

A Experimental Setup

A.1 Implementation Details

Molecules are encoded using CLAMP (Seidl et al., 2023) into 768-dimensional vectors, while tissues are embedded using a heterogeneous graph neural network (GNN). Each tissue graph contains a variable number of samples, $G = 11,560$ genes, and $P = 1,493$ pathways. Pathway-pathway connectivity is defined by a curated binary adjacency matrix derived from Pathformer (Liu et al., 2024), producing a $P \times P$ graph. Pathway node features—including degree, clustering coefficient, and spectral embeddings—are combined into 10-dimensional vectors and projected to a 192-dimensional hidden space. Gene expression values are log-transformed, standardized, and encoded to 192-dimensional embeddings via a feedforward layer.

The GNN backbone consists of two HeteroGNN layers. SAGEConv (Hamilton et al., 2017) is applied to sample-gene edges to reduce memory usage given the large number of samples, while GATv2Conv (Brody et al., 2022) with 3 attention heads is applied to gene-pathway and pathway-pathway edges. Each layer includes residual connections, dropout ($\gamma = 0.1$), and layer normalization. Tissue embeddings are obtained by pooling over sample node embeddings with a multi-query attention module using 4 learned queries and 4 attention heads.

The training objective combines gene reconstruction ($\lambda_{\text{gene}} = 0.5$), tissue classification ($\lambda_{\text{cls}} = 0.5$), and pathway contrastive ($\lambda_{\text{pathway}} = 0.25$) losses (see Figure 3 for a sensitivity analysis of ROC-AUC performance to these loss weighting coefficients). The model is trained using the Adam optimizer (learning rate 1×10^{-3} , weight decay 1×10^{-5}) with a batch size of 256, early stopping with a patience of 5 epochs, and modality dropout of 0.25 applied independently to molecular and tissue embeddings. Pretraining runs for 100 epochs, followed by fine-tuning for up to 15 epochs on a single NVIDIA A100 GPU. The final model is selected based on the checkpoint achieving the highest validation ROC-AUC. The complete training process requires approximately 12 hours of computation time.

A.2 Datasets

Table 2: Overview of the human-only compound-tissue-contextual activity prediction dataset. For each tissue, the table shows the total number of compound-tissue pairs, the number and percentage of positive (active) and negative (inactive) examples.

Tissue	Total	Positive	Negative	Pos (%)	Neg (%)	# Assays
Brain	316	163	153	51.6	48.4	4
Breast	21832	3495	18337	16.0	84.0	6
Cervix	25693	1470	24223	5.7	94.3	9
Kidney	59357	2486	56871	4.2	95.8	39
Liver	143285	8457	134828	5.9	94.1	337
Lung	14520	1632	12888	11.2	88.8	22
Ovary	15938	1854	14084	11.6	88.4	2
Prostate	3528	841	2687	23.8	76.2	4
Skin	21780	2933	18847	13.5	86.5	42

We evaluate our model on publicly available, human-specific compound-tissue-contextual activity data drawn from the EdelweissData 3.2 release, which integrates curated assay and transcriptomic measurements from Open TG-GATEs (Igarashi et al., 2015) and DrugMatrix (Svoboda et al., 2019). For our benchmarks, we use a panel of standard cell-based toxicity assays drawn from public screening platforms, each defined on a common set of small molecules. Only human samples were retained; non-human samples, replicates, and incomplete assay entries were removed prior to integration to ensure consistency in compound-tissue mapping.

Activity labels in EdelweissData 3.2 correspond to experimentally measured molecular responses classified as active (1) or inactive (0) according to the EPA-standardized hit-call procedure. These labels are derived directly from the curated datasets and integrate potency, efficacy, concentration-response curve shape, and model confidence, providing a consistent binary summary of the biological response across diverse assays.

Table 3: Summary of assays in the human-only compound–tissue-contextual activity prediction dataset. Each assay is a cell-based or biochemical experiment in human cell systems (including established cell lines and, for some endpoints, primary cell systems), measured at the indicated post-treatment time, with standardized concentration–response fitting and binary hit calls.

Tissue	Cell Type	Plate	Readout	Direction	Target
Brain	Cortical membranes	96-well	Radioligand	Up	GPCR / Rhodopsin-like
Breast	MDA-kb2	1536-well	Luciferase	Up	Nuclear / Steroidal
Cervix	HeLa	1536-well	Viability	Down	Cell cycle / Cytotoxicity
Kidney	HEK293T	1536-well	Background	Up	Artifact / Background
Liver	HepG2	24-well	mRNA (RT-PCR)	Down	Nuclear / Non-steroidal
Lung	Bronchial epithelial	96-well	ELISA	Up	Cytokine / Chemotactic
Ovary	VM7	1536-well	Luciferase	Up	Nuclear / Steroidal
Prostate	22Rv1	96/384-well	Growth	Down	Cell cycle / Cytotoxicity
Skin	Keratinocytes/fibroblasts	96-well	ELISA	Down	Protease / MMP

Each assay in the panel is processed using the EPA `tcp1` pipeline, which fits concentration–response curves and assigns harmonized activity calls. While originally developed for toxicology, these assays capture generalizable perturbational responses—including inflammation, stress-response activation, and transcriptional modulation—that are informative for modeling tissue-specific molecular activity.

The assays in our panel are all human cell-based (predominantly established lines, with some primary cell systems) and cover multiple tissues and cell types. All activity labels follow this format: cell-based assays in human or primary cell systems, with standardized concentration–response fitting, and summarized with a binary hit call per compound per endpoint. Table 3 provides a representative example of one assay per tissue, including assay details such as cell type, plate format, readout type, directionality (increase or decrease in signal relative to control), and target family. The full panel contains additional assays and endpoints that follow the same design principles.

From this harmonized dataset, we selected nine tissue-specific tasks. The compounds are split by scaffold (80% train, 10% validation, 10% test), with no overlap within or between tissues. To further ensure generalization, we performed a global scaffold split across all tissue-specific tasks, preventing the same compound scaffold from appearing in multiple tissues across training, validation, or test sets.

This panel of public, human-specific assays provides a standardized, biologically grounded benchmark for evaluating molecular representation models. Previous work has used subsets of these assays for similar classification tasks (Heusinkveld et al., 2018); here, we extend the benchmark across multiple tissues and compound sets. Data are available at <https://ui.staging.kit.cloud.douglasconnect.com/datasets?q=%7B%22searchAnywhere%22%3A%22summary%22%7D>.

It is worth noting that many of these public, human-specific assays rely on established cell lines rather than primary tissue. While such cell lines do not fully capture all aspects of in vivo tissue biology and microenvironmental context (Jiang et al., 2016), they are widely used preclinical models that provide reproducible measurements of perturbational responses. Throughout this work, we therefore treat each assay-defined cell line as an operational proxy for its tissue of origin and focus on modeling assay- and cell-type-specific tissue activity, rather than the complete physiological behavior of healthy tissues in vivo.

Tissue gene expression from GTEx v10 (Consortium, 2020) (available at https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression) is modeled per sample. To avoid leakage from donors appearing in multiple tissues, splits are made at the donor level and combined with scaffold-based compound splits, ensuring no overlap and true generalization. To verify that no GTEx donor or assay compound overlaps across different stages of training, all donor IDs and compound scaffolds were globally cross-checked to prevent leakage between pretraining, fine-tuning, and evaluation. For pretraining, all available GTEx tissues are used to learn general tissue-level embeddings. During fine-tuning, only tissues with measured compound activity are included. Assay-derived tissue labels are mapped directly to GTEx tissue names (details in our code repository for reproducibility), and donor-level splits are coordinated with scaffold-based compound splits to prevent leakage across pretraining and fine-tuning. During evaluation on the downstream activity prediction

Table 4: Number of trainable parameters for Espresso and baseline models.

Model	# Trainable Parameters
Espresso	60,238,966
BioBERT	108,311,041
CLAMP	55,588,352
GROVER	48,514,284
KA-GNN	2,376,704
Mol-BERT	44,104,705
MolFormer	45,557,762

task, the model uses the tissue embeddings learned as fixed representations, without accessing any new donor-specific expression data, ensuring that predictions rely solely on precomputed tissue-level features.

Pathway information comes from Pathformer (Liu et al., 2024), aggregating KEGG (Kanehisa & Goto, 2000), PID (Schaefer et al., 2009), Reactome (Croft et al., 2010), and BioCarta (Nishimura, 2001), filtered to 1,493 pathways based on gene number, overlap ratio, and sub-pathway counts. Gene-pathway associations are stored in a binary matrix of size $11,560 \times 1,493$ (genes \times pathways), where entry (g, p) indicates whether gene g belongs to pathway p . Pathformer also computed pathway crosstalk adjacency using BinoX (Ogris et al., 2017), which quantifies the association between pathways by analyzing relationships among their genes in genome-wide functional association networks. In their implementation, FunCoup v3.0 (Schmitt et al., 2014) was used as the functional network, with a link weight cutoff of 0.75, a minimum group size of 15, and 100 iterations of the sampling method. The resulting pathway crosstalk network is provided as a $1,493 \times 1,493$ adjacency matrix.

A.3 Baselines

We benchmark Espresso against six representative molecular representation models spanning language-based, graph-based, and multi-modal paradigms: BioBERT (Lee et al., 2020), CLAMP (Seidl et al., 2023), GROVER (Rong et al., 2020), KA-GNN (Bresson et al., 2025), Mol-BERT (Li & Jiang, 2021), and MolFormer (Wu et al., 2023). All models were evaluated per tissue using the same train/validation/test partitions as Espresso, ensuring strict comparability across architectures. Pretrained molecular encoders were fine-tuned on the tissue-specific molecular activity dataset with early stopping based on validation loss to prevent overfitting. All baselines share identical data splits, random seeds, and batch sampling strategies to ensure fair comparison. Fine-tuning employed standard binary cross-entropy loss and Adam optimization. No class weighting or data augmentation was applied unless specified in the original implementation. The number of trainable parameters for Espresso and all baseline models is reported in Table 4.

- **BioBERT (Lee et al., 2020):** A biomedical language model pretrained on PubMed abstracts and fine-tuned for tissue-specific compound activity prediction. Fine-tuning was performed using textual prompts that specified both the compound and the target tissue, reflecting the specific biological activity being modeled. For example, a prompt could be: “Predict the response of compound X in lung epithelial cells, measured by cytokine release” where compound X refer to a specific compound from our dataset, and the target tissue refers to the corresponding biological assay. Fine-tuning was conducted on our training dataset, utilizing the same compounds and their respective tissue-specific activity labels. The model was trained for up to three epochs with early stopping based on validation loss.
- **CLAMP (Seidl et al., 2023):** A cross-modal model jointly pretrained on molecular SMILES and assay descriptions via contrastive learning. Then, it was further fine-tuned for molecular activity prediction using paired SMILES and tissue-specific textual prompts, optimizing all model parameters for up to three epochs with early stopping on validation loss.
- **GROVER (Rong et al., 2020):** A self-supervised graph neural network pretrained on millions of molecular graphs to capture local and global chemical context. Fine-tuned for up to 20 epochs with early stopping using graph representations derived from SMILES and evaluated per tissue.

Table 5: Performance of Espresso and baseline models for predicting molecular activity in human tissues, measured by PR-AUC. PR-AUC highlights performance under class imbalance, with higher values indicating better precision–recall trade-offs.

Model	Brain	Breast	Cervix	Kidney	Liver	Lung	Ovary	Prostate	Skin
Espresso	0.7417	0.5660	0.4136	0.2394	0.1917	0.4222	0.2431	0.3995	0.4628
BioBERT	0.6981	0.5055	0.3444	0.2261	0.1720	0.4252	0.2407	0.3879	0.4258
CLAMP	0.6774	0.3038	0.1078	0.0906	0.0828	0.2073	0.1659	0.3282	0.2112
GROVER	0.6610	0.3542	0.1895	0.1487	0.1341	0.3016	0.2134	0.3568	0.2897
KA-GNN	0.6896	0.4402	0.2679	0.1775	0.1549	0.3917	0.2094	0.4010	0.3963
Mol-BERT	0.6116	0.4915	0.3508	0.2322	0.1760	0.4252	0.2289	0.4298	0.4313
MolFormer	0.5445	0.1576	0.0449	0.0415	0.0523	0.1044	0.1056	0.2657	0.1315

Table 6: Performance of Espresso for predicting molecular activity in human tissues. Metrics include ROC-AUC, PR-AUC, sensitivity, specificity, and Brier score, all computed per tissue. DeLong p -values and Benjamini–Hochberg (BH) corrected p -values indicate the statistical significance of ROC-AUC differences compared to a baseline model. Higher ROC-AUC, PR-AUC, sensitivity, and specificity indicate better predictive performance, while lower Brier scores indicate improved calibration.

Model	Brain	Breast	Cervix	Kidney	Liver	Lung	Ovary	Prostate	Skin
ROC-AUC	0.6471	0.8233	0.9325	0.8764	0.8207	0.8763	0.7466	0.6953	0.8950
PR-AUC	0.7417	0.5660	0.4136	0.2394	0.1917	0.4222	0.2431	0.3995	0.4628
Sensitivity	0.7059	0.8301	0.8382	0.7316	0.7205	0.8125	0.6054	0.8111	0.8613
Specificity	0.4286	0.5732	0.8422	0.8273	0.7468	0.7547	0.7521	0.4160	0.7532
Brier Score	0.2888	0.2581	0.1147	0.1090	0.1477	0.1654	0.1868	0.3044	0.1612
DeLong p -value	0.9190	$< 10^{-4}$	$< 10^{-5}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-9}$	0.4507	0.1431	$< 10^{-12}$
BH-corrected p	0.9190	$< 10^{-4}$	$< 10^{-5}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-9}$	0.5008	0.1785	$< 10^{-12}$

- **KA-GNN (Bresson et al., 2025):** A heterogeneous graph neural network employing kernel-based attention mechanisms to model higher-order molecular interactions. Fine-tuned for up to 100 epochs with early stopping on molecular graphs derived from SMILES and evaluated per tissue.
- **Mol-BERT (Li & Jiang, 2021):** A transformer-based molecular language model (ChemBERTa) pre-trained on large SMILES corpora to learn chemical syntax and substructure representations. Fine-tuned for up to three epochs with early stopping on all training molecules combined and evaluated per tissue.
- **MolFormer (Wu et al., 2023):** A hierarchical transformer model pretrained on large-scale molecular datasets using multi-scale attention to encode both local and global chemical dependencies. Fine-tuned for up to three epochs with early stopping on the combined training set and evaluated per tissue.

One might argue that Espresso’s performance gains partly stem from access to tissue context rather than architectural design alone. To address this, we include ablations where molecular encoders are paired with simplified tissue representations (Section 5.2). These analyses show that while context contributes to improved predictions, the hierarchical architecture of Espresso provides additional, measurable benefits beyond context alone.

Collectively, these baselines encompass a diverse range of molecular learning strategies—from biomedical language models and self-supervised graph networks to heterogeneous GNNs—providing a comprehensive and rigorous foundation for evaluating the molecular generalization and biological grounding of Espresso.

B Additional Experimental Results

We evaluated Espresso for tissue-specific molecular activity prediction across nine human tissues. Performance was assessed using multiple metrics—ROC-AUC, PR-AUC, sensitivity, specificity, Brier score—and

Table 7: Performance of Expresso for predicting molecular activity in human tissues, measured by ROC-AUC. Results show Expresso and baseline models, with 95% confidence intervals computed over 5 runs, each initialized with a different random seed. Higher values indicate better accuracy.

	Model	Brain	Breast	Cervix	Kidney	Liver
	Expresso	0.6471 ± 0.006	0.8233 ± 0.007	0.9325 ± 0.009	0.8764 ± 0.011	0.8207 ± 0.005
Baselines	BioBERT	0.6297 ± 0.014	0.7867 ± 0.015	0.8948 ± 0.008	0.8340 ± 0.003	0.7796 ± 0.021
	CLAMP	0.5546 ± 0.018	0.7849 ± 0.020	0.8815 ± 0.007	0.8038 ± 0.011	0.7712 ± 0.016
	GROVER	0.5844 ± 0.012	0.6018 ± 0.010	0.6060 ± 0.014	0.6056 ± 0.009	0.6062 ± 0.017
	KA-GNN	0.6008 ± 0.018	0.7538 ± 0.020	0.8461 ± 0.007	0.7878 ± 0.011	0.7491 ± 0.016
	Mol-BERT	0.5630 ± 0.009	0.7760 ± 0.013	0.8840 ± 0.011	0.8324 ± 0.004	0.7777 ± 0.012
	MolFormer	0.5882 ± 0.010	0.7886 ± 0.014	0.8989 ± 0.012	0.8356 ± 0.008	0.7834 ± 0.012
Ablation	Mol-BERT Encoder	0.6387 ± 0.002	0.7654 ± 0.006	0.8885 ± 0.007	0.8323 ± 0.012	0.7764 ± 0.013
	ChemGPT Encoder	0.6465 ± 0.004	0.5699 ± 0.008	0.6056 ± 0.006	0.5563 ± 0.011	0.5762 ± 0.007
	No Tissue Encoder	0.5756 ± 0.018	0.7871 ± 0.014	0.8895 ± 0.005	0.8061 ± 0.002	0.7739 ± 0.017
	Non-Graph Encoder	0.5672 ± 0.011	0.7855 ± 0.005	0.8884 ± 0.002	0.8291 ± 0.008	0.7801 ± 0.007
	One-Hot Encoder	0.6076 ± 0.004	0.8084 ± 0.006	0.8962 ± 0.012	0.8534 ± 0.009	0.7885 ± 0.004
	Text-Based Encoder	0.6092 ± 0.018	0.8007 ± 0.014	0.8993 ± 0.005	0.8312 ± 0.002	0.7849 ± 0.017
	Mean Tissue Expression	0.6176 ± 0.005	0.7891 ± 0.009	0.8886 ± 0.011	0.8325 ± 0.002	0.7812 ± 0.013
	No Pretraining	0.6180 ± 0.010	0.8001 ± 0.011	0.9110 ± 0.002	0.8552 ± 0.009	0.7924 ± 0.007
	Only $\mathcal{L}_{\text{activity}}$	0.6176 ± 0.011	0.8097 ± 0.009	0.8941 ± 0.010	0.8493 ± 0.004	0.7938 ± 0.008
	$\mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{gene}}$	0.6361 ± 0.014	0.8209 ± 0.013	0.9106 ± 0.006	0.8672 ± 0.015	0.8129 ± 0.004
	$\mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{cls}}$	0.6134 ± 0.012	0.8126 ± 0.015	0.9074 ± 0.007	0.8481 ± 0.016	0.7941 ± 0.007
	$\mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{pathway}}$	0.6218 ± 0.013	0.8141 ± 0.006	0.9112 ± 0.015	0.8553 ± 0.004	0.8052 ± 0.007
	Only Sample Nodes	0.5966 ± 0.006	0.7876 ± 0.017	0.8979 ± 0.009	0.8208 ± 0.015	0.7801 ± 0.003
	No Gene Nodes	0.6050 ± 0.003	0.8220 ± 0.005	0.9054 ± 0.016	0.8578 ± 0.013	0.7998 ± 0.014
	No Pathway Nodes	0.6387 ± 0.014	0.8091 ± 0.013	0.9044 ± 0.005	0.8616 ± 0.001	0.7949 ± 0.013

statistical comparison to the strongest baseline model per tissue via paired DeLong tests, with Benjamini-Hochberg (BH) correction applied for multiple testing. All scores represent the mean across test samples. Detailed results are summarized in Table 6. ROC-AUC results across baselines and Expresso, including 95% confidence intervals over 5 runs, are reported in Tables 7 and 8.

Expresso achieved high ROC-AUC values across tissues, reaching the highest performance in cervix (0.9325) and skin (0.8950). PR-AUC values reflect moderate performance in tissues with imbalanced classes, indicating the model’s ability to detect rare active molecular responses. Sensitivity and specificity are generally balanced, suggesting reliable discrimination between active and inactive molecules. Importantly, as shown in Table 5, Expresso consistently matches or exceeds the PR-AUC performance of most baseline models across tissues. The tissues in which PR-AUC remains low are the same for Expresso and the baselines, indicating that these limitations stem from intrinsic class imbalance and task difficulty rather than model-specific failure.

Brier scores are low in most tissues, demonstrating good probabilistic calibration. Paired DeLong tests with BH correction show that Expresso significantly outperforms the strongest baseline in most tissues, while a few tissues with smaller effect sizes did not reach statistical significance. 95% confidence intervals for ROC-AUC further support the robustness of these improvements. Notably, the tissues with higher DeLong p -values—primarily brain, ovary, and prostate—share two characteristics that help explain this pattern. First, they contain substantially fewer positive (active) compound-tissue pairs (Table 2), which increases the variance of ROC-AUC estimates and reduces statistical power, even when performance gains are present. Second, the strongest baseline model performs comparatively well in these tissues, narrowing the performance gap and making statistically significant differences harder to detect. Consequently, while Expresso still shows consistent numerical improvements in these cases, these differences do not reach significance after BH correction. This highlights how statistical significance is shaped jointly by dataset imbalance and the magnitude of model improvements across tissues.

Table 8: Performance of Expresso for predicting molecular activity in human tissues, measured by ROC-AUC. Results show Expresso and baseline models, with 95% confidence intervals computed over 5 runs, each initialized with a different random seed. Higher values indicate better accuracy.

	Model	Lung	Ovary	Prostate	Skin
	Expresso	0.8763 ± 0.008	0.7466 ± 0.004	0.6953 ± 0.014	0.8950 ± 0.011
Baselines	BioBERT	0.8526 ± 0.006	0.7234 ± 0.016	0.6848 ± 0.025	0.8622 ± 0.001
	CLAMP	0.8411 ± 0.018	0.6899 ± 0.007	0.6547 ± 0.004	0.8648 ± 0.020
	GROVER	0.6017 ± 0.013	0.6020 ± 0.006	0.6208 ± 0.015	0.6017 ± 0.005
	KA-GNN	0.8242 ± 0.018	0.6935 ± 0.007	0.6681 ± 0.004	0.8122 ± 0.008
	Mol-BERT	0.8582 ± 0.010	0.7137 ± 0.008	0.6753 ± 0.017	0.8672 ± 0.014
	MolFormer	0.8458 ± 0.015	0.7049 ± 0.004	0.6588 ± 0.017	0.8683 ± 0.009
Ablation	Mol-BERT Encoder	0.8534 ± 0.003	0.7032 ± 0.005	0.6635 ± 0.009	0.8741 ± 0.006
	ChemGPT Encoder	0.5247 ± 0.015	0.5812 ± 0.004	0.5854 ± 0.002	0.5866 ± 0.012
	No Tissue Encoder	0.8173 ± 0.007	0.6954 ± 0.008	0.6781 ± 0.013	0.8413 ± 0.002
	Non-Graph Encoder	0.8321 ± 0.011	0.7149 ± 0.005	0.6509 ± 0.002	0.8480 ± 0.007
	One-Hot Encoder	0.8568 ± 0.005	0.7425 ± 0.011	0.6862 ± 0.010	0.8649 ± 0.003
	Text-Based Encoder	0.8499 ± 0.002	0.7222 ± 0.005	0.6994 ± 0.007	0.8718 ± 0.012
	Mean Tissue Expression	0.8416 ± 0.006	0.7100 ± 0.001	0.6718 ± 0.014	0.8666 ± 0.009
	No Pretraining	0.8579 ± 0.013	0.7242 ± 0.001	0.6759 ± 0.012	0.8663 ± 0.004
	Only $\mathcal{L}_{\text{activity}}$	0.8557 ± 0.012	0.7147 ± 0.009	0.6707 ± 0.011	0.8743 ± 0.010
	$\mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{gene}}$	0.8695 ± 0.003	0.7322 ± 0.014	0.6738 ± 0.015	0.8812 ± 0.013
	$\mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{cls}}$	0.8592 ± 0.015	0.7132 ± 0.004	0.6733 ± 0.013	0.8743 ± 0.016
	$\mathcal{L}_{\text{activity}} + \mathcal{L}_{\text{pathway}}$	0.8695 ± 0.013	0.7200 ± 0.002	0.6786 ± 0.014	0.8731 ± 0.015
	Only Sample Nodes	0.8370 ± 0.007	0.7016 ± 0.014	0.6803 ± 0.016	0.8578 ± 0.005
	No Gene Nodes	0.8626 ± 0.016	0.7372 ± 0.003	0.6717 ± 0.004	0.8776 ± 0.016
	No Pathway Nodes	0.8596 ± 0.007	0.7328 ± 0.012	0.6559 ± 0.013	0.8733 ± 0.015

Table 9: Predicted activity scores of selected FDA-approved compounds across human tissues. Higher scores indicate higher predicted tissue-specific activity.

Compound	Brain	Breast	Cervix	Kidney	Liver	Lung	Ovary	Prostate	Skin
Sorafenib	0.4408	0.6120	0.3948	0.6911	0.7371	0.5680	0.6120	0.3108	0.6865
Fulvestrant	0.2134	0.8126	0.4457	0.4182	0.6202	0.6160	0.5804	0.8216	0.6743
Donepezil	0.9120	0.6809	0.3159	0.1524	0.3318	0.4215	0.3561	0.6914	0.5006
Talampanel	0.3284	0.3857	0.5027	0.4496	0.2774	0.1984	0.5392	0.3947	0.6563

Overall, these results indicate that Expresso effectively integrates molecular and tissue-specific embeddings to deliver accurate, well-calibrated predictions across diverse human tissues, with statistically robust gains over baseline models.

C In-Depth Analysis of FDA-Approved Compound Predictions

To further illustrate Expresso’s tissue-specific prioritization, we provide representative examples of four FDA-approved drugs, with predicted activity scores across human tissues summarized in Table 9.

Sorafenib, a multi-kinase inhibitor used primarily for liver cancer, shows the highest predicted activity in liver (0.7371), consistent with its main therapeutic target. Moderate predictions in kidney (0.6911) and lung (0.5680) may reflect off-target interactions, while lower scores in brain (0.4408) and prostate (0.3108) correspond to tissues of limited clinical relevance.

Fulvestrant, an estrogen receptor antagonist for breast cancer, is predicted most active in breast (0.8126), with elevated scores in prostate (0.8216) and ovary (0.5804) reflecting estrogen receptor expression in reproductive tissues. Moderate activity in liver (0.6202) and lung (0.6160) may capture metabolic or off-target effects, and low activity in brain (0.2134) aligns with limited central nervous system exposure.

Table 10: ROC-AUC of Expresso for each tissue under increasing levels of random noise applied to pathway features. Higher values indicate better predictive performance. This table allows direct comparison of model robustness across tissues as pathway information is progressively corrupted.

Noise %	Brain	Breast	Cervix	Kidney	Liver	Lung	Ovary	Prostate	Skin
0	0.6471	0.8233	0.9325	0.8764	0.8207	0.8763	0.7466	0.6953	0.8950
20	0.6050	0.8020	0.8923	0.8369	0.7886	0.8507	0.7103	0.6974	0.8628
40	0.5882	0.8002	0.8932	0.8350	0.7889	0.8508	0.7130	0.6903	0.8639
60	0.5714	0.7901	0.8931	0.8236	0.7811	0.8334	0.7158	0.6871	0.8346

Donepezil, a CNS-active acetylcholinesterase inhibitor for Alzheimer’s disease, is strongly predicted in brain (0.9120). Moderate scores in prostate (0.6914) and breast (0.6809) may reflect minor pathway-level effects, while low activity in kidney (0.1524), liver (0.3318), and cervix (0.3159) corresponds to tissues with little therapeutic relevance.

Talampanel, an AMPA receptor antagonist explored for neurological conditions, shows modest predicted activity in brain (0.3284), consistent with its limited efficacy. Higher predictions in skin (0.6563) and ovary (0.5392) likely reflect off-target effects or tissue-specific pathway correlations, rather than direct clinical relevance.

These results illustrate that Expresso effectively identifies the primary target tissues for each compound while reflecting plausible secondary tissue interactions. Predicted activity patterns are broadly consistent with known pharmacology, demonstrating the model’s capacity to integrate molecular structure with tissue-specific biological context.

D Assessing Out-of-Distribution Effects in Pathway Ablation

To address concerns that pathway ablation could produce out-of-distribution (OOD) inputs for Expresso, we systematically evaluated the OOD likelihood of ablated graphs and their potential impact on interpretability. Ablated graphs were created by removing individual pathway groups, as in the main pathway-level analysis, and then processed through Expresso’s hierarchical tissue graph encoder to obtain fixed-length embeddings that capture tissue-specific and functional context. To assess whether these perturbed graphs remain within the distribution of training data, we trained a binary discriminator to distinguish in-distribution tissue graphs from held-out tissue graphs not seen during training. The discriminator was a two-layer multilayer perceptron with ReLU activations and dropout, operating on the embeddings produced by Expresso’s graph encoder. Original training graphs were labeled as in-distribution, while held-out tissue graphs were labeled as OOD.

When applied to the ablated graphs, the discriminator classified 91% of them as in-distribution, indicating that pathway ablation generally produces perturbations within the model’s learned distribution. This suggests that the observed tissue-specific importance patterns in the ablation study are not significantly confounded by OOD effects and that the pathway knockouts provide meaningful insight into the model’s learned functional dependencies. Overall, these results support the validity of the ablation-based interpretability analysis and demonstrate that the generated perturbations preserve the biological plausibility captured by Expresso’s tissue graph representations.

E Pathway Network Noise Robustness

To evaluate how sensitive Expresso is to perturbations in the curated biological priors, we systematically introduced random noise into the pathway–pathway crosstalk matrix A_{pp} . At each noise level k , a fraction k of matrix entries was flipped—transforming existing edges to non-edges and vice versa—to emulate uncertainty, incompleteness, or potential annotation errors in biological pathway databases.

Results in Table 10 demonstrate that Expresso maintains stable performance even under substantial perturbations, with AUC values decreasing only marginally up to $k = 0.4$ and remaining above 0.78 on average at

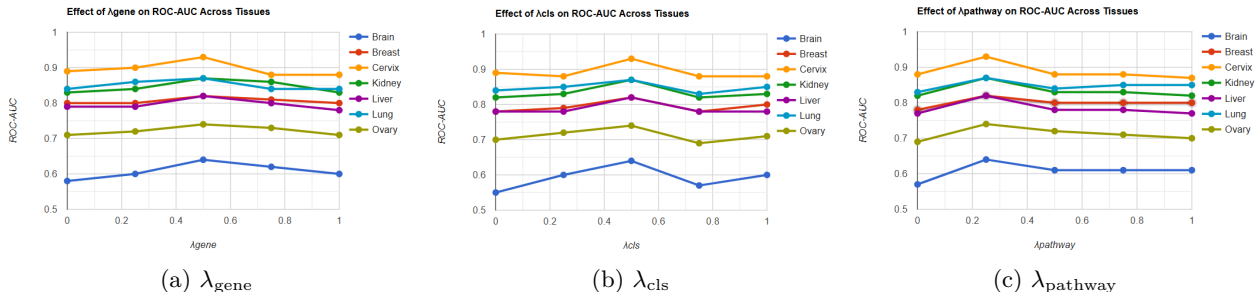


Figure 3: Sensitivity of Expresso’s ROC-AUC performance to loss weighting coefficients. Panels show ROC-AUC across tissues when varying λ_{gene} , λ_{cls} , or λ_{pathway} , with all other loss terms held constant.

Table 11: Leave-one-tissue-out evaluation of Expresso for molecular activity prediction across human tissues. Values report ROC-AUC for each tissue, indicating model predictive performance when the corresponding tissue is excluded from training.

Model	Brain	Breast	Cervix	Kidney	Liver	Lung	Ovary	Prostate	Skin
Expresso	0.6471	0.8233	0.9325	0.8764	0.8207	0.8763	0.7466	0.6953	0.8950
Leave One Tissue Out	0.6176	0.7807	0.8797	0.8157	0.7693	0.8363	0.6966	0.6753	0.8421

$k = 0.6$. This robustness reflects the model’s capacity to integrate redundant biological information across multiple graph relations (sample–gene, gene–pathway, and pathway–pathway), rather than relying on any single network structure.

Overall, these findings indicate that Expresso learns biologically grounded representations that generalize beyond specific pathway annotations. Its resilience to structural noise suggests strong internal regularization and biological consistency, both essential for deployment in real-world settings where pathway resources are inherently noisy and incomplete.

F Sensitivity Analysis of Loss Weighting

We assess the robustness of Expresso to the choice of auxiliary loss weights λ_{gene} , λ_{cls} , and λ_{pathway} via the sensitivity analysis shown in Figure 3. In each experiment, one loss weight is varied while all other terms in the objective are held fixed, and performance is evaluated using ROC-AUC across tissues.

Performance changes smoothly over a wide range of values for each coefficient, with no sharp degradation under moderate deviations from the selected defaults. This indicates that model performance is not highly sensitive to precise tuning of individual auxiliary loss weights.

For most tissues, a noticeable drop in performance is observed when a given loss weight is set to zero, suggesting that each auxiliary objective provides useful training signal even when assigned a small weight. Overall, the qualitative trends are consistent across tissues, indicating that the effects of loss weighting reflect general model behavior rather than tissue-specific artifacts.

G Tissue Scalability

The proposed framework constructs a separate heterogeneous graph for each tissue, enabling independent encoding of tissue-specific molecular and functional features, which allows the method to naturally scale to additional tissue types or cell lines without changes to the model architecture. In our experiments, we used all available data for the tissues under study. While larger datasets would increase computational and memory requirements, the model design does not impose fundamental limitations on scalability.

Table 12: Structural complexity of tissue-specific heterogeneous graphs used by Expresso.

Model	Brain	Breast	Cervix	Kidney	Liver	Lung	Ovary	Prostate	Skin
# Nodes	13312	13571	13080	13161	13319	13661	13250	13339	13708
# Edges	6512459	12500539	1148619	3021339	6674299	14581339	5079019	7136699	15667979
# Samples	255	514	23	104	262	604	193	282	651
# Genes	11560	11560	11560	11560	11560	11560	11560	11560	11560
# Pathways	1497	1497	1497	1497	1497	1497	1497	1497	1497
# $\mathcal{E}_{s \leftrightarrow g}$	2947800	5941840	265880	1202240	3028720	6982240	2231080	3259920	7525560
# $\mathcal{E}_{g \leftrightarrow p}$	86485	86485	86485	86485	86485	86485	86485	86485	86485
# $\mathcal{E}_{p \leftrightarrow p}$	443889	443889	443889	443889	443889	443889	443889	443889	443889

To further assess the model’s generalization to unseen biological contexts, we conducted a leave-one-tissue-out evaluation. In this setup, Expresso was pretrained and fine-tuned on all but one tissue, and subsequently evaluated on the held-out tissue without any additional retraining. This setting mimics the realistic scenario of predicting molecular activity in a new tissue for which no training data are available.

Results summarized in Table 11 demonstrate that Expresso maintains strong cross-tissue generalization. While performance generally decreases compared to in-tissue training, the model preserves competitive ROC-AUC scores across most tissues (e.g., 0.8363 in lung, 0.8157 in kidney, 0.8421 in skin), indicating robust transferability of learned molecular and functional representations. These findings support the claim that Expresso can be effectively extended to novel tissues with minimal or no re-training, highlighting its scalability and biological adaptability.

H Model Complexity

To examine the structural scale of the proposed framework, we quantified the composition of each tissue-specific heterogeneous graph used by Expresso. As summarized in Table 12, each graph integrates multiple biological layers connected through biologically grounded relations. While all tissues share a common gene and pathway backbone (11,560 genes and 1,497 pathways), the overall graph size varies substantially due to differences in sample availability and tissue-specific connectivity patterns. For instance, lung and skin exhibit the largest graphs (over 13,000 nodes and 15 million edges), reflecting their larger sample cohorts and denser molecular associations, whereas cervix shows a markedly smaller structure due to limited data. These observations demonstrate that Expresso effectively scales across tissues of diverse structural complexity without requiring architecture modifications.

I Limitations

Although we report performance at the level of *tissues*, our model currently learns fixed tissue-level embeddings from bulk GTEx profiles and does not condition on donor-specific expression at inference time. This design simplifies data requirements and improves robustness, but may under-capture within-tissue heterogeneity, batch effects, and microenvironmental influences. We partially mitigate these issues by using (i) global compound (scaffold) splits across all tissues, (ii) donor-wise splits during pretraining, and (iii) leave-one-tissue-out evaluations; nonetheless, extending the framework to incorporate cell-type-resolved expression and to assess domain shift to independent expression resources remains important future work.

In addition, many of the tissue labels in our pharmacology panel are implemented as single in vitro assays on cancer-derived or immortalized cell lines. Consequently, our model captures *assay- and cell-type-specific tissue activity* defined by these systems, rather than the full physiology of healthy human tissues as profiled in GTEx. While cancer cell lines are widely used surrogates in preclinical drug discovery, they can differ from primary tissues in their molecular and pathway-level signatures (Jiang et al., 2016), and our predictions should therefore be interpreted as preclinical, assay-defined tissue-context signals for screening and hypothesis generation, not as direct surrogates for clinical efficacy or safety in patients. Bridging these in vitro readouts to healthy-tissue physiology and in vivo outcomes is outside the scope of this work and an important direction for future research.