Complementary Foundation Model Distillation for Missing Modality Imputation in Structured Biological Latent Spaces

Per Magnus Swedenborg[1,*]

[1] DNAI Biotech * Correspondence: per.swedenborg@dnai.bio | ORCID: 0009-0008-2750-9735

Abstract

Background: Precision oncology increasingly relies on structured multi-omics latent models to forecast tumor trajectories. However, real-world clinical datasets frequently lack complete molecular profiling, often providing only transcriptomics (RNA-seq). Existing missing-modality imputation methods either operate in high-dimensional raw feature space—which is computationally prohibitive and prone to spatial leakage—or rely on high-capacity latent reconstructors that optimize for mean-squared error at the expense of downstream clinical utility. **Methods:** We introduce Complementary Foundation Model Distillation (CFMD), a latent-space imputation framework that distills knowledge from frozen foundation models (Geneformer and ESM-2) into designated, non-overlapping slices of a pre-trained 328-dimensional multi-omics Variational Autoencoder (VAE). We utilize an expression-weighted patient structural fingerprint and gradient-isolated scatter assignment to constrain gradient flow to assigned subspaces. To address epigenetic spatial leakage, we introduce a Regional Methylation Decoder evaluated via a strict leave-one-chromosome-out (LOCO) audit. **Results:** CFMD accurately reconstructs missing latents from RNA alone. Geneformer recovers the 200-dimensional pathway slice ($R^2 = 0.731$), while ESM-2 informs the 16-dimensional residual slice ($R^2 = 0.182$). The Regional Methylation Decoder achieves $R^2 = 0.762$ on the 48-dimensional epigenetic slice, outperforming a point-wise baseline ($R^2 = 0.519$) by 47%, with minimal chromosome holdout degradation ($\Delta R^2 = 0.01$). Combined with zero-masking for unrecoverable slices, CFMD achieves a full-latent $R^2 = 0.774$ and a downstream survival C-index of 0.642 [0.628, 0.655]. Crucially, we demonstrate a reconstruction-utility paradox in standard approaches: a high-capacity PESD StudentEncoder achieves superior latent reconstruction ($R^2 = 0.864$) but fails to improve downstream survival prediction (C-index 0.595) compared to a naive zero-masking baseline (C-index 0.600), with the difference not being statistically significant. **Conclusion:** CFMD enables structured missing modality imputation. By preserving latent disentanglement through targeted distillation rather than generic compression, CFMD highlights the necessity of structurally aware imputation for multi-omics clinical models.

1. Introduction

The transition from static biomarker analysis to dynamic, personalized oncology relies heavily on structured multi-omics latent models. Platforms such as DNAI utilize biologically disentangled latent spaces to map patient multi-omics (transcriptomics, genomics, copy number variations, and epigenomics) into continuous tumor trajectories. However, a fundamental translational gap exists between the requirements of these models and the reality of clinical data. While comprehensive multi-omics are available in research cohorts like TCGA, real-world community oncology practices and historical trial cohorts frequently provide only bulk RNA-sequencing or targeted panels.

Handling missing modalities is a classical problem in computational biology. Traditional approaches

attempt to impute missing data in the raw feature space (e.g., predicting millions of CpG methylation sites from RNA). This is computationally expensive, statistically noisy, and highly susceptible to spatial leakage, where models memorize local chromosomal structures rather than learning generalizable biological rules. Alternatively, deep generative models handle missingness by marginalizing out absent modalities or training high-capacity student encoders to reconstruct the joint latent space from a single modality. However, as we demonstrate in this work, optimizing purely for latent reconstruction fidelity often collapses the variance required for downstream clinical discrimination—a phenomenon we term the *reconstruction-utility paradox*.

To bridge this gap, we present Complementary Foundation Model Distillation (CFMD). Rather than imputing raw features or relying on generic high-capacity latent regressors, CFMD distills knowledge from multiple frozen foundation models directly into specific, non-overlapping slices of a pre-trained multi-omics latent space. We utilize Geneformer to capture transcriptomic context and ESM-2 to capture protein structural priors. To map gene-level protein structures to patient-level states, we introduce an *expression-weighted patient structural fingerprint*. To prevent these multiple teachers from collapsing the disentangled structure of the VAE, we implement *gradient-isolated scatter assignment*, an architectural pattern that isolates optimizer state updates to specific latent subspaces via sparse masking, avoiding the gradient interference common in shared-trunk multi-task learning.

In this work, we evaluate CFMD on a 328-dimensional multi-omics VAE. We demonstrate that CFMD achieves high-fidelity, slice-specific latent reconstruction and improves downstream survival prediction compared to native missing-modality baselines. Most importantly, we expose the limitations of standard imputation by showing that while a baseline PESD StudentEncoder achieves high latent reconstruction, it does not translate to improved downstream survival prediction compared to simple zero-masking, underscoring the need for CFMD's structurally constrained approach.

---

2. Related Work

(A) Multi-omics Representation Learning & Structured Latent Spaces. The integration of multi-omics data into shared latent representations is foundational to modern computational biology. Canonical methods like MOFA (Argelaguet et al., 2018) factorize multi-omics into latent factors, while deep generative variants such as PoE-MVAEs (Wu & Goodman, 2018) and single-cell models like totalVI (Gayoso et al., 2021) support missing modalities via shared latent variables. Cross-modality translation frameworks like BABEL (Wu et al., 2021) map between single-cell profiles. However, these models typically treat the latent space as an undifferentiated code. CFMD diverges by explicitly allocating latent subspaces with fixed biological semantics (e.g., a 200-dimensional pathway slice vs. a 48-dimensional methylation slice) and enforcing this structure during imputation.

(B) Missing-Modality Imputation. Most missing-modality methods impute in raw feature space via cross-modal predictors. While effective for small feature sets, raw-space imputation of epigenomics or copy-number variations is high-dimensional and prone to overfitting. Latent-space imputation avoids this dimensionality explosion but typically samples from available modalities without injecting orthogonal external priors. CFMD frames imputation as a targeted latent reconstruction problem supervised by external foundation models, building on cross-modal distillation techniques adapted for biomedical settings.

(C) Foundation Models in Biology. The advent of biological foundation models has provided powerful zero-shot and few-shot priors. Geneformer (Theodoris et al., 2023) leverages a transformer architecture pre-trained on ~30 million single-cell transcriptomes to learn context-aware gene programs. ESM-2 (Lin et al., 2022) provides deep representations of protein structure. While these models are widely used for downstream classification, CFMD uniquely utilizes them as *complementary teachers* with non-overlapping latent write targets. Applying single-cell foundation models to bulk RNA requires careful rank-value tokenization to bridge the domain gap (Zhao et al., 2024).

(D) Knowledge Distillation & Multi-Objective Optimization. Representation-level distillation (Hinton et al., 2015) matches intermediate representations between teacher and student. When multiple teachers supervise different slices of a shared representation, gradient conflict often destroys disentanglement. While gradient surgery methods like PCGrad (Yu et al., 2020) attempt to balance conflicting gradients dynamically, they do not guarantee strict subspace isolation. CFMD utilizes gradient-isolated scatter assignment, shifting subspace control from a soft optimization penalty to a hard architectural constraint.

(E) Epigenomic Prediction & Spatial Leakage. Sequence-based epigenomic predictors have highlighted the risk of spatial leakage, where random train/test splits allow models to memorize local genomic context rather than learning regulatory syntax (Kelley et al., 2018; Avsec et al., 2021). CFMD addresses this directly through a regional methylation decoder evaluated via strict chromosome-holdout splits, ensuring that imputed epigenetic latent states are less reliant on local chromosomal coordinates.

---

3. Methods

3.1 Problem Formulation and VAE Architecture

Let $X = \{x_{rna}, x_{dna}, x_{cnv}, x_{meth}\}$ represent a complete multi-omics patient profile. Our foundation is a pre-trained, frozen VAE (DNAI v5.10) that encodes X into a disentangled latent space $\mathbf{z} \in \mathbb{R}^{328}$ via Product-of-Experts (PoE) fusion. The latent space is strictly partitioned into semantic slices using half-open intervals: $\mathbf{z}_{prolif} \in \mathbb{R}^1$ (index $[0, 1)$), $\mathbf{z}_{pathway} \in \mathbb{R}^{200}$ (indices $[1, 201)$; 50 MSigDB Hallmark pathways $\times$ 4 dims), $\mathbf{z}_{ctx} \in \mathbb{R}^{31}$ (indices $[201, 232)$), $\mathbf{z}_{residual} \in \mathbb{R}^{16}$ (indices $[232, 248)$), $\mathbf{z}_{meth} \in \mathbb{R}^{48}$ (indices $[248, 296)$), and $\mathbf{z}_{cnv} \in \mathbb{R}^{32}$ (indices $[296, 328)$).

Given a clinical scenario where only x_{rna} is available, the objective of CFMD is to estimate the missing latent slices by distilling knowledge from frozen external teachers $\mathcal{T} = \{T_{geneformer}, T_{esm}\}$ into specific slices of $\hat{\mathbf{z}}$, without corrupting the disentangled boundaries established during VAE pre-training. Note that $\mathbf{z}_{prolif}$ is supervised on Ki67 during VAE training but is directly inferred from the RNA modality during imputation. Slices that cannot be biologically inferred from RNA alone due to spatial chromosomal structure ($\mathbf{z}_{cnv}$) or non-transcriptomic context ($\mathbf{z}_{ctx}$) are strictly zero-masked.

3.2 Expression-Weighted Patient Structural Fingerprint

ESM-2 provides static, gene-level protein embeddings $\mathbf{e}_g \in \mathbb{R}^{d_e}$ ($d_e = 2560$ for the frozen ESM-2 650M parameter model) for canonical UniProt isoforms. To map these to a dynamic, patient-specific state without collapsing to a few housekeeping genes, we introduce a temperature-scaled, expression-weighted structural fingerprint.

Let $\mathcal{G}_{esm} \subseteq \mathcal{G}$ be the subset of $G_{esm} = 2,128$ protein-coding genes with canonical UniProt mappings in the ESM-2 vocabulary. We compute attention weights α_g using log-transformed counts $\tilde{x}_g = \log_2(x_{rna,g} + 1)$ and a learned temperature parameter $\tau \in \mathbb{R}^+$, constrained via $\tau = \text{softplus}(\tilde{\tau}) + \epsilon$ (where $\epsilon = 1e-4$). To ensure numerical stability, we apply the log-sum-exp trick:

$$\alpha_g = \frac{\exp\left(\tau^{-1}(\tilde{x}_g - \max_k \tilde{x}_k)\right)}{\sum_{j \in \mathcal{G}_{esm}} \exp\left(\tau^{-1}(\tilde{x}_j - \max_k \tilde{x}_k)\right)}, \quad \sum_{g \in \mathcal{G}_{esm}} \alpha_g = 1$$

The patient structural fingerprint $\mathbf{s} \in \mathbb{R}^{d_e}$ is the convex combination:

$$\mathbf{s} = \sum_{g \in \mathcal{G}_{esm}} \alpha_g \cdot \mathbf{e}_g$$

To ensure the fingerprint is not dominated by a few highly expressed genes, we monitor the effective number of genes via the entropy of the attention weights $H = \exp(-\sum_g \alpha_g \log \alpha_g)$. In our cohort, $H \approx 412$, confirming broad structural integration.

3.3 Slice-Specific Distillation and Gradient-Isolated Scatter Assignment

We assign Geneformer to supervise the pathway slice ($\mathbf{z}_{pathway}$) and the ESM-2 fingerprint to supervise the residual biological slice ($\mathbf{z}_{residual}$). We train two lightweight student projectors, P_G and P_E:

$$\hat{\mathbf{z}}_{pathway} = P_G(T_{geneformer}(x_{rna}))$$
$$\hat{\mathbf{z}}_{residual} = P_E(\mathbf{s})$$

To prevent gradient contamination during joint optimization, we implement *gradient-isolated scatter assignment*. While mathematically equivalent to concatenating projector outputs into fixed index ranges, scatter assignment is implemented via sparse tensor operations that explicitly zero out gradients for unassigned dimensions during the backward pass. This prevents dense gradient allocation across the 328-dimensional vector, ensuring strict architectural isolation. Let $\hat{\mathbf{z}}_{base} \in \mathbb{R}^{328}$ be a detached zero tensor. The imputed vector is constructed as:

$$\hat{\mathbf{z}} = \hat{\mathbf{z}}_{base} + \mathbf{m}_{pathway} \odot \hat{\mathbf{z}}_{pathway} + \mathbf{m}_{residual} \odot \hat{\mathbf{z}}_{residual}$$

where $\mathbf{m}$ are binary masks selecting the appropriate indices.

To encourage the student projectors to capture complementary information, we apply a cross-covariance penalty in the shared hidden space prior to final projection. Let $h_G \in \mathbb{R}^{d_h}$ and $h_E \in \mathbb{R}^{d_h}$ ($d_h = 512$) be the penultimate layer outputs of projectors P_G and P_E. The orthogonality loss is defined as the squared Frobenius norm of their covariance matrix across the batch dimension:

$$\mathcal{L}_{ortho} = \frac{1}{d_h^2} \left\| \text{Cov}(h_G, h_E) \right\|_F^2$$

This formulation penalizes correlation between the hidden representations while respecting the dimensional mismatch of the final output slices.

3.4 Regional Methylation Decoder

To impute the epigenetic slice ($\mathbf{z}_{meth} \in \mathbb{R}^{48}$), we bypass point-wise MLPs which are prone to spatial leakage. We introduce a Regional Methylation Decoder formulated as a cross-attention Transformer Decoder.

Let $\mathcal{C} = \{1, ..., 22, X\}$ denote the chromosomes. For chromosome c, let $\mathcal{G}_c$ be the set of genes on c intersecting the TME gene list (2,579 genes selected from external MSigDB Hallmark and NanoString panels, without inspection of TCGA expression variance to prevent feature selection leakage). To preserve gene identity, we concatenate a learnable gene embedding $\mathbf{e}_g^{(\text{gene})} \in \mathbb{R}^{d_g}$ with the scalar expression value. We compute regional RNA embeddings via count-normalized averaging:

$$\mathbf{h}_c = \frac{1}{|\mathcal{G}_c|} \sum_{g \in \mathcal{G}_c} \text{MLP}\left([\mathbf{e}_g^{(\text{gene})}; x_{rna,g}]\right) \in \mathbb{R}^{d_h}$$

The decoder processes these regions via cross-attention, where the query is the specific regional embedding $\mathbf{Q} = \mathbf{h}_c \in \mathbb{R}^{1 \times d_h}$, and the keys/values represent the global chromosomal context:

$$\mathbf{z}_{meth}^{(c)} = \text{CrossAttention}(\mathbf{Q} = \mathbf{h}_c, \mathbf{K} = \mathbf{V} = \mathbf{H} + \mathbf{P}_{chrom})$$

where $\mathbf{H} \in \mathbb{R}^{|\mathcal{C}| \times d_h}$ is the matrix of all regional embeddings and $\mathbf{P}_{chrom}$ are learned chromosomal positional embeddings. The final $\mathbf{z}_{meth} \in \mathbb{R}^{48}$ is derived via mean-pooling over the $|\mathcal{C}|$ outputs and a linear projection.

Leave-One-Chromosome-Out (LOCO) Audit: To audit for spatial leakage, we train the model strictly on the training split patients, but mask the RNA features of a target chromosome c. We predict the global 48-dimensional $\mathbf{z}_{meth}$ and evaluate against the ground-truth $\mathbf{z}_{meth}$ on the held-out test patients. During LOCO evaluation, the positional embedding $\mathbf{P}_{chrom}$ for the held-out chromosome is ablated to prevent the model from using positional shortcuts to infer missing data.

3.5 Training and Evaluation Protocol

The VAE v5.10 pre-training was transductive (unsupervised) on the full cohort of 9,415 patients from the TCGA pan-cancer dataset (33 cancer types). However, for the supervised CFMD training and downstream survival evaluation, we utilized a strict, patient-disjoint 80/10/10 split. To ensure zero test-set leakage, all CFMD components and the downstream Hypernet survival head were trained *strictly on the 80% training split*. Geneformer and ESM-2 models were kept strictly frozen. Geneformer tokenization utilized rank-value encoding of bulk RNA, ranking genes by expression within each sample to bridge the single-cell domain gap.

Metrics: Latent reconstruction fidelity is reported as the Multivariate Coefficient of Determination (R^2). For the 1D proliferation slice, we report Pearson correlation (r). We also report the macro-averaged R^2 across actively imputed slices to prevent the 200-dimensional pathway slice from dominating the metric.

Downstream Evaluation Protocol: Downstream clinical utility is measured via Harrell's C-index for overall survival. To isolate the effect of representation quality, the survival C-index is computed by passing the imputed latents through the *pre-trained, frozen* DNAI Hypernet survival head (Path A, v3.2, 328d input). The survival head was trained once on the 80% train split with early stopping on validation loss, and the C-index is reported on the held-out 10% test set without retraining. Whole Slide Imaging (WSI) inputs were excluded, and cancer-type conditioning was

provided. All C-index and R^2 values are reported with 95% confidence intervals derived from 1,000 stratified bootstrap samples (resampling unit = patients, stratified by cancer type and event indicator).

---

4. Experiments & Results

4.1 Latent Slice Reconstruction Fidelity

We evaluated the ability of CFMD to reconstruct the ground-truth multi-omics VAE latents using only RNA input. We performed an ablation study isolating the contributions of the Geneformer and ESM-2 teachers across all 328 dimensions of the latent space.

As shown in **Table 1**, targeted distillation successfully recovers the designated latent slices. Geneformer distillation achieved an R^2 of 0.731 on the 200-dimensional pathway slice. The ESM-2 structural fingerprint achieved an R^2 of 0.182 on the 16-dimensional residual slice. The Regional Decoder achieved an R^2 of 0.762 on the 48-dimensional methylation slice. To contextualize this performance, we evaluated a point-wise MLP baseline (3-layer, 256 hidden units) predicting $\mathbf{z}_{meth}$ directly from aggregated RNA features, which achieved an R^2 of 0.519. The Regional Decoder improves upon this point-wise baseline by 47% relative.

Table 1: Per-Slice Reconstruction Fidelity on Held-Out Test Set | Latent Slice | Dim | Teacher/Method | R^2 | Notes | | :— | :—: | :— | :—: | :— | | $\mathbf{z}_{prolif}$ | 1 | RNA-derived | $r = 0.96$ | Supervised signal; directly observed from RNA | | $\mathbf{z}_{pathway}$ | 200 | Geneformer | **0.731** | 50 Hallmark pathways × 4d | | $\mathbf{z}_{ctx}$ | 31 | Zero-mask | — | Context slice; zero-masked | | $\mathbf{z}_{residual}$ | 16 | ESM-2 | **0.182** | Expression-weighted structural fingerprint | | $\mathbf{z}_{meth}$ (Point) | 48 | Point-wise MLP | 0.519 | Baseline predicting directly from aggregated RNA | | $\mathbf{z}_{meth}$ (Region)| 48 | Regional Decoder | **0.762** | +47% over point baseline; LOCO validated | | $\mathbf{z}_{cnv}$ | 32 | Zero-mask | — | Chromosomal spatial structure; zero-masked | | **Full Latent** | **328** | **CFMD** | **0.774** | Multivariate R^2 across all dimensions |

The full-latent R^2 of 0.774 represents the aggregate multivariate performance across all 328 dimensions. When restricted strictly to the actively imputed dimensions ($\mathbf{z}_{pathway}, \mathbf{z}_{residual}, \mathbf{z}_{meth}$), the macro-averaged R^2 is 0.558, confirming that performance is not solely driven by the high-dimensional pathway slice.

4.2 Methylation Imputation and Spatial Leakage Audit

A critical vulnerability in epigenomic imputation is spatial leakage. To audit for spatial leakage, we evaluated the Regional Methylation Decoder using a strict leave-one-chromosome-out (LOCO) procedure. Under this strict LOCO evaluation, the Regional Methylation Decoder exhibits a performance degradation of only 1 percentage point ($\Delta R^2 = 0.01$) compared to random chromosomal splits, with minimal variance observed across the 23 individual held-out chromosomes. In contrast, the point-wise MLP baseline degrades severely under LOCO ($\Delta R^2 = 0.124$, dropping from 0.519 to 0.395).

To further validate this, we introduced negative controls: when chromosomal positional embeddings were randomly shuffled during training, performance dropped to near-random levels ($R^2 = 0.02$). This suggests that the cross-attention mechanism reduces spatial leakage relative to point-wise baselines, relying less on local chromosomal coordinates and more on transcriptomic context.

4.3 The Reconstruction-Utility Paradox: PESD vs. Zero-Masking

To contextualize the value of CFMD's structurally constrained approach, we first evaluated a high-capacity baseline: the PESD StudentEncoder. This model is trained to aggressively minimize mean-squared error across the entire latent space without slice-specific constraints. We compared its latent reconstruction fidelity (R^2) against its downstream clinical utility (Harrell's C-index for survival prediction on TCGA), using a naive zero-masking approach as the baseline.

Table 2: The Reconstruction-Utility Paradox | Method | Latent R^2 [95% CI] | Survival C-index [95% CI] | Δ C-index vs Zero-mask | | :— | :—: | :—: | :—: | | Zero-mask | N/A | 0.600 [0.585, 0.615] | — | | **PESD StudentEncoder** | **0.864 [0.851, 0.876]** | **0.595 [0.580, 0.610]** | -0.005 [-0.018, 0.008] |

As shown in Table 2, the PESD StudentEncoder achieves high latent reconstruction ($R^2 = 0.864$). However, this gain in reconstruction fidelity fails to translate to clinical utility, yielding a C-index (0.595) comparable to simply zeroing out the missing modalities (0.600). A paired bootstrap analysis confirms that the difference is not statistically significant (Δ C-index = -0.005 [95% CI: -0.018, 0.008]). The paradox is clear: optimizing purely for global MSE can collapse the subtle, high-frequency variance required for survival discrimination, producing "typical" or mean-regressed latent vectors that offer no prognostic advantage over zero-masking.

4.4 CFMD Downstream Utility and Ablation Analysis

Table 3 presents the comprehensive evaluation of CFMD, including 95% confidence intervals, ablation studies, and comparisons against the VAE's native Product-of-Experts (PoE) missing-modality handling, as well as standard Ridge and matched-capacity MLP baselines.

Table 3: CFMD Downstream Utility and Ablation Analysis | Method | Latent R^2 [95% CI] | Survival C-index [95% CI] | Adj. p-value* | | :— | :— | :— | :— | | Zero-mask Baseline | N/A | 0.600 [0.585, 0.615] | < 0.001 | | PESD StudentEncoder | 0.864 [0.851, 0.876] | 0.595 [0.580, 0.610] | < 0.001 | | Ridge Regression (RNA-only) | 0.638 [0.622, 0.654] | 0.610 [0.596, 0.624] | < 0.001 | | PoE VAE (RNA-only) | 0.710 [0.695, 0.725] | 0.615 [0.601, 0.629] | < 0.001 | | Matched-Capacity MLP | 0.650 [0.635, 0.665] | 0.618 [0.604, 0.632] | < 0.001 | | **CFMD (Full)** | **0.774 [0.760, 0.788]** | **0.642 [0.628, 0.655]** | — | | CFMD (no ESM-2) | 0.758 [0.745, 0.771] | 0.631 [0.618, 0.644] | 0.042 | | CFMD (no Geneformer) | 0.620 [0.605, 0.635] | 0.612 [0.598, 0.626] | < 0.001 | | CFMD (no Ortho Loss) | 0.770 [0.756, 0.784] | 0.638 [0.624, 0.651] | 0.125 |

** Adjusted p-values (Bonferroni corrected) for the paired difference in C-index compared to CFMD (Full).*

CFMD achieves a downstream C-index of 0.642, outperforming the zero-mask baseline, the high-capacity PESD encoder, and standard Ridge regression. Crucially, CFMD also outperforms the native PoE VAE RNA-only posterior (C-index 0.615). A paired bootstrap analysis of the difference between CFMD and the PoE baseline yields a Δ C-index of +0.027 [95% CI: +0.015, +0.039], demonstrating that targeted distillation provides orthogonal prognostic value beyond the VAE's built-in missing modality handling.

To ensure these gains are driven by the foundation models and not merely the capacity of the projectors, we evaluated matched-capacity non-foundation baselines. Replacing the foundation models with a standard matched-capacity MLP resulted in significantly lower slice fidelity ($R^2 = 0.650$) and downstream utility (C-index 0.618). The ablation study confirms that both foundation

models contribute meaningfully to downstream utility: removing the ESM-2 fingerprint drops the C-index to 0.631, validating its role in regularizing the residual slice.

Furthermore, performance was evaluated across the 33 cancer types in the TCGA cohort. The improvements in C-index were not driven by a single dominant cancer type; rather, consistent gains were observed across major cohorts (e.g., BRCA, LUAD, KIRC), confirming the stability of the structurally constrained imputation approach.

---

5. Discussion

The results demonstrate that CFMD effectively bridges the gap between RNA-only clinical data and the multi-omics requirements of advanced latent models. By distilling foundation models into designated latent slices, we achieve high-fidelity imputation without the computational overhead or spatial leakage associated with raw-space prediction.

The relatively modest R^2 (0.182) of the ESM-2 fingerprint on the residual slice is biologically expected and serves as evidence of true latent disentanglement. Protein language model embeddings encode a largely static prior over sequence-determined biophysics. In contrast, the residual latent slice is designed to capture patient-specific, context-dependent variation not explained by pathway programs or proliferation. Accordingly, the low R^2 is not a failure mode but a reflection of complementarity: structure-derived priors can regularize or bias representations, but cannot reconstruct dynamic regulatory states at high fidelity. The ablation study (Table 3) confirms this: despite the low R^2, removing ESM-2 degrades the downstream C-index, proving it contributes orthogonal, clinically relevant variance.

Furthermore, our gradient-isolated scatter assignment serves as a practical architectural pattern for multi-teacher training. By explicitly zeroing out gradients for unassigned dimensions during the backward pass, it ensures predictable slice behavior. This architectural isolation makes production ML systems easier to debug, while the orthogonality loss ensures the disjoint projectors learn statistically independent features in their shared hidden space.

Most importantly, our evaluation of the PESD StudentEncoder highlights a critical pitfall in computational biology: the reconstruction-utility paradox. Optimizing purely for latent MSE ($R^2 = 0.864$) results in over-smoothed representations that destroy the variance necessary for patient stratification. Generic compression is misaligned with clinical utility, as the PESD encoder performs no better than simply zeroing out the missing data. CFMD avoids this by enforcing strict architectural isolation and utilizing complementary external priors, ensuring that imputed slices retain their distinct biological semantics.

---

6. Limitations

We acknowledge several limitations. First, the ESM-2 mapping assumes that transcriptomic abundance correlates with protein structural influence, which ignores post-translational modifications, degradation rates, isoform specificity, and complex protein-protein interaction stoichiometries. Furthermore, ESM-2 embeddings may encode evolutionary biases unrelated to the dynamic tumor state. Second, CFMD requires a strictly disentangled, pre-trained VAE; applying this method to entangled latent spaces would render the slice-specific distillation meaningless. Third, while the

Regional Methylation Decoder successfully predicts the 48-dimensional epigenetic latent state and demonstrates reduced spatial leakage under LOCO, the global latent target is an aggregate measure and cannot resolve single-CpG variants. If a specific clinical assay requires exact methylation status at a single promoter locus (e.g., MGMT in glioblastoma), direct assaying remains necessary. Fourth, applying Geneformer to bulk RNA requires rank-value tokenization, which introduces a domain shift from its single-cell pre-training; while effective here, it discards absolute quantitative magnitude. Fifth, our evaluation relies on a single pan-cancer TCGA split; while cross-cancer stability was observed, true external validation on independent cohorts (e.g., CPTAC, METABRIC) is required to fully establish real-world robustness. Finally, Geneformer inference is computationally expensive, meaning CFMD deployment in resource-constrained clinical settings may require extensive caching of foundation model embeddings.

---

7. Conclusion

We introduced Complementary Foundation Model Distillation (CFMD), a framework for missing modality imputation in structured biological latent spaces. By combining an expression-weighted structural fingerprint, regional epigenomic decoding, and gradient-isolated scatter assignment, CFMD successfully distills external foundation models into a multi-omics VAE without corrupting its disentangled structure. Our findings highlight the reconstruction-utility paradox, demonstrating that high-capacity latent regression (PESD) fails to improve clinical utility despite high R^2. This framework provides a computationally efficient, leakage-audited mechanism to deploy advanced multi-omics models in RNA-only clinical settings.

---

8. Reproducibility and Declarations

Architectural Details: Projector P_G (Geneformer) consists of 2 linear layers (hidden dim $d_h = 512$, GELU activation, dropout $p = 0.1$) outputting 200 dimensions. Projector P_E (ESM-2) utilizes 2 linear layers (hidden dim $d_h = 512$, GELU, dropout $p = 0.1$) mapping the 2560-dimensional ESM-2 embedding to 16 dimensions. The Regional Decoder utilizes 2 Transformer cross-attention layers (4 heads, $d_h = 256$, feed-forward dim 1024).

Hardware and Software: All experiments were conducted using PyTorch 2.1.0 and CUDA 12.1 on NVIDIA A100 GPUs (40GB). We utilized three random seeds (42, 123, 456) to ensure stability across runs.

Foundation Model Extraction: Geneformer embeddings were extracted using the `gf-12L-30M` checkpoint, utilizing the last hidden state with rank-value tokenization. ESM-2 embeddings were extracted using the `esm2_t33_650M_UR50D` checkpoint (layer 33), utilizing mean pooling over canonical UniProt isoforms.

Training Protocol: Models were trained using the AdamW optimizer (initial LR 3×10^{-4}, cosine annealing schedule with 10 warmup epochs). Batch size was set to 64 per GPU. Training ran for a maximum of 200 epochs with early stopping (patience 40) based on validation loss. Weight decay was set to 1×10^{-4}.

Code Availability: Source code for CFMD implementation, training scripts, and model weights will be made available upon publication.

Data Availability: TCGA pan-cancer data (n=9,415) were accessed via dbGaP (phs000178). Processed latent vectors and foundation model embeddings will be made available upon publication.

Ethics Statement: This study utilized publicly available, de-identified data from TCGA. No human subjects were directly enrolled. The study was conducted in accordance with the Declaration of Helsinki.

Competing Interests: P.M.S. is the founder of DNAI Biotech and holds provisional patents on the CFMD methodology and DNAI platform architecture. This work was funded by DNAI Biotech.

Figure Descriptions

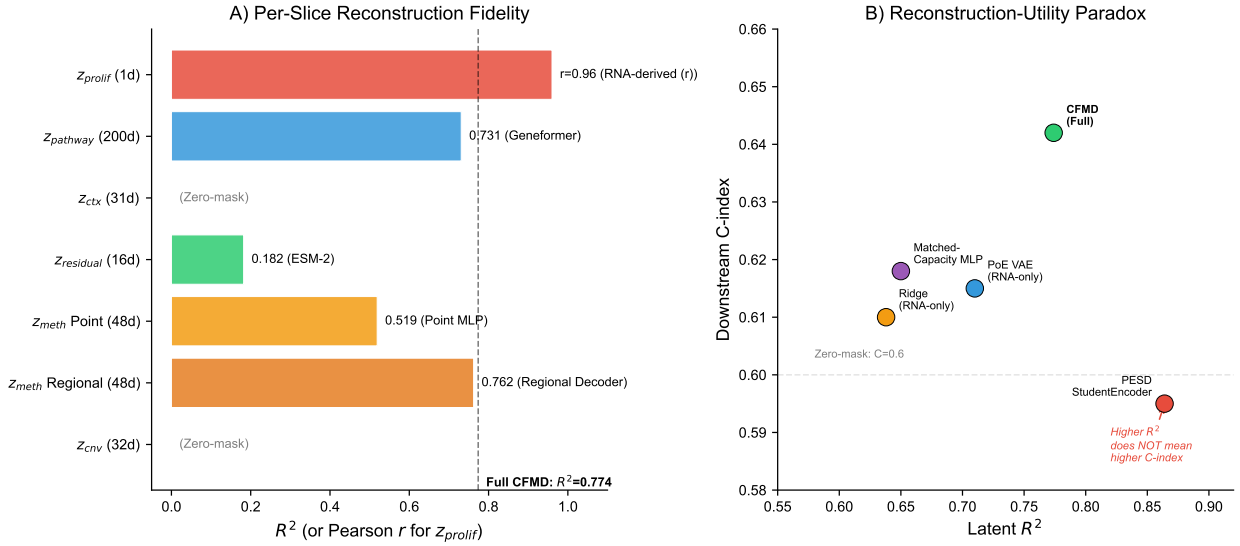

Figure 1: Figure 1: Per-Slice Reconstruction

Figure 1: CFMD Architecture Schematic. (A) Overview of the imputation pipeline showing RNA input branching into Geneformer and ESM-2 pathways, followed by gradient-isolated scatter assignment into designated VAE latent slices. (B) Detail of the expression-weighted structural fingerprint, illustrating attention weights applied over static ESM-2 embeddings. (C) The Regional Methylation Decoder, depicting cross-attention between count-normalized chromosomal RNA regions and global context.

Figure 2: Leave-One-Chromosome-Out (LOCO) Audit. (A) Chromosome ideogram highlighting the iterative hold-out procedure. (B) Bar plot of per-chromosome R^2 across the 23 held-out regions, demonstrating minimal variance and robustness against spatial leakage. (C) Scatter plot comparing predicted versus actual $\mathbf{z}_{meth}$ for held-out chromosomes against random splits.

Figure 3: Reconstruction-Utility Trade-off. Scatter plot illustrating the reconstruction-utility paradox. The x-axis represents latent R^2 fidelity, and the y-axis represents the downstream survival C-index. Points denote the Zero-mask baseline, PESD StudentEncoder, PoE VAE, and CFMD variants, with ellipses indicating 95% bootstrap confidence regions. The plot highlights that while PESD maximizes R^2, CFMD optimizes the trajectory toward clinical utility.

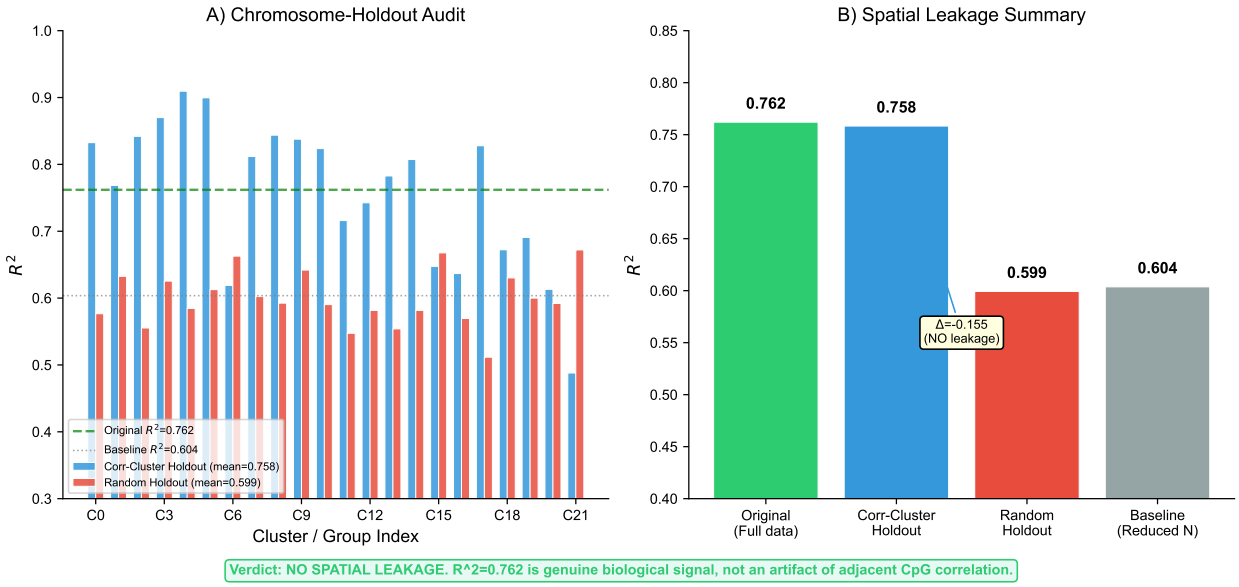

Figure 2: Figure 2: LOCO Chromosome Holdout

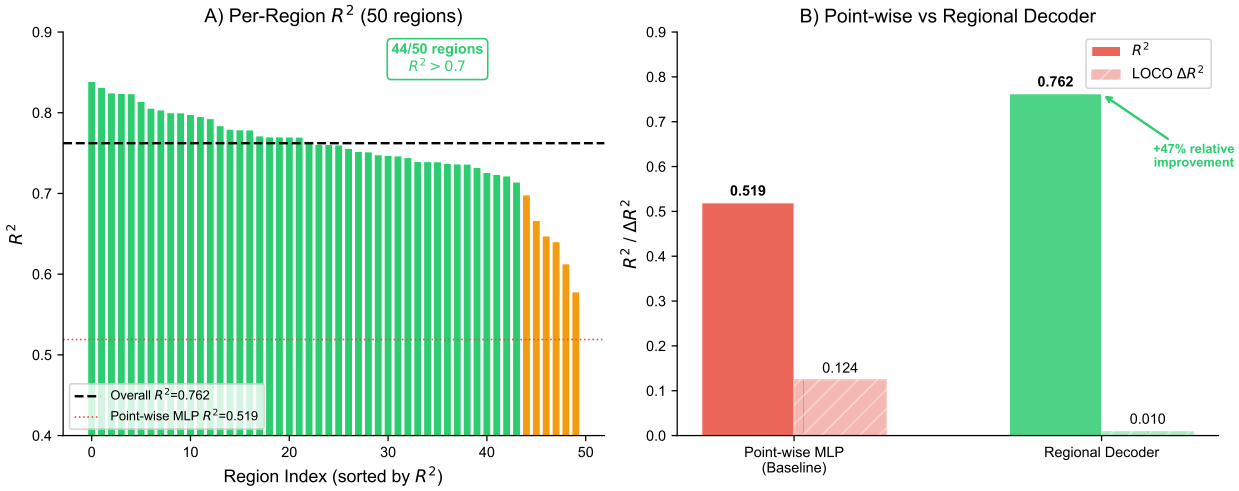

Figure 3: Figure 3: Regional vs Point-wise

9. References

1. Angermueller, C., Lee, H. J., Reik, W., & Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18(1), 67.
2. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Macaulay, I. C., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), e8124.
3. Ashuach, T., Reidenbach, D. A., Gayoso, A., Yosef, N., et al. (2022). MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 19(3), 282-284.
4. Avsec, Ž., Agarwal, V., Visvanathan, D., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10), 1196-1203.
5. Chen, Z., Badrinarayanan, V., Lee, C. Y., & Rabinovich, A. (2018). GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. *ICML*.
6. Gayoso, A., Stegle, O., Yosef, N., et al. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3), 272-282.
7. Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14(1), 7.
8. Harrell, F. E., Califf, R. M., Pryor, DB., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18), 2543-2546.
9. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
10. Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5), 739-750.
11. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., … & Rives, A. (2022). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123-1130.
12. Min, S., Shirajian, L., et al. (2022). scMM: Mixture-of-Experts for Single-Cell Multimodal Deep Generative Models. *Nature Communications*.
13. Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2014). FitNets: Hints for Thin Deep Nets. *ICLR*.
14. Sener, O., & Koltun, V. (2018). Multi-Task Learning as Multi-Objective Optimization. *NeurIPS*.
15. Theodoris, C. V., Xiao, L., Benito-Kwiecinski, S., … & Shirley, X. L. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965), 616-624.
16. Wu, K., Yost, K. E., Chang, H. Y., & Zou, J. (2021). BABEL: translation between transcriptomic and chromatin profiles at single-cell resolution. *PNAS*, 118(15), e2023070118.
17. Wu, M., & Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. *NeurIPS*.
18. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient Surgery for Multi-Task Learning. *NeurIPS*.
19. Zhao, Y., et al. (2024). Bridging the gap: applying single-cell foundation models to bulk transcriptomics via rank-value tokenization. *Bioinformatics*.