SCOR: A SECTION-CENTERED OBJECTIVE FOR GENE EXPRESSION PREDICTION FROM HISTOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting spatial gene expression from hematoxylin and eosin (H&E) histology is typically optimized with spot-wise regression, even though many spatial observations are grouped within the same tissue section. Viewing the task as grouped regression, conventional mean squared error (MSE) decomposes exactly into within-section residual variation and section-mean residual error. This exposes an implicit objective-design choice: MSE assigns equal weight to error components defined at different levels of the grouped data structure. We propose the Section-Centered Objective for Regression (SCOR), which retains full weight on within-section residual variation while explicitly controlling the contribution of the section-mean residual component, without changing the histology representation, predictor architecture, or inference procedure. SCOR further admits a profiled regularized section-offset interpretation that gives its weighting parameter a direct shrinkage meaning. Controlled diagnostics show that preserving the section-defined grouping is important for reproducing the benefit: pairwise residual differencing alone is insufficient, and the effect persists beyond local spatial proximity. A fixed-sample analysis further shows that training-section replication modulates its magnitude. With a single fixed objective setting, SCOR improves over MSE on the majority of eight HEST-1K tasks, transfers the same setting without retuning to HER2ST and STNet, and extends to ridge regression. These results identify grouped-error weighting as an overlooked design choice in histology-to-transcriptomics regression.

1 INTRODUCTION

Spatial transcriptomics (ST) measures gene expression while preserving spatial location within tissue, enabling molecular organization to be studied in direct relation to histological morphology (Yoshikawa et al., 2026; Chelebian et al., 2025; Gulati et al., 2025). Hematoxylin and eosin (H&E) images are inexpensive, routinely acquired, and available at substantially greater scale than spatial molecular assays (Xu et al., 2026; Vorontsov et al., 2026). Predicting spatial gene expression directly from histology has therefore emerged as an important problem in computational pathology (Wang et al., 2025). A defining feature of this prediction setting, however, is that the training observations are grouped: many spatial spots are observed from the same tissue section. Conventional spot-wise regression does not explicitly distinguish prediction error shared across a section from error that varies among spots within that section. This raises a basic question that is largely independent of model architecture: how should the grouping structure of the training data influence the regression objective?

Most progress in H&E-to-ST prediction has focused on improving the information represented by the predictor, including spatial dependencies and graph structure (Pang et al., 2021; Zeng et al., 2022; Yang et al., 2023; 2024), multiscale and broader tissue context (Jia et al., 2024; Chung et al., 2024; Ganguly et al., 2025; Jeong et al., 2026), and cross-modal or generative modeling (Xie et al., 2023; Huang et al., 2025). Recent work has also begun to reconsider the supervision itself. MISO (Schmauch et al., 2025) uses within-sample cosine-based supervision to reduce sensitivity to inter-sample variability, while COAST (Byeon et al., 2026) combines absolute regression with signed differential regression between target and context spots. STRank (Nishimura et al., 2026) learns relative expression ordering rather than absolute expression levels. These methods modify supervision through similarity, differential, or ranking relations. In contrast, SCOR retains squared regression

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and asks how its error is allocated across the within-section and section-mean components induced by the grouped data structure. This distinction exposes a more basic question: what weighting of grouped prediction error is already implicit in conventional squared regression itself?

We revisit mean squared error (MSE) from this perspective. Because spots are grouped within tissue sections, squared prediction error admits an exact decomposition into a within-section component, which measures residual variation among spots after removing the section mean, and a section-mean component, which measures the squared residual shared at the section level. Aggregated across sections, conventional MSE is exactly the sum of these two components. This identity reveals an implicit objective-design choice: ordinary spot-wise MSE assigns the same coefficient to within-section residual variation and section-mean residual error, although they describe prediction error at different levels of the data hierarchy. Importantly, section-mean residual structure should not be interpreted as purely technical noise, since aggregate expression can itself contain histologically predictable biological signal (Schmauch et al., 2020). The issue is therefore not that section-level error should be removed, but that its relative influence on learning is fixed implicitly by MSE.

We introduce the Section-Centered Objective for Regression (SCOR), which exposes this implicit weighting and makes it controllable. SCOR retains within-section residual variation at full weight while assigning a coefficient $c \in [0, 1]$ to the section-mean residual component, continuously interpolating between conventional MSE and complete section centering. Tissue-section identity is required only during training to construct the objective; SCOR introduces no predictor parameters and requires no change to the histology representation, prediction architecture, or inference procedure. Moreover, the formulation is not merely an ad hoc reweighting. For $0 \leq c < 1$, SCOR is exactly equivalent to profiling ridge-regularized section-specific residual offsets, with $c = \alpha/(1+\alpha)$, giving the objective weight a direct shrinkage interpretation; conventional MSE at $c = 1$ is recovered in the limiting case $\alpha \to \infty$. SCOR therefore controls how strongly the section-mean residual component influences parameter learning without assuming that this component is biologically irrelevant.

The formulation leads to several testable predictions. First, if the section-defined grouping used by SCOR is important, a matched pairwise objective that preserves same-section relationships should reproduce its benefit more closely than one that replaces them with cross-section relationships. Second, if the effect can be explained solely by local spatial proximity, the benefit should diminish substantially when same-section partners are spatially far apart. Third, if repeated group structure is important for learning, SCOR's advantage should vary with the amount of training-section replication even when the total number of training spots is held fixed. We test these predictions alongside benchmark evaluation using ridge regression and multilayer perceptrons across eight HEST-1K tasks (Jaume et al., 2024) and the independent HER2ST (Andersson et al., 2021) and STNet (He et al., 2020) datasets. A single objective setting selected on one development task is held fixed across the remaining prediction tasks and datasets. Together, these experiments test not only whether SCOR improves prediction, but whether its behavior follows the grouped-error structure from which it is derived. Our contributions are:

- We identify an overlooked objective-design implication of the grouped structure of H&E-to-ST regression: conventional spot-wise MSE exactly combines within-section residual variation and section-mean residual error with a fixed relative weighting. We introduce SCOR to make this weighting explicit and controllable without modifying the representation, predictor, or inference procedure.

- We provide complementary interpretations and analyses of SCOR through profiled regularized section-specific residual offsets, its stochastic minibatch behavior, and a same-section pairwise representation that recovers SCOR in expectation at $\lambda = 1 - c$. These results characterize what SCOR changes relative to conventional MSE rather than treating it as an ad hoc alternative loss.

- We evaluate SCOR across ten prediction tasks from the benchmark datasets HEST-1K, HER2ST, and STNet and across linear and nonlinear predictors. Controlled experiments show that preserving the section-defined training grouping is important for reproducing the gain under the matched diagnostic, that the effect persists beyond local spatial neighborhoods, and that its magnitude is modulated by training-section replication, while held-out total MSE generally does not increase when correlation improves.

2 PROPOSED METHOD

We formulate H&E-to-ST prediction as grouped regression, where multiple spatial observations are nested within each tissue section. We first show that conventional spot-wise MSE implicitly fixes the relative weighting of within-section and section-mean residual errors. We then introduce SCOR to make this weighting explicit and describe its full-section and minibatch forms. Finally, we provide complementary section-offset and pairwise interpretations that motivate the controlled experiments in Section 3.2.

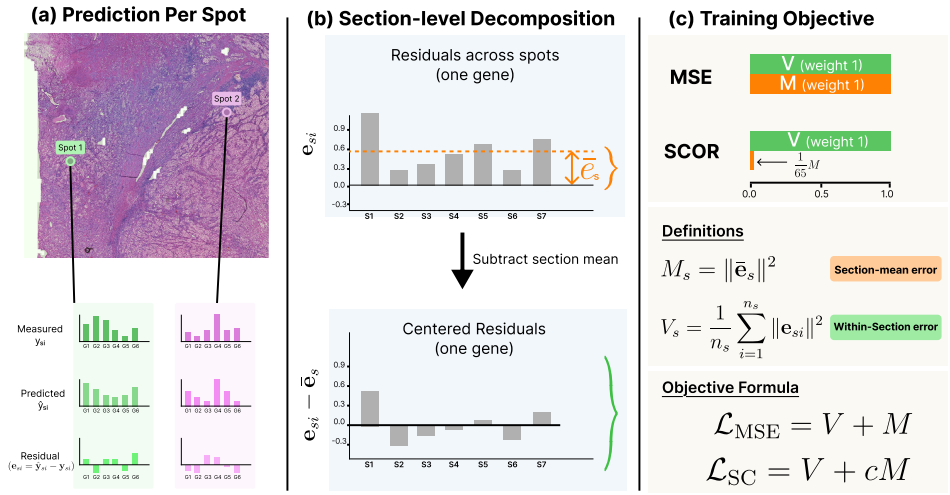


Figure 1: Overview of the proposed section-centered regression formulation. (a) For each spatial spot within a tissue section, the prediction residual is defined as the difference between predicted and measured gene expression. (b) The squared prediction error within a section decomposes exactly into a within-section component V_s, obtained after removing the section-mean residual, and a section-mean component M_s. (c) Conventional MSE assigns unit weight to both components, whereas SCOR retains full weight on V and scales the contribution of M by $c \in [0, 1]$.

2.1 GROUPED REGRESSION AND ERROR DECOMPOSITION

Let $s \in \{1, \dots, S\}$ index tissue sections and $i \in \{1, \dots, n_s\}$ index spots within section s. Each spot has a histology-derived representation $\mathbf{x}_{si} \in \mathbb{R}^d$ and a gene-expression target $\mathbf{y}_{si} \in \mathbb{R}^G$. A predictor $f_\theta : \mathbb{R}^d \to \mathbb{R}^G$ produces $\hat{\mathbf{y}}_{si} = f_\theta(\mathbf{x}_{si})$, with residual $\mathbf{e}_{si} = \hat{\mathbf{y}}_{si} - \mathbf{y}_{si}$. Conventional spot-wise mean squared error is

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \|\mathbf{e}_{si}\|^2, \qquad N = \sum_{s=1}^{S} n_s. \tag{1}$$

Because the observations are grouped within tissue sections, the squared error contributed by section s admits the exact decomposition

$$\frac{1}{n_s} \sum_{i=1}^{n_s} \|\mathbf{e}_{si}\|^2 = \underbrace{\frac{1}{n_s} \sum_{i=1}^{n_s} \|\mathbf{e}_{si} - \bar{\mathbf{e}}_s\|^2}_{V_s} + \underbrace{\|\bar{\mathbf{e}}_s\|^2}_{M_s}, \tag{2}$$

where $\bar{\mathbf{e}}_s = \frac{1}{n_s} \sum_i \mathbf{e}_{si}$. Here, V_s measures residual variation among spots after removing the section mean, whereas M_s measures the squared section-mean residual.

To preserve the spot weighting of Eq. 1 when aggregating across sections, we define

$$V = \frac{1}{N} \sum_s n_s V_s, \qquad M = \frac{1}{N} \sum_s n_s M_s.$$

Eq. 2 then gives

$$\mathcal{L}_{\mathrm{MSE}} = V + M$$

exactly. This identity exposes an implicit objective choice in conventional spot-wise regression: MSE assigns equal coefficients to within-section residual variation and section-mean residual error even though they describe prediction error at different levels of the grouped data structure.

The section-mean component should not be interpreted as purely technical nuisance. Aggregate gene expression can itself contain histologically predictable signal (Schmauch et al., 2020), so M may contain both biologically informative and other section-level variation. Our objective therefore does not assume that section-level residual structure should be removed; instead, it makes its relative influence on training explicit.

2.2 Section-Centered Objective

Motivated by the decomposition above, we define the Section-Centered Objective for Regression (SCOR) as

$$\mathcal{L}_{\mathrm{SC}} = V + cM, \qquad 0 \le c \le 1. \tag{3}$$

When $c = 1$, Eq. 3 reduces exactly to conventional MSE. Decreasing c progressively reduces the contribution of the section-mean residual while retaining the full within-section component. At $c = 0$, the objective depends only on within-section residual variation and is invariant to a common residual offset shared by all spots in a section. SCOR therefore interpolates continuously between conventional spot-wise regression and complete section centering without changing the histology representation, predictor architecture, or inference procedure.

For predictors trained over complete sections, including the ridge-regression setting considered here, Eq. 3 can be evaluated directly. For stochastic predictors, a complete section is generally not represented in each optimization step. Let $\mathcal{B}$ denote a minibatch, $\mathcal{B}_s \subseteq \mathcal{B}$ the spots from section s, $m_s = |\mathcal{B}_s|$, and

$$\bar{\mathbf{e}}_{s,\mathcal{B}} = \frac{1}{m_s} \sum_{i \in \mathcal{B}_s} \mathbf{e}_{si}.$$

The minibatch form of SCOR is

$$\mathcal{L}_{\mathrm{SC}}^{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{s} \left[\sum_{i \in \mathcal{B}_s} \|\mathbf{e}_{si} - \bar{\mathbf{e}}_{s,\mathcal{B}}\|^2 + c\, m_s \|\bar{\mathbf{e}}_{s,\mathcal{B}}\|^2 \right]. \tag{4}$$

When $c = 1$, Eq. 4 reduces exactly to conventional minibatch MSE. For $c < 1$, estimating the section mean from sampled spots makes the stochastic objective an approximation to the full-section objective. Appendix A.1 shows that, conditional on the number m_s of sampled spots from a section, the expected coefficient on the section-mean component remains exactly c, while the effective coefficient on the within-section component approaches one as more spots from the same section are represented in the minibatch; the exact full-section objective is recovered when the complete section is included.

2.3 Interpretations and Testable Implications

Profiled section-offset interpretation. SCOR admits an equivalent interpretation through ridge-penalized section-specific residual offsets. Let $\mathbf{b}_s \in \mathbb{R}^G$ denote an offset associated with section s, and consider

$$\mathcal{J}(\theta, \{\mathbf{b}_s\}) = \frac{1}{N} \sum_{s} \sum_{i} \|\mathbf{e}_{si} + \mathbf{b}_s\|^2 + \frac{\alpha}{N} \sum_{s} n_s \|\mathbf{b}_s\|^2, \qquad \alpha \ge 0. \tag{5}$$

Optimizing Eq. 5 with respect to each $\mathbf{b}_s$ gives

$$\mathbf{b}_s^* = -\frac{1}{1+\alpha} \bar{\mathbf{e}}_s, \tag{6}$$

and profiling out these offsets yields

$$\min_{\{\mathbf{b}_s\}} \mathcal{J} = V + \frac{\alpha}{1+\alpha} M. \tag{7}$$

For any finite $\alpha \geq 0$, profiling the offsets therefore yields SCOR with

$$c = \frac{\alpha}{1+\alpha}. \tag{8}$$

so that $0 \leq c < 1$. Conversely, every $c \in [0, 1)$ corresponds to $\alpha = c/(1-c)$, while conventional MSE at $c = 1$ is recovered in the limiting case $\alpha \to \infty$. Equivalently, $\mathbf{b}_s^* = -(1-c)\bar{\mathbf{e}}_s$, so $1-c$ can be interpreted as the fraction of the section-mean residual absorbed by the profiled offset. The offsets are introduced only for this interpretation and are analytically eliminated; they are not predictor parameters and are not required at inference. The complete derivation is given in Appendix A.2.

Same-section pairwise interpretation. The within-section component also admits a pairwise representation. For spots sampled from the same section within a minibatch, an appropriately finite-sample-corrected residual-difference term provides an unbiased estimator of the corresponding within-section component. Let $\widetilde{\mathcal{P}}^{\mathcal{B}}_{\text{same}}$ denote this corrected same-section estimator and define

$$\mathcal{L}^{\mathcal{B}}_{\text{pair}}(\lambda) = (1-\lambda)\mathcal{L}^{\mathcal{B}}_{\text{MSE}} + \lambda\widetilde{\mathcal{P}}^{\mathcal{B}}_{\text{same}}, \qquad 0 \leq \lambda \leq 1. \tag{9}$$

As shown in Appendix A.3,

$$\mathbb{E}\left[\mathcal{L}^{\mathcal{B}}_{\text{pair}}(\lambda)\right] = V_{\mathcal{B}} + (1-\lambda)M_{\mathcal{B}}.$$

Setting $\lambda = 1 - c$ therefore recovers minibatch SCOR in expectation. This view is used only as a diagnostic tool: SCOR itself is implemented directly through centered residuals and does not require explicit spot pairs.

These interpretations yield three empirical implications. First, if the section-defined grouping used by SCOR is important, preserving same-section relationships in the pairwise diagnostic should behave differently from a matched control in which partners are drawn across sections. Second, if the observed benefit is explained solely by local spatial proximity, it should diminish substantially as same-section partners become spatially farther apart. Third, if replicated grouping structure is important, the advantage of SCOR should vary with the number of represented training sections even when the total number of training spots is held fixed. Sections 3.2 and 3.3 test these implications under controlled settings.

3 Experimental Setup

We evaluate SCOR around three questions: whether it improves prediction across tasks and datasets, whether its behavior depends on preserving the section-defined grouping used by the objective rather than on pairwise differencing or local spatial proximity alone, and what data conditions modulate its effect. All objective comparisons are conducted under matched representations, predictors, data splits, and evaluation procedures unless explicitly stated otherwise.

Common setup. The evaluation comprises ten prediction tasks across three datasets. HEST-1K (Jaume et al., 2024) contributes eight tasks, including ccRCC, IDC, LYMPH, PAAD, PRAD, READ, SKCM, and LUAD, while HER2ST (Andersson et al., 2021) and STNet (He et al., 2020) provide two independent breast-cancer evaluation settings. The primary experiments use frozen UNI2-h (Chen et al., 2024) representations extracted from spatially aligned H&E patches. The resulting 1536-dimensional features are standardized and projected to 256 dimensions using PCA, which is applied uniformly across predictor families as an experimental control rather than as a requirement of SCOR. We use an MLP as the primary nonlinear predictor and ridge regression as a controlled linear setting. HEST-1K retains its benchmark-defined gene panels and log-transformed observed counts without library normalization, whereas HER2ST and STNet retain their established 250-gene panels with expression library-normalized within the selected panel at each spot and then log transformed. Dataset characteristics and complete training settings are provided in Table 4 and Table 3, respectively.

We report pooled and section-level Pearson correlation coefficient (PCC). Pooled PCC concatenates all held-out spots within a fold before computing and averaging gene-wise correlations, and therefore reflects both within- and between-section expression variation when multiple sections are present. Section-level PCC instead computes gene-wise correlation within each held-out section before averaging across sections, thereby removing between-section variation from the correlation

calculation and measuring recovery of expression variation among spots within an unseen section. For folds containing a single held-out section, the two measures are identical. Throughout the paper, ΔPCC denotes an absolute PCC difference, $\mathrm{PCC}_A - \mathrm{PCC}_B$, whereas percentage improvements reported for the grouping and spatial diagnostics denote relative changes, computed as $100 \times (\mathrm{PCC}_A - \mathrm{PCC}_B)/\mathrm{PCC}_B$. Neural experiments use fixed random seeds. Preprocessing parameters are fit using training data only, while checkpoint selection and early stopping use validation data; test observations are excluded from all preprocessing and model selection.

3.1 Benchmark and Cross-Dataset Evaluation

We use ccRCC as the sole development task for selecting the SCOR objective setting because its folds provide substantial training-section replication and multiple held-out sections, allowing both pooled and section-level validation PCC to be informative. We sweep the profiled-offset parameter over $\alpha \in \left\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}\right\}$ and inspect both validation metrics across the sweep. As α decreases, section-level validation PCC reaches a plateau around $\alpha = 1/64$, while pooled validation PCC continues to improve. Based on this joint validation behavior, we select $\alpha = 1/64$, corresponding through Eq. 8 to $c = 1/65$. This setting is then fixed before evaluation on all remaining prediction tasks and datasets.

For the primary MLP benchmark, we apply the fixed SCOR setting to all eight HEST-1K tasks using three fixed random seeds and the frozen benchmark folds. We compare against conventional MSE, direct PCC optimization, MISO (Schmauch et al., 2025), PairSTRank (Nishimura et al., 2026), and ListSTRank (Nishimura et al., 2026). All objectives use the same histology representations, prediction targets, data splits, predictor architecture, preprocessing, optimization budget, and evaluation procedure. Baseline-specific objective definitions and hyperparameters follow their published formulations or official implementations where available, and no test-set performance is used for objective-specific model selection.

HER2ST and STNet are reserved for cross-dataset evaluation and do not contribute to objective-setting selection. We apply the same fixed SCOR objective setting without retuning c, while preserving each dataset's native target construction and frozen patient-grouped folds. This provides a zero-retuning test of whether the objective setting selected on ccRCC transfers across datasets with different tissue composition, target construction, and hierarchical organization.

3.2 Section-Defined Grouping and Spatial Controls

Following the pairwise interpretation in Section 2.3, we construct matched ccRCC diagnostics that preserve pairwise residual differencing while varying whether SCOR's section-defined grouping is retained. Each eligible anchor is paired with eight partners sampled independently with replacement from the current minibatch, excluding the anchor. Same-section partners are drawn from the anchor's section, whereas cross-section partners come from a different section. The two arms otherwise share the same anchors, predictor, initialization, minibatches, optimization budget, and number of partners, so at a fixed pairwise weight λ they differ only in the grouping relation. The cross-section arm is therefore a matched control that breaks the section-defined grouping rather than an equivalent form of SCOR. The diagnostic objective follows Eq. 9. For same-section pairing, the theoretical correspondence to SCOR fixes $\lambda_{\mathrm{same}} = 1 - c = 64/65$ for $c = 1/65$. Because this equivalence does not hold across sections, we select λ_{cross} independently using section-level validation PCC over $\phi \in \left\{\frac{1}{64}, \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, 32\right\}$, with $\lambda = 1/(1+\phi)$, yielding $\lambda_{\mathrm{cross}} = 4/5$. Test performance is evaluated over five fixed random seeds using pooled and section-level PCC.

To test whether same-section behavior is explained by local spatial proximity, we partition candidate partners within each training section into near, intermediate, and far distance tertiles. We retain anchors with at least one partner in every tertile and use the same anchor set across conditions. All three conditions use the fixed $\lambda_{\mathrm{same}} = 64/65$ and otherwise matched training settings, differing only in partner distance. We compare pooled and section-level PCC across the three distance conditions and against conventional MSE.

3.3 Error Characterization and Section Replication

To characterize the error component reweighted by SCOR, we decompose held-out residuals from the MSE-trained predictor into V and M and define the section-mean error share as $R_M = M/(V + M)$. We also compare held-out V, M, and total MSE between MSE and SCOR to assess whether correlation changes are accompanied by changes in within-section residual variation or absolute prediction error. We test training-section replication on ccRCC by varying the number of represented sections over $K \in \{1, 2, 4, 8, 16\}$ while holding the number of sampled training spots fixed within each fold–seed combination. Spots are distributed equally across the selected sections, with MSE and SCOR using identical observations at each K. We evaluate batch sizes of 256 and 512 and full-batch training, allowing comparison between minibatch SCOR, where section means are estimated from sampled observations, and the corresponding complete-sample regime.

4 Results

4.1 Predictive Performance Across Tasks and Datasets

Across the eight HEST-1K tasks, SCOR achieves higher mean pooled PCC than conventional MSE on five tasks, matches MSE on LYMPH to the reported precision, and is lower on SKCM and LUAD (Table 1 and Fig. 2). The same task-wise pattern is observed under section-level evaluation. Among the evaluated objectives under the common benchmark protocol, SCOR obtains the highest unweighted average pooled PCC, with an average of 0.4488 compared with 0.4456 for MISO and 0.4442 for conventional MSE. SCOR also achieves the highest task-level mean pooled PCC on five of the eight HEST-1K tasks. Notably, both tasks on which SCOR is lower than MSE are in the single-training-section regime, motivating the controlled section-replication analysis in Section 4.3. Because STRank was developed under a different target and gene-selection regime, PairSTRank and ListSTRank in Table 1 are evaluated under the common HEST-1K protocol to isolate objective differences. Appendix E additionally provides a ccRCC implementation sanity check under the original CONCH-v1 (Lu et al., 2024) setting and evaluates all objectives under a STRank-aligned regime. Ranking supervision becomes substantially more competitive under Spearman evaluation, while SCOR remains above MSE on seven of eight tasks under both Spearman and Pearson correlation and achieves the highest mean under both metrics.

Table 1: MLP pooled PCC across eight HEST-1K tasks. Values are mean ± standard deviation across seeds. Avg. is the unweighted mean of the eight task-level means. Best and second-best results are **bold** and underlined, respectively.

Objective	ccRCC	IDC	LYMPH	PAAD	PRAD	READ	SKCM	LUAD	Avg.
MSE	$0.2601_{\pm 0.0082}$	$0.5867_{\pm 0.0027}$	$0.2626_{\pm 0.0039}$	$\underline{0.5132}_{\pm 0.0055}$	$0.3872_{\pm 0.0092}$	$0.2453_{\pm 0.0098}$	$\mathbf{0.6985}_{\pm 0.0074}$	$\mathbf{0.6000}_{\pm 0.0019}$	0.4442
MISO	$\underline{0.2632}_{\pm 0.0197}$	$\underline{0.5891}_{\pm 0.0012}$	$\mathbf{0.3014}_{\pm 0.0010}$	$0.5039_{\pm 0.0057}$	$0.3719_{\pm 0.0259}$	$\underline{0.2500}_{\pm 0.0101}$	$\underline{0.6962}_{\pm 0.0085}$	$0.5890_{\pm 0.0032}$	$\underline{0.4456}$
Direct PCC	$0.2508_{\pm 0.0243}$	$0.5809_{\pm 0.0031}$	$0.2656_{\pm 0.0037}$	$0.5096_{\pm 0.0019}$	$\underline{0.3933}_{\pm 0.0174}$	$0.2462_{\pm 0.0091}$	$0.6913_{\pm 0.0109}$	$0.5772_{\pm 0.0041}$	0.4393
PairSTRank	$0.2162_{\pm 0.0193}$	$0.3825_{\pm 0.0068}$	$\underline{0.2681}_{\pm 0.0039}$	$0.3722_{\pm 0.0019}$	$0.3107_{\pm 0.0128}$	$0.1961_{\pm 0.0029}$	$0.4290_{\pm 0.0044}$	$0.3746_{\pm 0.0136}$	0.3187
ListSTRank	$0.1057_{\pm 0.0091}$	$0.3108_{\pm 0.0068}$	$0.1307_{\pm 0.0018}$	$0.3127_{\pm 0.0069}$	$0.1575_{\pm 0.0137}$	$0.1926_{\pm 0.0040}$	$0.3999_{\pm 0.0141}$	$0.2702_{\pm 0.0107}$	0.2350
SCOR	$\mathbf{0.2738}_{\pm 0.0078}$	$\mathbf{0.5967}_{\pm 0.0029}$	$0.2626_{\pm 0.0034}$	$\mathbf{0.5206}_{\pm 0.0045}$	$\mathbf{0.3973}_{\pm 0.0088}$	$\mathbf{0.2511}_{\pm 0.0153}$	$0.6890_{\pm 0.0113}$	$\underline{0.5989}_{\pm 0.0020}$	$\mathbf{0.4488}$

The fixed SCOR setting also transfers to the independent HER2ST and STNet datasets without retuning, improving both pooled and section-level PCC over conventional MSE on each dataset (Fig. 2 and Table 2). On HER2ST, SCOR achieves the highest pooled PCC among the evaluated objectives, while MISO is marginally higher in section-level PCC. On STNet, SCOR achieves the highest value under both metrics. The gains are positive across all three fixed random seeds despite differences in target construction and dataset organization from HEST-1K.

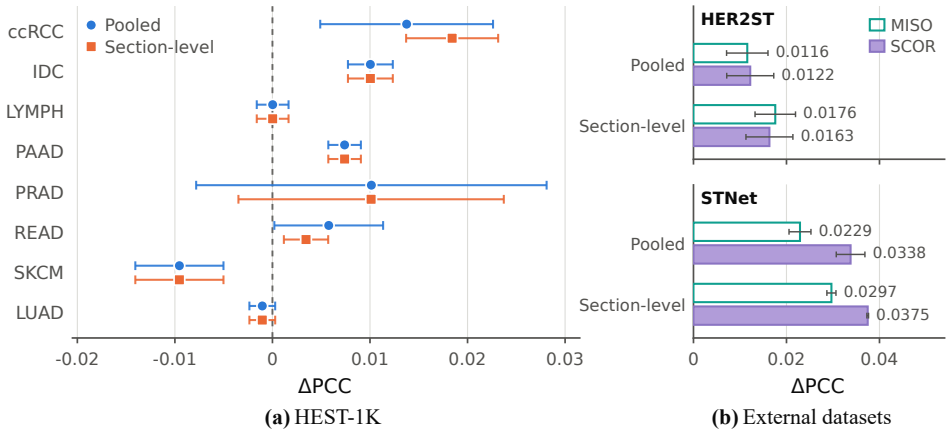
The effect is not confined to the nonlinear MLP predictor. Under ridge regression, SCOR improves section-level PCC over conventional MSE on eight of the ten prediction tasks and pooled PCC on seven based on the underlying unrounded results (Appendix Table 6), providing a controlled linear setting in which only the objective changes.

4.2 Dependence on the Section-Defined Grouping

The ccRCC pairwise diagnostic shows that preserving the section-defined grouping used by SCOR matters under the matched residual-difference formulation. Relative to the cross-section control,

Table 2: MLP cross-dataset evaluation on HER2ST and STNet using the SCOR setting selected on ccRCC without further tuning. Values are mean ± standard deviation across seeds. Best and second-best results are **bold** and underlined.

Objective	HER2ST		STNet	
	Pooled	Section-level	Pooled	Section-level
MSE	0.2191±0.0051	0.1624±0.0043	0.1204±0.0002	0.0826±0.0010
MISO	0.2307±0.0012	**0.1800**±0.0003	0.1433±0.0022	0.1122±0.0011
Direct PCC	0.2247±0.0029	0.1629±0.0010	0.1220±0.0024	0.0806±0.0002
SCOR	**0.2313**±0.0030	0.1787±0.0015	**0.1542**±0.0030	**0.1200**±0.0012



(a) HEST-1K **(b)** External datasets

Figure 2: Objective differences from conventional MSE across benchmark and external datasets. ΔPCC denotes the absolute PCC difference from MSE. (a) SCOR differences in pooled and section-level PCC across the eight HEST-1K tasks. (b) MISO and SCOR differences on HER2ST and STNet, with SCOR using the setting selected on ccRCC without further tuning.

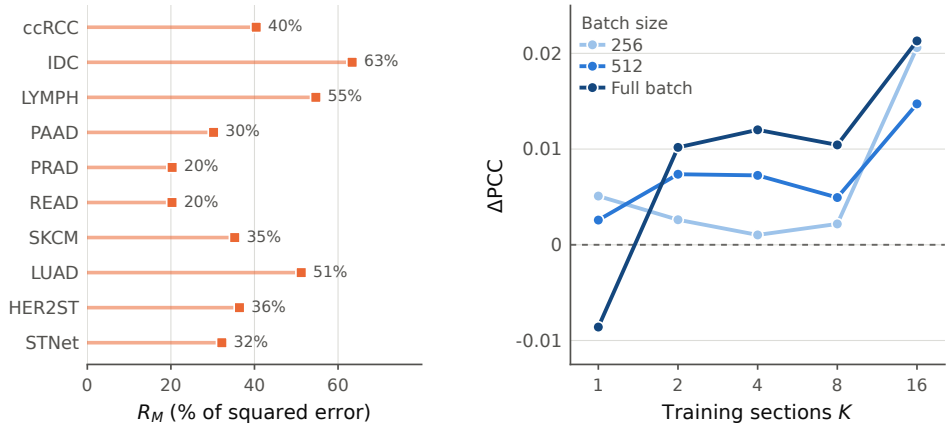
same-section pairing improves pooled PCC by 2.7% and section-level PCC by 6.6%, with the same-section advantage positive in four of five and all five fixed random seeds, respectively. In contrast, the cross-section control remains at or near the conventional MSE baseline under both metrics. The same-section arm uses the theory-matched weight $\lambda_{\text{same}} = 64/65$, whereas the cross-section arm is allowed its own validation-selected weight $\lambda_{\text{cross}} = 4/5$. Thus, the weaker cross-section result cannot be attributed to forcing that control to use a weight selected for same-section pairing. Together with the matched anchors, predictor, initialization, minibatches, optimization budget, and number of partners, this result indicates that pairwise residual differencing alone is insufficient to reproduce the same benefit when the section-defined grouping is broken. Because ccRCC contains one tissue section per patient, this diagnostic does not disentangle tissue-section identity from patient identity and therefore should not be interpreted as establishing tissue section as the unique causal grouping. Rather, it tests whether preserving the section-defined grouping used to construct SCOR is important under the controlled pairwise formulation.

Local spatial proximity modulates the effect but is insufficient to explain it. Relative to the farthest same-section distance tertile, the nearest tertile improves pooled PCC by only 0.5% but improves section-level PCC by 4.3%, indicating that spatial proximity can influence the magnitude of the benefit. Importantly, partners in the farthest tertile, separated by a median of $4.55\,\text{mm}$, still improve over conventional MSE by 2.8% in pooled PCC and 4.6% in section-level PCC, with the section-level gain positive across all five fixed random seeds. The persistence of the gain at this separation indicates that local spatial proximity alone does not account for the behavior of same-section pairing.

4.3 Error Structure and the Role of Section Replication

The component reweighted by SCOR is substantial under conventional MSE training: across the ten prediction tasks, the held-out section-mean error share $R_M = M/(V + M)$ ranges from 20.3%

to 63.4%, with a median of 35.8% (Fig. 3a). Thus, the section-mean residual component accounts for an appreciable fraction of held-out squared error without implying that it is uninformative. The correlation gains generally do not require sacrificing absolute prediction accuracy. Total held-out MSE decreases under SCOR on seven of the ten tasks; among the seven tasks with higher pooled PCC, six also have lower MSE, with PAAD as the only exception. SKCM and LUAD are the remaining tasks with increased MSE, and both also have lower pooled PCC than conventional MSE.



(a) Section-mean error share **(b)** Controlled section replication

Figure 3: Error structure and the role of replicated training groups. (a) Section-mean error share $R_M = M/(V + M)$ under conventional MSE across the ten prediction tasks. (b) Pooled ΔPCC of SCOR relative to MSE as the number of represented ccRCC training sections K varies while the total number of sampled training spots is held fixed.

The controlled section-replication experiment further shows that SCOR's advantage varies with repeated grouping structure. Holding the number of sampled ccRCC training spots fixed, we vary the represented sections over $K \in \{1, 2, 4, 8, 16\}$. Across batch sizes of 256 and 512 and full-batch training, the largest pooled gain occurs at $K = 16$, with positive improvement in all five section-subsampling repeats (Fig. 3b). At $K = 1$, full-batch training contains no replicated section-level structure, whereas minibatch centering remains distinct from full-batch MSE because it is computed from sampled subsets, consistent with Eq. 4 and Appendix A.1. A descriptive cross-task analysis shows a similar relationship between naturally available training-section replication and SCOR gains (Appendix C.1). Overall, these results indicate that SCOR's advantage is modulated by training-section replication, with the largest gain at the highest tested replication level.

5 Conclusion

We identified an overlooked objective-design choice in H&E-to-ST regression: conventional spot-wise MSE implicitly fixes the relative weighting of within-section residual variation and section-mean residual error. SCOR makes this weighting explicit without changing the histology representation, predictor architecture, or inference procedure. Across benchmark and independent datasets, SCOR improves prediction under both nonlinear and linear predictors, while controlled diagnostics show that preserving the section-defined grouping is important for reproducing the gain and that local spatial proximity alone is insufficient to explain it. Training-section replication further modulates the magnitude of the effect. A limitation is that tissue-section and patient identity coincide in the ccRCC grouping diagnostic, preventing their effects from being disentangled. Overall, our results highlight grouped-error weighting as an important objective-design consideration for histology-to-transcriptomics regression.

AI Use Statement

Generative AI tools were used to assist with language editing, manuscript organization, and code development. The authors determined the research questions and experimental design, verified the

implementation and all reported results, and take full responsibility for the scientific content and conclusions of the paper.

REFERENCES

Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, et al. Spatial deconvolution of her2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1):6012, 2021.

Keunho Byeon, Sunhong Park, Jeewoo Lim, and Jin Tae Kwak. Coast: Context-aware differential learning for gene expression prediction in spatial transcriptomics. *arXiv preprint arXiv:2607.09166*, 2026.

Eduard Chelebian, Christophe Avenel, and Carolina Wählby. Combining spatial transcriptomics with tissue morphology. *Nature Communications*, 16(1):4452, 2025.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11591–11600. IEEE, 2024.

Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. Merge: Multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15611–15620. IEEE, 2025.

Gunsagar S Gulati, Jeremy Philip D'Silva, Yunhe Liu, Linghua Wang, and Aaron M Newman. Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nature reviews Molecular cell biology*, 26(1):11–31, 2025.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834, 2020.

Tinglin Huang, Tianyu Liu, Mehrtash Babadi, Wengong Jin, and Rex Ying. Scalable generation of spatial transcriptomics from histology images via whole-slide flow matching. *arXiv preprint arXiv:2506.05361*, 2025.

Guillaume Jaume, Paul Doucet, Andrew H Song, Ming Y Lu, Cristina Almagro-Pérez, Sophia J Wagner, Anurag J Vaidya, Richard J Chen, Drew F Williamson, Ahrong Kim, et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833, 2024.

Taejin Jeong, Joohyeok Kim, Jinyeong Kim, Chanyoung Kim, and Seong Jae Hwang. Feast: Fully connected expressive attention for spatial transcriptomics. *arXiv preprint arXiv:2603.25247*, 2026.

Yuran Jia, Junliang Liu, Li Chen, Tianyi Zhao, and Yadong Wang. Thitogene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1):bbad464, 2024.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Kazuya Nishimura, Haruka Hirose, Ryoma Bise, Kaito Shiku, and Yasuhiro Kojima. Learning relative gene expression trends from pathology images in spatial transcriptomics. *Advances in Neural Information Processing Systems*, 38, 2026.

Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv*, pp. 2021–11, 2021.

Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature communications*, 11(1):3877, 2020.

Benoît Schmauch, Loïc Herpin, Antoine Olivier, Thomas Duboudin, Rémy Dubois, Lucie Gillet, Alexandre Filiot, Jean-Baptiste Schiratti, Valentina Di Proietto, Delphine Le Corre, et al. A deep learning-based multiscale integration of spatial omics with tumor morphology. *Nature Communications*, 16(1):11674, 2025.

Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H Bernhard, Ran A Godrich, Juan A Retamero, et al. End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*, pp. 1–11, 2026.

Chuhan Wang, Adam S Chan, Xiaohang Fu, Shila Ghazanfar, Jinman Kim, Ellis Patrick, and Jean YH Yang. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nature Communications*, 16(1):1544, 2025.

Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36:70626–70637, 2023.

Kexin Xu, Hashem Koohy, and Agne Antanaviciute. Predicting spatial transcriptomics from histology images: Progress, challenges, and directions. *EXO-Beyond the Cell*, 1(3):202619–202619, 2026.

Yan Yang, Md. Zakir Hossain, Eric A. Stone, and Shafin Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pp. 5028–5037. IEEE, 2023. doi: 10.1109/WACV56688.2023.00501. URL https://doi.org/10.1109/WACV56688.2023.00501.

Yan Yang, Md Zakir Hossain, Eric Stone, and Shafin Rahman. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recognition*, 145:109966, 2024. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2023.109966. URL https://www.sciencedirect.com/science/article/pii/S0031320323006647.

Takahisa Yoshikawa, Roman D Bülow, Peter Boor, and Motoko Yanagita. Next-generation kidney tissue analysis—spatial omics and digital pathology. *Nature Reviews Nephrology*, pp. 1–19, 2026.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5):bbac297, 2022.

APPENDIX

A THEORETICAL DETAILS

A.1 MINIBATCH APPROXIMATION OF SCOR

For a tissue section s, let the full-section residual mean and error components be $\bar{\mathbf{e}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{e}_{si}$, $V_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\|\mathbf{e}_{si} - \bar{\mathbf{e}}_s\|^2$, and $M_s = \|\bar{\mathbf{e}}_s\|^2$. Suppose m_s spots are sampled uniformly without replacement from the n_s spots of section s, and let $\bar{\mathbf{e}}_{s,\mathcal{B}}$ denote their sample mean. The expected squared deviation around the sampled mean is

$$\mathbb{E}\left[\frac{1}{m_s}\sum_{i\in\mathcal{B}_s}\|\mathbf{e}_{si} - \bar{\mathbf{e}}_{s,\mathcal{B}}\|^2\right] = \frac{n_s(m_s-1)}{m_s(n_s-1)}V_s,$$

while the expected squared sampled mean is

$$\mathbb{E}\left[\|\bar{\mathbf{e}}_{s,\mathcal{B}}\|^2\right] = M_s + \frac{n_s - m_s}{m_s(n_s-1)}V_s.$$

Combining these terms gives the expected per-spot contribution of section s under minibatch SCOR:

$$\mathbb{E}\left[\frac{1}{m_s}\left(\sum_{i\in\mathcal{B}_s}\|\mathbf{e}_{si} - \bar{\mathbf{e}}_{s,\mathcal{B}}\|^2 + c\,m_s\|\bar{\mathbf{e}}_{s,\mathcal{B}}\|^2\right)\right] = \gamma_s V_s + cM_s, \tag{10}$$

where

$$\gamma_s = 1 - \frac{(1-c)(n_s - m_s)}{m_s(n_s-1)}. \tag{11}$$

Thus, for a fixed minibatch section count m_s, minibatch estimation leaves the coefficient on M_s exactly equal to c, while the effective coefficient on V_s is slightly attenuated from one. Since $(n_s - m_s)/(n_s - 1) \le 1$,

$$\gamma_s \ge 1 - \frac{1-c}{m_s}.$$

At $m_s = 1$, $\gamma_s = c$; when $m_s = n_s$, $\gamma_s = 1$ and the exact full-section objective is recovered. More generally, γ_s approaches one as more observations from the same section are represented in the minibatch. This analysis makes explicit why the stochastic implementation becomes a closer approximation to full-section SCOR as within-section minibatch coverage increases.

A.2 PROFILED SECTION-OFFSET DERIVATION

Starting from Eq. 5, write each residual as $\mathbf{e}_{si} = (\mathbf{e}_{si} - \bar{\mathbf{e}}_s) + \bar{\mathbf{e}}_s$. Since $\sum_i(\mathbf{e}_{si} - \bar{\mathbf{e}}_s) = \mathbf{0}$, the cross term vanishes and

$$\sum_i\|\mathbf{e}_{si} + \mathbf{b}_s\|^2 = \sum_i\|\mathbf{e}_{si} - \bar{\mathbf{e}}_s\|^2 + n_s\|\bar{\mathbf{e}}_s + \mathbf{b}_s\|^2. \tag{12}$$

Substituting Eq. 12 into Eq. 5 gives

$$\mathcal{J} = V + \frac{1}{N}\sum_s n_s\left[\|\bar{\mathbf{e}}_s + \mathbf{b}_s\|^2 + \alpha\|\mathbf{b}_s\|^2\right]. \tag{13}$$

The optimization over $\mathbf{b}_s$ is separable across sections. Differentiating the bracketed term gives

$$2(\bar{\mathbf{e}}_s + \mathbf{b}_s) + 2\alpha\mathbf{b}_s = \mathbf{0},$$

which yields Eq. 6,

$$\mathbf{b}_s^* = -\frac{1}{1+\alpha}\bar{\mathbf{e}}_s.$$

Substitution gives

$$\bar{\mathbf{e}}_s + \mathbf{b}_s^* = \frac{\alpha}{1+\alpha}\bar{\mathbf{e}}_s$$

and

$$\|\bar{\mathbf{e}}_s + \mathbf{b}_s^*\|^2 + \alpha\|\mathbf{b}_s^*\|^2 = \frac{\alpha}{1+\alpha}\|\bar{\mathbf{e}}_s\|^2.$$

Therefore,

$$\min_{\{\mathbf{b}_s\}} \mathcal{J} = V + \frac{\alpha}{1+\alpha} M,$$

which gives Eq. 8. For $0 \le c < 1$, $\alpha = c/(1-c)$, and Eq. 6 becomes

$$\mathbf{b}_s^* = -(1-c)\bar{\mathbf{e}}_s. \tag{14}$$

Thus, $1-c$ corresponds to the fraction of the section-mean residual absorbed by the profiled offset. At $c = 0$, the section-mean residual is fully absorbed; as $c \to 1$, the offset is increasingly suppressed and the objective approaches conventional MSE. The offsets are analytically profiled out and are introduced only to interpret the objective; they are not additional predictor parameters and are not used at inference.

A.3 Pairwise Representation of SCOR

We derive the pairwise representation used in the section-defined grouping diagnostic while accounting explicitly for the exclusion of self-pairs. Consider the subset $\mathcal{B}_s$ of a minibatch containing $m_s = |\mathcal{B}_s| \ge 2$ spots from section s, with residuals $\{\mathbf{e}_{si}\}_{i\in\mathcal{B}_s}$ and minibatch section mean

$$\bar{\mathbf{e}}_{s,\mathcal{B}} = \frac{1}{m_s}\sum_{i\in\mathcal{B}_s} \mathbf{e}_{si}. \tag{15}$$

The corresponding within-section component is

$$V_{s,\mathcal{B}} = \frac{1}{m_s}\sum_{i\in\mathcal{B}_s} \|\mathbf{e}_{si} - \bar{\mathbf{e}}_{s,\mathcal{B}}\|^2. \tag{16}$$

Same-section pairwise representation without self-pairs. Let I be sampled uniformly from $\mathcal{B}_s$, and conditional on $I = i$, let J be sampled uniformly from $\mathcal{B}_s \setminus \{i\}$. Thus, $I \ne J$ by construction. The expected pairwise residual-difference term is

$$\mathbb{E}_{I,J:J\neq I}\left[\frac{1}{2}\|\mathbf{e}_{sI} - \mathbf{e}_{sJ}\|^2\right] = \frac{1}{2m_s(m_s-1)}\sum_{i\in\mathcal{B}_s}\sum_{\substack{j\in\mathcal{B}_s\\ j\neq i}} \|\mathbf{e}_{si} - \mathbf{e}_{sj}\|^2. \tag{17}$$

Because the omitted self-pair terms are identically zero,

$$\sum_i\sum_{j\neq i}\|\mathbf{e}_{si} - \mathbf{e}_{sj}\|^2 = \sum_i\sum_j \|\mathbf{e}_{si} - \mathbf{e}_{sj}\|^2.$$

Using

$$\frac{1}{2m_s^2}\sum_i\sum_j \|\mathbf{e}_{si} - \mathbf{e}_{sj}\|^2 = V_{s,\mathcal{B}}, \tag{18}$$

we obtain

$$\mathbb{E}_{I,J:J\neq I}\left[\frac{1}{2}\|\mathbf{e}_{sI} - \mathbf{e}_{sJ}\|^2\right] = \frac{m_s}{m_s-1}V_{s,\mathcal{B}}. \tag{19}$$

Thus, excluding self-pairs introduces the finite-sample factor $m_s/(m_s-1)$. Multiplying by $(m_s-1)/m_s$ removes this factor:

$$\mathbb{E}_{I,J:J\neq I}\left[\frac{m_s-1}{m_s}\frac{1}{2}\|\mathbf{e}_{sI} - \mathbf{e}_{sJ}\|^2\right] = V_{s,\mathcal{B}}. \tag{20}$$

The corrected same-section residual-difference term is therefore an unbiased estimator of the minibatch within-section component.

Define

$$V_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_s m_s V_{s,\mathcal{B}}, \qquad M_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_s m_s\|\bar{\mathbf{e}}_{s,\mathcal{B}}\|^2.$$

The conventional minibatch MSE satisfies

$$\mathcal{L}_{\mathrm{MSE}}^{\mathcal{B}} = V_{\mathcal{B}} + M_{\mathcal{B}}. \tag{21}$$

Let $\widetilde{\mathcal{P}}_{\mathrm{same}}^{\mathcal{B}}$ denote the corrected pairwise estimator formed using the same spot weighting as the minibatch objective. By Eq. 20,

$$\mathbb{E}\left[\widetilde{\mathcal{P}}_{\mathrm{same}}^{\mathcal{B}}\right] = V_{\mathcal{B}}. \tag{22}$$

Consequently,

$$\mathbb{E}\left[\mathcal{L}_{\mathrm{pair}}^{\mathcal{B}}(\lambda)\right] = (1-\lambda)(V_{\mathcal{B}} + M_{\mathcal{B}}) + \lambda V_{\mathcal{B}} \tag{23}$$

$$= V_{\mathcal{B}} + (1-\lambda)M_{\mathcal{B}}. \tag{24}$$

Setting $\lambda = 1 - c$ therefore gives

$$\mathbb{E}\left[\mathcal{L}_{\mathrm{pair}}^{\mathcal{B}}(1-c)\right] = V_{\mathcal{B}} + cM_{\mathcal{B}},$$

which recovers minibatch SCOR in expectation. The same-section diagnostic in Section 3.2 implements this sampling scheme by drawing partners from the other spots of the same section represented in the current minibatch.

Why cross-section pairing is not equivalent to SCOR. Consider an anchor from section s and a partner from a different section $t \neq s$. Their residual difference can be decomposed as

$$\mathbf{e}_{si} - \mathbf{e}_{tj} = (\mathbf{e}_{si} - \bar{\mathbf{e}}_{s,\mathcal{B}}) - (\mathbf{e}_{tj} - \bar{\mathbf{e}}_{t,\mathcal{B}}) + (\bar{\mathbf{e}}_{s,\mathcal{B}} - \bar{\mathbf{e}}_{t,\mathcal{B}}). \tag{25}$$

Averaging independently over the anchor and partner sections eliminates the centered cross terms and gives

$$\mathbb{E}\left[\frac{1}{2}\|\mathbf{e}_{sI} - \mathbf{e}_{tJ}\|^2\right] = \frac{1}{2}V_{s,\mathcal{B}} + \frac{1}{2}V_{t,\mathcal{B}} + \frac{1}{2}\|\bar{\mathbf{e}}_{s,\mathcal{B}} - \bar{\mathbf{e}}_{t,\mathcal{B}}\|^2. \tag{26}$$

The cross-section pairwise term therefore contains an additional between-section mean-residual component and does not estimate the within-section component targeted by SCOR. The cross-section arm is consequently used only as a matched control that preserves the pairwise residual-difference mechanism while breaking the section-defined grouping.

B Experimental and Reproducibility Details

B.1 Implementation Details

All preprocessing, training, validation, and evaluation procedures are held fixed across objectives unless otherwise stated. HEST-1K data and benchmark splits follow the released benchmark. For HER2ST and STNet, patient-grouped folds are held fixed across experiments. Histology representations are extracted using the official UNI2-h model, and competing objectives are implemented according to their published formulations or official implementations where available. The primary training settings are summarized in Table 3.

B.2 Dataset and Benchmark Characteristics

The evaluation spans ten prediction tasks across three source datasets, comprising 170 tissue sections and 260,895 spatial observations. HEST-1K contributes eight tasks spanning kidney, breast, lymph node, pancreas, prostate, bowel, skin, and lung tissue, with four Visium and four Xenium tasks. HER2ST and STNet provide independent breast-cancer datasets measured with an earlier spatial transcriptomics platform. Table 4 summarizes the complete evaluated data.

The benchmark covers substantial variation in measurement technology and hierarchical organization. Individual tasks range from two to 68 tissue sections and from approximately 3,000 to 74,000 spatial observations, while the number of sections contributed by each patient also varies considerably. This variation provides evaluation settings with substantially different amounts of replicated section-defined structure, which is relevant to the controlled analyses of SCOR.

Table 3: Implementation and training hyperparameters used in the primary benchmark experiments. G denotes the number of genes in the prediction target. Diagnostic and sensitivity analyses use the seed counts specified in their respective experimental descriptions.

Setting	Value
Histology encoder	UNI2-h, frozen
Input representation	1536 dimensions
Feature standardization	StandardScaler, fit on training data only
PCA dimension	256
MLP input dimension	256
MLP hidden layers	$512 \to 512$
Normalization	LayerNorm
Activation	GELU
Dropout	0.3
Output dimension	G
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Batch size	4096 spots
Maximum epochs	150
Early-stopping patience	20 epochs
Checkpoint criterion	Maximum validation pooled PCC
Primary random seeds	2021, 7, 13
SCOR shrinkage parameter	$\alpha = 1/64$
Equivalent section-mean weight	$c = 1/65$
Minibatch construction	Uniform random permutation of training spots
Section means	Computed within each minibatch
Ridge solver	LSQR
Ridge intercept	None
Ridge regularization coefficient	$100/(256G)$
Ridge SCOR computation	Full training sections

Table 4: Characteristics of the prediction tasks used in the evaluation. Counts refer to the complete evaluated task before fold-specific train, validation, and test partitioning.

Source	Task	Technology	Sections	Patients	Spots	Folds
HEST-1K	ccRCC	Visium	24	24	74,220	6
HEST-1K	IDC	Xenium	4	4	35,536	4
HEST-1K	LYMPH	Visium	4	4	19,964	4
HEST-1K	PAAD	Xenium	3	3	7,571	3
HEST-1K	PRAD	Visium	23	2	62,710	2
HEST-1K	READ	Visium	4	2	8,407	2
HEST-1K	SKCM	Xenium	2	2	3,034	2
HEST-1K	LUAD	Xenium	2	2	5,206	2
HER2ST	Breast	Legacy	36	8	13,620	4
STNet	Breast	Legacy	68	23	30,627	6

B.3 Validation and Controlled Replication Protocols

For HEST-1K neural training, validation is carved from the training fold before preprocessing. Whole tissue sections are held out for validation when at least three training sections are available. When fewer training sections are available, a 15% spot-level carve from the training fold is used for checkpoint selection and early stopping. HER2ST and STNet use their fixed validation sections. Checkpoint selection and early stopping use pooled validation PCC, and all feature standardization and PCA fitting exclude validation and test observations as appropriate.

The primary MLP experiments use minibatches of 4096 spots drawn from a uniform random permutation of the training spots within each epoch. For SCOR, section means are computed from

the observations belonging to each section within the current minibatch. Ridge regression instead evaluates the objective over the complete training sections.

For the controlled ccRCC section-replication experiment, the number of represented training sections is varied over $K \in \{1, 2, 4, 8, 16\}$ while the total number of sampled training spots is held fixed within each fold–seed combination. The fixed total ranges from 1184 to 1424 spots across combinations. Spots are distributed equally across the selected sections, and MSE and SCOR receive exactly the same sampled observations at every K. Each condition is evaluated with minibatch sizes of 256 and 512 and with full-batch optimization. This design changes the amount of replicated section-defined structure while holding the number of training observations fixed.

C Additional SCOR Analysis

C.1 Cross-task Section Replication

The controlled ccRCC section-replication experiment isolates the effect of training-section replication within a single prediction task by varying the number of represented training sections while holding the tissue, target, representation, predictor, number of training observations, and evaluation setting fixed. We complement this controlled analysis with a descriptive cross-task comparison asking whether a similar pattern is visible across prediction tasks that naturally differ in the amount of section replication available during training. Because these tasks also differ in tissue type, measurement technology, target construction, and other dataset characteristics, this analysis is not intended to isolate a causal effect of section replication; rather, it tests whether the cross-task pattern is consistent with the controlled ccRCC finding.

For each prediction task, we summarize training-section replication by the minimum number of training sections available across its evaluation folds and compare this quantity with the improvement of SCOR over conventional MSE.

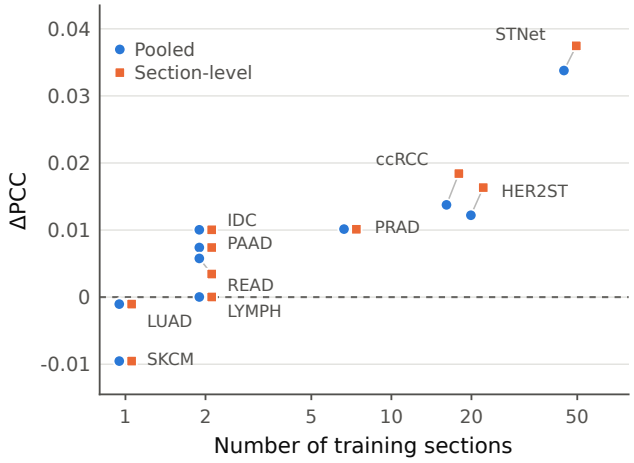


Figure 4: SCOR improvement over conventional MSE as a function of training-section replication across the ten prediction tasks. The horizontal axis reports the minimum number of training sections available across folds for each task.

Across the ten prediction tasks, greater training-section replication is generally associated with larger SCOR gains under both pooled and section-level PCC (Fig. 4). STNet, HER2ST, and ccRCC combine relatively high training-section replication with some of the largest gains, whereas LUAD and SKCM have only one training section per fold and are the two HEST-1K tasks on which SCOR is lower than MSE. The relationship is not strictly monotonic, and the tasks differ along multiple dimensions beyond section replication. The cross-task comparison should therefore be interpreted descriptively, but its overall pattern is consistent with the controlled ccRCC experiment in which the number of represented training sections is varied while the number of training observations is held fixed.

C.2 Objective-Setting Selection and c-Sensitivity

C.2.1 Development-Set Selection

SCOR is parameterized through the profiled-offset interpretation in Eq. 8. Objective-setting selection is performed using ccRCC development data only. We sweep

$$\alpha \in \left\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}\right\}$$

and examine both pooled and section-level validation PCC. No fixed numerical threshold is imposed for defining equivalent or comparable validation performance. Instead, the operating point is selected from the joint validation curves: section-level PCC reaches a plateau around $\alpha = 1/64$, while pooled PCC is still improving. We therefore select

$$\alpha = \frac{1}{64},$$

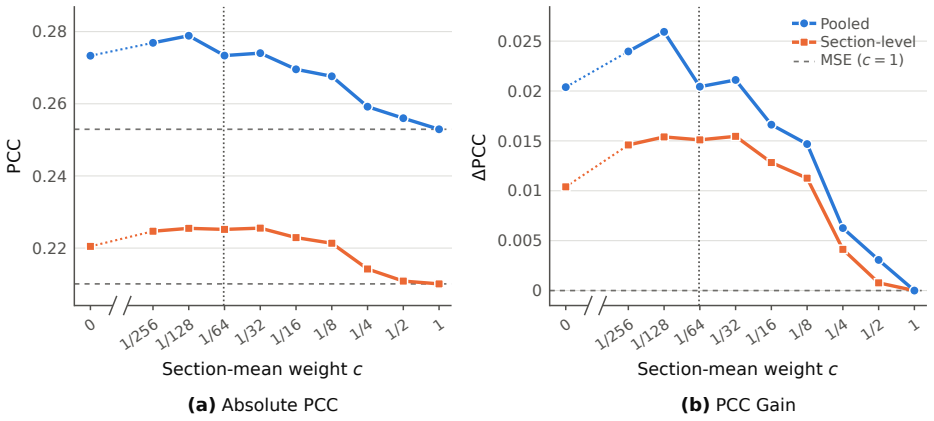
which corresponds to

$$c = \frac{\alpha}{1+\alpha} = \frac{1/64}{1+1/64} = \frac{1}{65}.$$

This setting is fixed before evaluation on the remaining HEST-1K tasks and before all HER2ST and STNet experiments. No test-set results are used to revise the selected value. The post-hoc c-sensitivity analysis in Section C.2.2 is conducted only after this manuscript setting has been fixed and is not used for objective-setting selection.

C.2.2 Post-hoc Sensitivity to the Section-Mean Weight

After fixing the manuscript setting $c = 1/65$ from the development procedure above, we perform a separate post-hoc sensitivity analysis on ccRCC to assess whether SCOR depends strongly on the precise value of the section-mean weight c. Ten values from $c = 0$ to $c = 1$ are evaluated using eight fixed random seeds paired across objective settings and the same six frozen folds. SCOR is broadly stable across the low-c region (Fig. 5). From $c = 1/256$ to $c = 1/32$, gains over MSE remain similar under both pooled and section-level PCC, and $c = 1/64$ improves both metrics in all eight seeds. Above this region, the gain decreases progressively as c approaches one, where $\mathcal{L}_{\mathrm{SC}} = V + cM$ reduces exactly to conventional MSE.



(a) Absolute PCC **(b)** PCC Gain

Figure 5: Post-hoc sensitivity of SCOR to the section-mean weight c on ccRCC. (a) Pooled and section-level PCC across the sweep, with $c = 1$ corresponding to conventional MSE. (b) Corresponding absolute ΔPCC from MSE. The selected manuscript setting $c = 1/65$ is not a point on the post-hoc sweep grid; the dotted vertical line marks $c = 1/64$, the nearest evaluated value.

The $c = 0$ boundary clarifies why SCOR downweights rather than removes the section-mean component. At $c = 0$, the loss becomes $\mathcal{L}_{\mathrm{SC}} = V$, so a residual shared uniformly across all spots in a

section is unpenalized. Such an offset can leave within-section Pearson correlation unchanged while substantially shifting the section-average prediction. As shown in Table 5, removing M entirely retains a PCC gain but reduces the section-level improvement and causes held-out section-mean error to increase sharply. A small positive weight on M therefore constrains section-wide offsets while retaining the correlation benefit associated with strong downweighting.

Table 5: Selected values from the post-hoc sensitivity analysis of SCOR to the section-mean weight c on ccRCC. ΔPCC denotes the absolute PCC difference from conventional MSE using paired seeds. M denotes the held-out section-mean squared error.

c	ΔPCC pooled	ΔPCC section	Held-out M
0	+0.0204	+0.0104	0.678
1/256	+0.0240	+0.0146	0.284
1/128	+0.0259	+0.0154	0.222
1/64	+0.0204	+0.0151	0.187
1/32	+0.0211	+0.0155	0.175
1	0	0	0.181

Differences among neighboring low-c settings are modest relative to the observed seed-to-seed variability, supporting a broad region of stable performance rather than a sharply defined optimum. The numerically highest pooled gain at $c = 1/128$ should therefore not be interpreted as evidence for a uniquely optimal setting. The manuscript setting $c = 1/65$ was selected using validation data before this sensitivity analysis and lies within the stable low-c region. Because the sweep evaluates held-out test folds after the manuscript setting had already been fixed, it is reported only as a post-hoc sensitivity analysis and is not used to revise the selected value of c.

D Additional Predictor Results

D.1 Ridge Regression

We evaluate SCOR with ridge regression to test whether its effect extends beyond the nonlinear MLP predictor used in the primary benchmark. Both MSE and SCOR use the same frozen folds, UNI2-h representations, PCA preprocessing, and fixed ridge specification, with no additional hyperparameter tuning. Because ridge regression is deterministic, each task is fit once under the matched protocol.

Table 6: Ridge regression results under conventional MSE and SCOR across the ten prediction tasks. ΔPCC denotes the absolute difference $\mathrm{PCC}_{\mathrm{SCOR}} - \mathrm{PCC}_{\mathrm{MSE}}$. Values are displayed to four decimal places; improvement counts reported in the text are based on the underlying unrounded results.

Task	Pooled PCC			Section-level PCC		
	MSE	SCOR	ΔPCC	MSE	SCOR	ΔPCC
ccRCC	0.2420	**0.2898**	+0.0478	0.1979	**0.2174**	+0.0195
IDC	0.5805	**0.5820**	+0.0015	0.5805	**0.5820**	+0.0015
LYMPH	**0.2487**	0.2425	-0.0062	**0.2487**	0.2425	-0.0062
PAAD	**0.4887**	0.4876	-0.0011	**0.4887**	0.4876	-0.0011
PRAD	**0.3746**	0.3588	-0.0158	0.3182	**0.3198**	+0.0016
READ	0.2143	**0.2163**	+0.0020	0.2365	**0.2393**	+0.0028
SKCM	0.6510	**0.6511**	+0.0001	0.6510	**0.6511**	+0.0001
LUAD	0.5470	0.5470	+0.0000	0.5470	0.5470	+0.0000
HER2ST	0.2071	**0.2072**	+0.0001	0.1473	**0.1598**	+0.0125
STNet	0.1279	**0.1441**	+0.0162	0.0799	**0.1087**	+0.0288

SCOR improves section-level PCC on eight of the ten prediction tasks and pooled PCC on seven when comparisons are made using the underlying unrounded results. The largest gains occur on ccRCC and STNet, while HER2ST also shows a marked section-level improvement. These results

show that the effect of section-centered error reweighting is not confined to the nonlinear MLP predictor.

E STRank Regime Sensitivity

STRank (Nishimura et al., 2026) was developed under a prediction regime that differs from the primary HEST-1K regression protocol used in our main experiments, including its target construction, gene selection, data splits, and histology representation. We therefore conduct a separate analysis under a STRank-aligned regime. We retain the STRank target construction, highly variable gene panels, official data splits, and training setup, while replacing its original CONCH-v1 (Lu et al., 2024) histology backbone with UNI2-h so that all objectives are compared using the same representation as in our main experiments. Specifically, the normalized expression target is

$$y_{sig} = \log\left(1 + 10^4 c_{sig} / \sum_g c_{sig}\right).$$

Because STRank explicitly supervises relative expression ordering, we report per-gene Spearman correlation in addition to Pearson correlation. Before the cross-task comparison, we performed an implementation sanity check using the original CONCH-v1 backbone on ccRCC. Under this original-backbone configuration, our MSE result matches the value reported by STRank, while PairSTRank and ListSTRank achieve slightly higher performance than their corresponding reported results. This ccRCC sanity check supports the fidelity of our STRank implementation and makes implementation failure an unlikely explanation for its weaker performance under the common HEST-1K protocol.

Under the STRank-aligned regime with the shared UNI2-h representation, ranking supervision becomes substantially more competitive than under the common HEST-1K protocol, while SCOR remains consistently effective across tasks (Table 7). Under Spearman evaluation, SCOR improves over conventional MSE on seven of eight tasks and achieves the highest mean performance of 0.3479. PairSTRank improves over MSE on four tasks and achieves the highest task-level score on PRAD, PAAD, and ccRCC, consistent with the relative-ordering supervision for which it was designed. Under Pearson evaluation, SCOR again improves over MSE on seven of eight tasks and achieves the highest mean performance of 0.3706, with the highest score on seven tasks. PairSTRank and ListSTRank each improve over MSE on three tasks. PRAD is the only task on which SCOR is lower than MSE under both Spearman and Pearson evaluation.

These results provide two complementary observations. First, the stronger performance of the ranking objectives under the STRank-aligned regime confirms that their behavior depends substantially on the target and evaluation setting, rather than reflecting an implementation failure in the primary benchmark. Second, SCOR does not rely on the original HEST-1K target construction or Pearson-based evaluation to improve over MSE. With the same fixed SCOR objective setting and the shared UNI2-h representation, the improvement persists on seven of eight tasks under both Spearman and Pearson correlation after adopting the STRank target, gene-selection, split, and training protocol. The comparison therefore highlights a distinction between the objectives: STRank changes the supervision toward relative expression ordering, whereas SCOR retains regression supervision while controlling how grouped residual error contributes to learning.

Table 7: Performance across the eight HEST-1K tasks under the STRank-aligned target, gene-selection, split, and training protocol with a shared UNI2-h histology backbone. Mean is the unweighted average across tasks. Best results within each task and metric are shown in **bold**.

		IDC	PRAD	PAAD	READ	ccRCC	LYMPH	SKCM	LUAD	Mean
Spearman	MSE	0.4896	0.4508	0.2147	0.2141	0.1437	0.3638	0.3112	0.3796	0.3209
	PairSTRank	0.4644	**0.4729**	**0.3021**	0.1844	**0.1690**	0.3479	0.2972	0.4288	0.3334
	ListSTRank	0.4289	0.4589	0.2880	0.1683	0.1651	0.3205	**0.3507**	0.4521	0.3291
	SCOR	**0.5044**	0.4368	0.2893	**0.2212**	0.1593	**0.3722**	0.3478	**0.4522**	**0.3479**
Pearson	MSE	0.4862	0.4465	0.2084	0.2577	0.1713	0.3136	0.3805	0.4477	0.3390
	PairSTRank	0.4191	**0.4602**	0.2339	0.2361	0.1969	0.2130	0.2926	0.3494	0.3001
	ListSTRank	0.3814	0.4470	0.2129	0.2149	0.1904	0.1734	0.3523	0.3559	0.2910
	SCOR	**0.4985**	0.4356	**0.2912**	**0.2688**	**0.2009**	**0.3563**	**0.4034**	**0.5104**	**0.3706**