACAM7
GroundTruth
CoxFormer
g
Regulation of Cell-Migration
Positive Regulation of-Cell Motility
Extracellular Structure-Organization
External Encapsulating-Structure Organization
Positive Regulation of-Cell Migration
Extracellular Matrix-Organization
Gene Ratio

Figure 5: **Reference-free pathological region detection in colorectal cancer liver metastasis using CoxFormer.** **(a)** Overview of the unsupervised, reference-free pipeline. A spatial generative framework is trained on spatially stable housekeeping genes to learn a "pseudo-healthy" baseline, then used to generate transcriptome-wide expression; residuals on non-housekeeping genes are used for DEG detection and spatial clustering to segment normal versus abnormal regions. **(b)** Four colorectal cancer liver metastasis sections used for benchmarking, spanning untreated and therapy-perturbed (PR) colorectal cancer (CRC) and liver metastasis (LM) tissues. **(c)** Aggregate abnormal-region detection performance across sections, comparing CoxFormer with reference-based spatial imputation methods (SpaGE, SpaOTsc, Tangram, gimVI, novoSpaRc, Seurat, LIGER, and stPlus). The left panel shows the mean ROC curve, and the right panel summarizes AUROC, recall, accuracy, precision, and F1 score. **(d)** Spatial segmentation maps on an untreated colorectal cancer section. CoxFormer delineates abnormal regions that closely match the ground-truth annotations, whereas competing methods show noisy or incorrect boundaries. **(e)** Representative housekeeping genes showing the agreement between measured expression and CoxFormer reconstruction, which provides a stable baseline for downstream deviation analysis. **(f)** Comparison of measured expression of representative CRC-associated marker genes (*CD44*, *GPA33*, *EPCAM*, *GUCA2A*, and *CEACAM7*) comparing measured expression with the CoxFormer-generated "pseudo-healthy" baseline, highlighting the disease-associated deviations within abnormal regions. **(g)** Gene Ontology biological process enrichment of the top differentially expressed genes between CoxFormer-segmented normal and abnormal regions. The processes are dominated by extracellular remodeling and motility-related programs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CoxFormersupplementarynotes.pdf
- CoxFormersupplementaryfigures.pdf
- nrreportingsummary.pdf