APTAMERBENCH: AUDITING GENERALIZATION ACROSS PATIENTS AND CROSS-MODAL CORRESPONDENCE IN PAIRED SINGLE-CELL PROFILING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate prediction in unseen patients does not by itself establish subtype discrimination within patients or the value of pairing measurements from the same cell. We introduce AptamerBench[1], a benchmark for cell typing from single-cell aptamer-binding (APT) profiles, with 361,792 paired cells across 40 participants, 293 aptamer features, and an operational RNA-derived hierarchy of five lineages and 27 subtypes. AptamerBench combines patient-disjoint evaluation, subject-balanced scoring, and controls retrained on shuffled APT profiles. Standard APT-only models achieve low subtype performance in new patients. With true-lineage decoding applied to both conditions, an MLP trained on real APT exceeds its matched control retrained on profiles shuffled within each patient and lineage by 0.160 in subject-balanced Macro-F1 (95% CI [0.132, 0.172]). This contrast remains positive under joint disease/acquisition-group holdout. High conditional scores can nevertheless coexist with poor discrimination within patients. Under the original patient-disjoint folds, real APT, shuffled APT, and context from other cells of the same patient each achieve approximately 0.87 subject-balanced Macro-F1 when distinguishing Memory CD8T-1 from NK cells, while the context model has zero minority-class recall in every patient. An RNA+APT gain over the RNA reference is not reproduced under group holdout, and the tested controls do not resolve an additional classification benefit from exact same-cell pairing. Label-independent retrieval nevertheless detects same-cell RNA–APT correspondence in held-out patients. An exploratory multicenter RNA/ADT analysis reproduces the same qualitative separation. AptamerBench provides reusable tests that distinguish generalization to new patients, conditional subtype association, within-patient discrimination, modality benefit, pairing-specific predictive value, and detectable cross-modal correspondence.

1 INTRODUCTION

High predictive performance in grouped multimodal data can arise for very different reasons. A model may generalize to observations from unseen groups while still failing to distinguish closely related instances within each group. Likewise, adding a second modality may improve prediction without demonstrating that preserving exact instance-level correspondence across modalities is itself useful. Conflating these quantities can lead to overly strong conclusions about cross-group generalization, fine-grained discrimination, or the value of multimodal pairing.

Clinical single-cell cohorts provide a particularly demanding instance of this problem: thousands of cells are nested within a much smaller number of patients and therefore share genetic, environmental, and technical structure (Zimmerman et al., 2021; Squair et al., 2021). Random cell-level splits can consequently produce optimistic estimates of generalization, motivating patient-disjoint evaluation (Fischer et al., 2024). Modern assays further provide paired transcriptomic and protein-related

[1]Code is available at https://anonymous.4open.science/r/AptamerBench-4851. The paired RNA and APT data are protected human-subject data. We plan to deposit them in a controlled-access repository after acceptance, pending final ethics and data-use approval, and will provide the accession number in the final manuscript.

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.

measurements from the same cell (Stoeckius et al., 2017; Hao et al., 2021), including barcode-linked aptamer measurements (Delley et al., 2018; Wu et al., 2025; Luo et al., 2026). Yet existing patient-disjoint and multimodal benchmarks typically report aggregate prediction scores that cannot distinguish within-patient discrimination, modality contribution, pairing-specific value, and detectable correspondence. AptamerBench is the first evaluation suite to operationalize and jointly test these quantities through matched estimands and controls.

We therefore argue that evaluation of paired single-cell prediction should separate several questions that are often treated as interchangeable: (i) whether a modality supports prediction in unseen patients; (ii) whether it contains subtype-associated information after patient and coarse cell-type composition are controlled; (iii) whether adding the modality improves prediction over an established reference modality; and (iv) whether that improvement specifically depends on preserving same-cell cross-modal correspondence. Detectable correspondence is itself a separate property: two modalities may retain recoverable same-cell information without that correspondence improving a particular downstream classification task.

To study these distinctions, we introduce AptamerBench, a patient-disjoint benchmark for single-cell aptamer-binding (APT) profiles paired with RNA measurements. AptamerBench contains 361,792 paired cells from 40 participants, 293 APT features, and RNA-derived operational labels organized into five parent groups and 27 terminal subtypes. The benchmark is organized around three evaluation questions. **Q1** measures ordinary APT-only prediction of parent and subtype in unseen patients using only the target cell's APT profile. **Q2** asks whether target-cell APT retains subtype-associated information after controlling for patient and parent composition, by comparing real APT with an independently trained patient-by-parent shuffled control under the same oracle true-parent decoding. **Q3** separates the predictive benefit of adding APT to RNA from the additional benefit, if any, of preserving exact same-cell RNA–APT correspondence. We further include a label-independent retrieval audit that directly tests whether paired RNA and APT profiles remain identifiable in held-out patients without using subtype, parent, or disease labels.

These analyses reveal several clear separations. Fine-grained APT-only prediction in unseen patients is weak, yet real APT still retains subtype-associated signal, outperforming the matched patient-by-parent shuffle by $+0.160$ in subject-balanced Macro-F1. Strong pooled scores can also mask poor within-patient recognition, including near-zero minority-class recall for some controls. Finally, although adding APT to RNA improves performance in the primary setting, this gain is not consistent across sensitivity analyses, and we do not resolve an additional classification benefit from preserving exact same-cell pairing. At the same time, label-independent retrieval clearly detects RNA–APT correspondence in held-out patients. The external multicenter RNA/ADT analysis reproduces this overall pattern. These separations motivate three contributions:

- **Estimand decomposition.** We distinguish generalization to unseen subjects, within-subject discrimination, conditional target information, modality benefit, and pairing-specific value as separate quantities in grouped multimodal prediction.
- **Matched diagnostics.** We design patient-disjoint splits, grouped shuffles, patient-context controls, and label-independent retrieval to probe these quantities with matched controls.
- **Empirical separations.** Across model families, hierarchy and label-granularity choices, group holdouts, and an external RNA/ADT cohort, we show that high pooled performance can coexist with weak within-subject recognition, while detectable cross-modal correspondence can coexist with no resolved pairing-specific classification advantage.

2 Related Work

Grouped evaluation, single-cell annotation, and hierarchy. Apt-seq (Delley et al., 2018) measures aptamer binding and gene expression in the same cell, while newer aptamer-based platforms extend this capability to clinical and perturbation settings (Wu et al., 2025; Luo et al., 2026). Related technologies such as CITE-seq (Stoeckius et al., 2017) jointly profile RNA and antibody-derived tags. For cell-type annotation, SingleR (Aran et al., 2019) and CellTypist (Domínguez Conde et al., 2022) transfer labels from annotated references, while prior benchmarks compare annotation (Abdelaal et al., 2019) and integration methods (Luecken et al., 2022) across datasets and technologies.

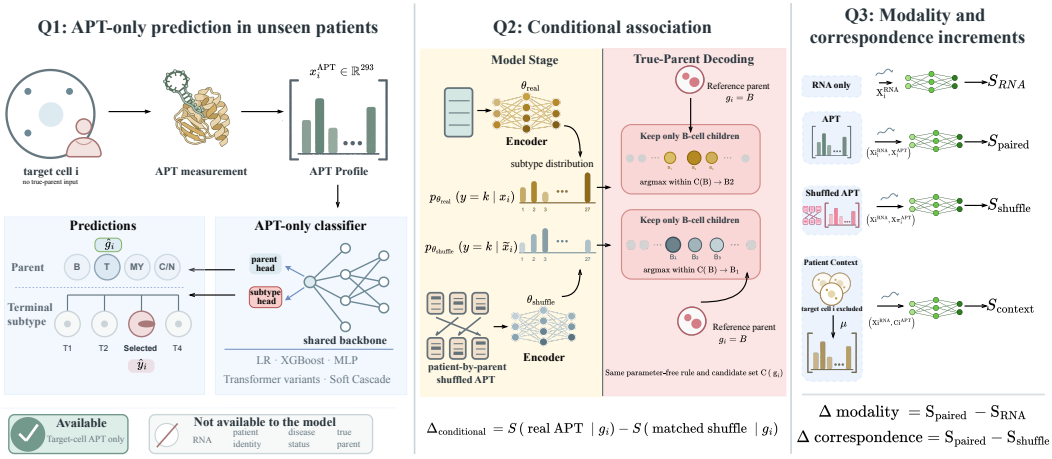


Figure 1: Overview of the AptamerBench evaluation design. **Q1.** Ordinary APT-only prediction uses only the target cell's APT profile from an unseen patient to predict its operational parent and terminal subtype. RNA, patient identity, disease status, and the reference parent are unavailable. **Q2.** Conditional subtype association compares real APT with a patient-by-parent shuffled control. The conditions are trained independently, and the same reference parent is used only during decoding to restrict predictions to its terminal children. **Q3.** Modality and correspondence analyses compare RNA alone, paired RNA and APT, grouped shuffled APT, and patient APT context. Patient context is constructed from other cells of the held-out patient while excluding the target cell.

More generally, evaluation is difficult when many observations are nested within fewer higher-level groups. In clinical single-cell cohorts, cells from the same patient share biological and technical structure, so cell-level splits can create pseudoreplication and optimistic estimates (Zimmerman et al., 2021; Squair et al., 2021). Patient-disjoint evaluation, as used in scTab (Fischer et al., 2024), addresses this leakage but does not identify whether pooled performance reflects between-patient structure or within-patient discrimination. Hierarchical methods including CHETAH (De Kanter et al., 2019), scHPL (Michielsen et al., 2021), and OnClass (Wang et al., 2021) exploit structured cell-type relationships. Unlike methods that use privileged information to improve deployable prediction (Vapnik & Vashist, 2009), AptamerBench uses reference parents only in controlled oracle diagnostics. Thus, our focus is not only whether prediction transfers across patients, but what that held-out-group performance actually demonstrates.

Multimodal contribution, interaction, and correspondence. WNN (Hao et al., 2021), MOFA+ (Argelaguet et al., 2020), totalVI (Gayoso et al., 2021), and MultiVI (Ashuach et al., 2023) integrate or jointly model multiple single-cell modalities, while multimodal benchmarks evaluate modality prediction, cross-modal alignment, and joint representation learning (Lance et al., 2022). More broadly, improved multimodal prediction need not imply useful interaction or dependence on instance-level pairing. EMAP (Hessel & Lee, 2020) compares multimodal models with additive unimodal projections, while SHAPE (Hu et al., 2022) and InterSHAP (Wenderoth et al., 2025) quantify modality contributions and interactions. AptamerBench connects these concerns by separating modality contribution, pairing-specific predictive value, and label-independent cross-modal correspondence, rather than treating them as interchangeable evidence of multimodal benefit.

3 AptamerBench: Data and Evaluation Design

Consider observations i nested within higher-level groups s_i, with prediction target y_i and paired modalities $x_i^{(A)}$ and $x_i^{(B)}$. Held-out-group prediction, within-group discrimination, conditional target information, modality contribution, and instance-level cross-modal correspondence are distinct predictive properties that cannot be inferred from a single aggregate score.

AptamerBench instantiates this setting with patients as groups, paired RNA and APT profiles as modalities, and RNA-derived operational cell types as targets. It contains 361,792 paired profiles

from 40 participants, with 293 APT features and 27 terminal subtypes, using patient-disjoint evaluation throughout. Q1 evaluates ordinary APT prediction in unseen patients, Q2 conditional subtype association after controlling for patient and parent composition, and Q3 modality benefit versus same-cell pairing-specific predictive value. Figure 1 summarizes the design; cohort, information-access, and partition details are provided in Appendices A and B.1.

3.1 Dataset and Label Hierarchy

Cohort and reference labels. The prediction targets are RNA-derived operational reference labels rather than error-free biological ground truth. Cohort processing, marker references, annotation records, and acquisition details are provided in Appendix A.

Hierarchies and label spaces. The primary hierarchy H_0 groups the 27 terminal labels into five operational parents: B, T, Myeloid, CD8T/NK, and Other. We additionally evaluate an expert-endorsed alternative hierarchy H_1 and a coarsened label space L_1 containing 15 terminal classes. These alternatives test whether the primary conditional result depends strongly on hierarchy choice or label granularity. Complete H_0, H_1, and L_1 mappings are provided in Appendix A.

Acquisition scope. All specimens were collected at one hospital and processed under a shared workflow. Disease and control groups were collected during distinct acquisition windows, so disease-associated and acquisition-associated effects cannot be separated. We therefore interpret AptamerBench as evaluating transfer to unseen patients within this cohort and protocol, rather than generalization across sites or acquisition periods.

3.2 Task Definitions

All cells from the same patient remain in the same partition (Zimmerman et al., 2021; Squair et al., 2021). For ordinary prediction, the model has access only to the target cell's input features. It does not receive patient identity, disease status, the reference parent, or measurements from other test cells. Some diagnostic analyses intentionally introduce privileged or transductive information to isolate specific quantities. We identify these settings explicitly below, and they should not be interpreted as deployment scenarios. Complete information-access conditions are listed in Appendix Table 3.

Q1: Prediction in unseen patients. Q1 asks whether the target cell's APT profile predicts its operational parent and terminal subtype in a patient excluded from model development. At test time, the model receives only the target-cell APT vector. We evaluate logistic regression, XGBoost (Chen & Guestrin, 2016), an MLP, and three Transformer-based models (Vaswani et al., 2017): a Flat Transformer, a hierarchical Transformer, and a Soft Cascade. All models use the same patient partitions, and none receives the reference parent as an ordinary input. Q1 measures ordinary APT-based transfer across representative model families rather than an unrestricted architecture search. Architecture and tuning details are provided in Appendix B.2.

Hierarchy-aware reference architectures. Because the terminal labels form an operational hierarchy, we include structured neural references that test whether explicitly modeling this structure improves ordinary APT-based prediction. The hierarchical Transformer allocates probability mass over parent and terminal nodes, whereas the Soft Cascade combines a global terminal classifier with parent-specific experts weighted by the model's predicted parent probabilities. Importantly, these models never observe the reference parent at inference time: the hierarchy must be inferred entirely from the target cell's APT profile. They therefore remain ordinary prediction models, in contrast to the privileged true-parent decoding used in Q2.

Inference decomposition. Q1 and Q2 define three progressively more controlled inference regimes. Flat ordinary prediction operates over the full terminal label space. Hierarchy-aware ordinary prediction can exploit the parent–child structure, but the relevant parent must still be inferred from the same target-cell APT profile. Finally, the Q2 oracle diagnostic supplies the reference parent only at decoding and restricts prediction to its terminal children. This progression separates the difficulty of learning and using hierarchical structure from the narrower question of whether APT

retains subtype-associated information once the parent is externally controlled; it does not represent a sequence of increasingly deployable models.

Q2: Conditional subtype association. Q2 asks whether target-cell APT retains subtype-associated predictive signal after patient and parent composition are controlled. We compare real APT with a matched control formed by permuting complete APT vectors within each patient-by-parent stratum. Terminal labels remain attached to their original cells, while the permutation breaks the original cell-to-APT correspondence within each stratum. This preserves patient and parent composition while disrupting finer cell-level APT–subtype association.

Let g_i denote the RNA-derived reference parent of cell i, and let $\mathcal{C}(g_i)$ denote the set of terminal subtypes belonging to that parent under the specified hierarchy. For either condition, let x_i denote the APT input presented to the model. For both real and shuffled conditions, g_i is used only at decoding to restrict predictions to $\mathcal{C}(g_i)$:

$$\hat{y}_i = \arg \max_{k \in \mathcal{C}(g_i)} p(y = k \mid x_i). \tag{1}$$

The real and shuffled conditions are trained and selected independently. For score S, we define

$$\Delta_{\text{conditional}}(S) = S(\text{real APT} \mid g_i) - S(\text{matched shuffle} \mid g_i). \tag{2}$$

A positive contrast indicates that real APT retains greater subtype-associated predictive information than the matched shuffle under the specified hierarchy. Because g_i is supplied at decoding, this is an oracle diagnostic rather than ordinary deployment performance. It should not be interpreted as a causal effect or evidence of a biological mechanism.

Q3: Modality benefit and same-cell correspondence. Q3 asks whether APT improves prediction beyond RNA alone and whether this gain depends on preserving the original same-cell RNA–APT pairing. Let S_{RNA}, S_{paired}, and $S_{\text{shuffle},G}$ denote the scores obtained using RNA alone, paired RNA+APT, and RNA+APT with APT shuffled within grouping G, respectively. We define

$$\Delta_{\text{modality}} = S_{\text{paired}} - S_{\text{RNA}}, \tag{3}$$

$$\Delta_{\text{correspondence},G} = S_{\text{paired}} - S_{\text{shuffle},G}, \tag{4}$$

where G specifies shuffling within either patient or patient-by-parent groups.

Δ_{modality} measures the predictive gain from adding APT to RNA, whereas $\Delta_{\text{correspondence},G}$ asks whether preserving the observed same-cell pairing improves prediction relative to the corresponding grouped-shuffle control. Because grouped shuffling can disrupt both APT–label association and cross-modal correspondence, $\Delta_{\text{correspondence},G}$ should be interpreted as a diagnostic contrast rather than a pure measure of cross-modal synergy or causality.

We additionally evaluate a patient-context control, with score S_{context}, constructed from APT measurements of other cells from the same held-out patient while excluding the target cell. This transductive control tests whether patient-level APT context can support similar predictive performance without observing the target cell's own APT profile. All Q3 conditions are independently trained and selected. Detailed constructions are provided in Appendix B.4.

Label-independent correspondence audit. The Q3 classification contrasts use an RNA-derived subtype target and therefore do not directly test whether same-cell RNA–APT correspondence is itself detectable. We examine this separately using an exploratory bidirectional retrieval task that uses no subtype, parent, or disease labels.

Dual encoders map RNA and APT profiles into a shared embedding space. The real condition is trained on observed same-cell pairs, whereas the control replaces each APT partner with a fixed-point-free permutation within the same training patient. At test time, all candidates are drawn from the same held-out patient, so patient identity alone cannot solve the task. For candidate-set size K, we define

$$\Delta_{\text{retrieval}}(K) = R@1_{\text{real}}(K) - R@1_{\text{shuffle}}(K). \tag{5}$$

The primary analysis uses $K = 64$ and patient-balanced Recall@1. Additional candidate sizes and retrieval metrics are reported in Appendix D.

Unified diagnostic view. Rather than collapsing multimodal performance into a single score, AptamerBench summarizes several distinct predictive properties through the diagnostic profile

$$\mathcal{D} = (S_{\text{ordinary}}, \Delta_{\text{conditional}}, \Delta_{\text{modality}}, \Delta_{\text{correspondence},G}, \Delta_{\text{retrieval}}(K)), \tag{6}$$

where S_{ordinary} denotes ordinary held-out-patient prediction performance. The remaining components respectively measure conditional subtype-associated information, modality contribution, pairing-specific classification value, and label-independent cross-modal correspondence. The components of $\mathcal{D}$ are diagnostic rather than additive: they use task-appropriate endpoints, are not averaged into a composite benchmark score, and a positive value in one component does not imply a positive value in another. Within-patient metrics provide a complementary view of whether the applicable classification signal supports discrimination among cells from the same patient.

3.3 Metrics and Statistical Inference

Our primary classification metric is subject-balanced (SB) pooled Macro-F1. To prevent patients with more cells from dominating the evaluation, we first normalize each patient's confusion matrix to equal total weight and then pool the normalized matrices before computing Macro-F1. We additionally report mean within-patient Macro-F1 to assess discrimination among cells from the same patient rather than across the pooled cohort.

For Q2, the primary metric is multichild true-parent SB Macro-F1, which restricts evaluation to parent groups with at least two terminal children. We complement this with support-aware within-patient-parent Macro-F1, computed only for patient–parent combinations containing at least two represented child classes. For the Memory CD8T-1 versus NK analysis, we additionally report within-patient balanced accuracy, AUROC, and minority-class recall.

All compared conditions use the same patient-disjoint folds and are selected and refitted independently. We compute confidence intervals using paired patient-level bootstrap resampling, with matched resamples and model runs across compared conditions (Field & Welsh, 2007). Group-holdout experiments use a separate leave-one-group-out protocol and are reported as robustness analyses rather than combined with the primary patient-disjoint evaluation. Complete metric definitions, eligibility rules, training procedures, and resampling details are provided in Appendices A.3 and B.1.

4 Results

Table 1 summarizes the primary Q1 and Q2 results, and Table 2 summarizes the Q3 multimodal comparisons. We first examine APT transfer and conditional subtype signal, then separate modality contribution from pairing-specific value, and finally evaluate robustness under stricter holdout and an external multicenter setting. Additional estimates and sensitivity analyses are provided in Appendix C.

4.1 APT Transfer and Conditional Signal

Ordinary transfer. Fine-grained APT-only prediction remains weak across all six model families under patient-disjoint evaluation. Fine-subtype SB Macro-F1 ranges from 0.134 to 0.148, while parent-level performance ranges from 0.296 to 0.342 (Table 1, Panel A). Differences among the tested architectures are small and unresolved, suggesting that the limited fine-grained transfer is not specific to a single classifier family.

Explicitly modeling the hierarchy does not materially alter this conclusion. The Soft Cascade obtains the highest fine-subtype point estimate (0.148), but its difference from the Flat Transformer is unresolved. Thus, the separation between ordinary Q1 prediction and Q2 oracle-conditional performance cannot be explained simply by the use of a flat classifier that ignores the hierarchy. The two settings instead differ critically in whether the relevant parent must be inferred from the target-cell APT profile or is externally controlled at decoding.

By comparison, changing the evaluation split has a substantially larger observed effect. Using the same MLP pipeline, random cell-level splitting increases fine-subtype SB Macro-F1 from 0.137 to 0.225. Thus, allowing cells from the same patients to appear in model development and testing

Table 1: Core AptamerBench results for ordinary APT prediction and conditional subtype association. Panel A reports patient-disjoint Q1 performance. Panel B compares real APT with the matched patient-by-parent shuffle under true-parent decoding. H_0 is the primary hierarchy, while H_1 and L_1 are sensitivity analyses. Differences are reported with paired 95% confidence intervals. Absolute scores are not directly comparable across label settings.

A. Ordinary APT prediction in unseen patients

Model	Fine SB Macro-F1 ↑	Parent SB Macro-F1 ↑
Logistic regression	0.134	0.296
XGBoost	0.134	0.326
MLP	0.137	0.333
Flat Transformer	0.147	0.342
Hierarchical Transformer	0.146	0.339
Soft Cascade	0.148	0.336

B. Conditional subtype association

Setting	Real	Shuffle	Pooled difference ↑	Within-patient-parent difference ↑
H0	0.376	0.216	$\mathbf{+0.160}$ $[\mathbf{0.132}, \mathbf{0.172}]$	$\mathbf{+0.100}$ $[\mathbf{0.086}, \mathbf{0.115}]$
H1	0.342	0.190	$+0.152$ $[0.124, 0.166]$	$+0.102$ $[0.090, 0.115]$
L1	0.471	0.330	$+0.141$ $[0.121, 0.156]$	$+0.136$ $[0.123, 0.149]$

produces substantially more optimistic performance than transfer to unseen patients. Together, these results show that ordinary APT-only prediction provides limited fine-grained generalization under the patient-disjoint setting, even when the classifier explicitly models the label hierarchy.

Conditional subtype signal. Weak ordinary transfer does not imply that APT lacks subtype-associated predictive information. Under the primary H_0 hierarchy, real APT achieves a multichild true-parent SB Macro-F1 of 0.376 compared with 0.216 for the matched patient-by-parent shuffle. The resulting conditional contrast is $+0.160$, with a 95% confidence interval of $[0.132, 0.172]$ (Table 1, Panel B). The corresponding support-aware within-patient-parent difference is $+0.100$.

The same pattern persists under the alternative hierarchy H_1 and the coarsened label space L_1, where the pooled contrasts remain positive at $+0.152$ and $+0.141$, respectively. Although the magnitude varies across operational parents, the overall contrast remains positive across these hierarchy and label-granularity sensitivity analyses. APT therefore retains subtype-associated predictive information after patient and parent composition are matched, even though ordinary fine-grained transfer remains weak.

Within-patient recognition. The conditional results above do not guarantee strong discrimination within individual patients. The Memory CD8T-1 versus NK comparison provides a clear example. Real APT, shuffled APT, and patient context all achieve approximately 0.87 pooled SB Macro-F1, with little separation in the pooled metric (Figure 2**a**).

The patient-level view is markedly different. The patient-context control has zero mean within-patient minority-class recall despite its high pooled score. Real APT shows a modest advantage over the matched shuffle in within-patient Macro-F1, but its minority-class recall remains only 0.0266. Thus, high pooled performance can coexist with poor recognition of a minority subtype within individual patients. This result motivates evaluating within-patient discrimination alongside subject-balanced pooled performance.

4.2 Modality and Pairing

Modality contribution. Adding APT to RNA produces a small improvement in the primary setting. With the 10,000-HVG MLP, SB Macro-F1 increases from 0.728 with RNA alone to 0.749 with paired RNA+APT, corresponding to $\Delta_{\text{modality}} = +0.0204$ (Table 2, Panel A).

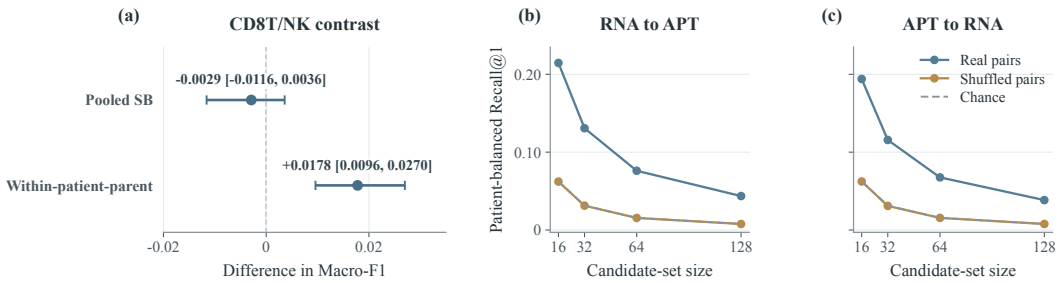


Figure 2: Patient-level diagnostics distinguish pooled performance, within-patient recognition, and label-independent correspondence. **a.** Real-minus-shuffle Macro-F1 contrasts for Memory CD8T-1 versus NK cells. Points show mean differences and paired 95% confidence intervals. **b and c.** Patient-balanced Recall@1 for bidirectional retrieval between RNA and APT. Shuffled training reassigns complete APT vectors within patients. Dashed curves show random-ranking chance.

Table 2: Q3 modality and pairing diagnostics. Panel A compares paired RNA+APT with RNA alone across evaluation settings. Panel B compares paired RNA+APT with pairing and patient-context controls under the primary 10k-HVG MLP. Differences are paired minus comparator and are reported with paired 95% confidence intervals. Intervals containing zero do not establish equivalence.

Setting or comparator	SB Macro-F1 (paired/comparator)	Difference [95% CI]
A. Modality contribution: paired RNA+APT minus RNA		
MLP, 10k HVGs	0.7488 / 0.7284	$+0.0204$ $[0.0004,\ 0.0374]$
MLP, 5k HVGs	0.724 / 0.727	-0.0037 $[-0.0146,\ 0.0165]$
scVI / totalVI	0.7507 / 0.7453	$+0.0054$ $[-0.0026,\ 0.0122]$
Disease/acq. holdout	0.67947 / 0.67948	-0.00001 $[-0.0084,\ 0.0060]$
B. Pairing controls: paired RNA+APT minus stated control		
Patient context	0.7488 / 0.7489	-0.00014 $[-0.01210,\ 0.00800]$
Patient-shuffled APT	0.7488 / 0.752	-0.0036 $[-0.0131,\ 0.0061]$
Patient$\times$parent shuffle	0.7488 / 0.751	-0.0017 $[-0.0059,\ 0.0039]$

The modality increment is inconsistent across settings: negative with 5,000 HVGs, small with scVI/totalVI, and approximately zero under joint disease/acquisition-group holdout. Because the operational targets are derived from RNA, the Q3 classification contrasts are target-specific and may favor the RNA reference; they should not be interpreted as a task-independent measure of APT information.

Pairing-specific classification value. The primary multimodal improvement does not establish that preserving exact same-cell pairing is responsible for the gain. Under the 10,000-HVG MLP, paired RNA+APT performs similarly to patient-shuffled APT, patient-by-parent-shuffled APT, and the patient-context control. All paired-minus-control intervals include zero (Table 2, Panel B).

The patient-context comparison is particularly informative. It achieves nearly the same pooled score using APT measurements from other cells of the held-out patient without observing the target cell's own APT profile. The exploratory scVI/totalVI comparisons reach the same qualitative conclusion. We therefore do not resolve an additional classification benefit from preserving exact same-cell pairing under the tested controls. These results do not establish equivalence and remain compatible with modest effects that the present cohort cannot resolve.

Same-cell correspondence is detectable. The lack of a resolved pairing-specific classification gain does not imply that RNA and APT lack same-cell correspondence. In the label-independent retrieval audit, real-pair training substantially outperforms shuffled-pair training in both directions. At $K = 64$, patient-balanced Recall@1 is 0.0761 for RNA-to-APT retrieval and 0.0676 for APT-

to-RNA retrieval, compared with 0.0155 in both directions after within-patient shuffled training (Figure 2**b,c**).

The real-pair advantage remains positive when distractors are selected to more closely match the true partner on basic QC characteristics (Appendix D.1). Because this sensitivity analysis uses the true partner's QC measurements to select distractors, it should be interpreted as a stress test rather than as evidence for the biological source of the correspondence.

Together, the classification and retrieval results separate two distinct claims. Same-cell RNA–APT correspondence is detectable in held-out patients, but its presence does not by itself establish an additional classification benefit from preserving exact pairing.

4.3 Robustness and External Evaluation

Joint disease and acquisition-group holdout. The main conditional finding persists under the stricter leave-one-group-out evaluation. The H_0 real-minus-shuffle contrast decreases from $+0.160$ under the primary patient folds to $+0.086$, but remains positive. In contrast, the multimodal increment is not reproduced. Paired RNA+APT and RNA alone achieve nearly identical pooled performance, with $\Delta_{\text{modality}} \approx 0$.

These results further separate conditional APT signal from multimodal classification gain. The former persists under the more demanding group shift, while the latter does not. Because disease groups coincide with acquisition windows, this analysis tests transfer across the observed joint groups without separating disease-associated from acquisition-associated effects.

Cross-assay framework transfer. We apply the same diagnostic decomposition to the independent Stephenson et al. RNA/ADT cohort (Stephenson et al., 2021): 639,482 paired cells from 118 subjects across three centers under leave-one-center-out evaluation. This setting changes the protein modality, cohort, label ontology, and evaluation grouping while preserving the evaluation logic. Because the released annotations used both RNA and surface-protein measurements, and the analysis followed inspection of the primary AptamerBench results, we treat it as exploratory framework transfer rather than biological replication of the APT findings.

The qualitative separation recurs. Under matched true-parent decoding, real ADT exceeds the independently retrained subject-by-parent shuffle by $+0.257$ in SB Macro-F1, with a within-subject-parent difference of $+0.284$; the contrast is positive in every held-out center. Paired RNA+ADT, however, shows no resolved pooled advantage over RNA alone or the grouped-shuffle controls. Label-independent retrieval detects same-cell correspondence, with real-minus-shuffle Recall@1 differences of $+0.317$ for RNA-to-ADT and $+0.262$ for ADT-to-RNA, both positive across held-out centers.

Thus, despite changes in assay, cohort, ontology, and grouping, the same diagnostic separation reappears: conditional subtype-associated information and detectable same-cell correspondence can coexist with an unresolved pairing-specific classification gain. This supports cross-assay portability of the framework without implying biological replication of the APT findings or direct comparison between APT and ADT assays. Complete analyses are reported in Appendices C.9, C.10, and F.

5 Conclusion

We introduced AptamerBench to separate predictive properties often conflated in grouped multimodal evaluation. APT-only subtype transfer to unseen patients is weak, yet real APT retains conditional subtype-associated signal relative to matched shuffles, while high pooled scores can mask poor within-patient recognition. Adding APT to RNA yields a small, setting-dependent gain, and our controls do not resolve additional classification value from exact same-cell pairing, although label-independent retrieval detects same-cell correspondence. The external RNA/ADT study reproduces this qualitative separation across assay, cohort, ontology, and grouping. More broadly, held-out-group prediction, within-group discrimination, conditional target information, modality contribution, pairing-specific value, and cross-modal correspondence are distinct properties that should be evaluated separately.

AI USE STATEMENT

Large language model tools assisted with manuscript organization, academic English editing, integration of completed result summaries into the text, and code for figures drawn from saved CSV results. The reported experimental scores and intervals were not generated by manuscript editing. The authors are responsible for verifying all source-to-claim links, scientific interpretations, and the final submission.

ETHICS STATEMENT

This study analyzes paired single-cell RNA and aptamer profiles derived from peripheral-blood specimens collected from 40 participants across five cancer groups and a healthy-control group. The study was conducted under approvals from the relevant institutional ethics committees. Protocol identifiers and institutional details are withheld during double-blind review and will be restored in the final version. Specimen collection and research use were conducted under the approved protocols.

All computational analyses used de-identified participant information. No direct personal identifiers were included in the benchmark data, and no attempt was made to re-identify participants. High-dimensional molecular profiles may nevertheless retain patient-specific information, so removal of direct identifiers does not eliminate all privacy risks.

The paired molecular data have not yet been deposited in a public repository. Subject to the applicable ethics approval and data-use requirements, we plan to deposit the shareable data in an appropriate controlled-access repository after acceptance. The final manuscript and public repository will provide the corresponding data accession or storage location when release is complete.

AptamerBench is intended for methodological research and benchmark evaluation, not for clinical diagnosis, treatment selection, or individual risk assessment. The cohort is small at the participant level and does not support claims of clinical validity, demographic fairness, or population-wide generalizability. Disease groups were processed during non-overlapping acquisition periods, so disease-associated and acquisition-associated effects cannot be separated in the available cohort. The RNA-derived subtype labels are operational reference annotations rather than independently measured biological ground truth. We therefore restrict our claims to predictive performance under the specified cohort, ontology, and patient-disjoint evaluation protocol.

REPRODUCIBILITY STATEMENT

The main paper and appendices specify the patient-disjoint splits, preprocessing, model configurations, selection rules, and evaluation metrics. An anonymized snapshot of the code is available at `https://anonymous.4open.science/r/AptamerBench-4851`, including the fixed patient folds, the operational ontology, experiment and evaluation scripts, and grouped prediction evaluators. These materials allow inspection of the reported protocols and re-running of the analyses once the data are available.

The paired RNA and APT data are protected human-subject data derived from peripheral-blood specimens and are not included in the anonymous repository. We plan to deposit them in a controlled-access repository after acceptance, pending final ethics and data-use approval, and will provide the accession number in the final manuscript.

Reproducing all AptamerBench experiments requires access to the cohort data and the complete training environment. Detailed provenance, compute, random seeds, model selection, and verification procedures are documented in Appendix E.

REFERENCES

Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology*, 20(1):194, 2019.

Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21(1):111, 2020.

Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature methods*, 20(8):1222–1231, 2023.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Jurrian K De Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, and Frank CP Holstege. Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *Nucleic acids research*, 47(16):e95–e95, 2019.

Cyrille L Delley, Leqian Liu, Maen F Sarhan, and Adam R Abate. Combined aptamer and transcriptome sequencing of single cells. *Scientific reports*, 8(1):2919, 2018.

César Domínguez Conde, Chao Xu, Louie B Jarvis, Daniel B Rainbow, Sara B Wells, Tamir Gomes, Sarah K Howlett, Ondrej Suchanek, Krzysztof Polanski, Hamish W King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

Can Ergen, Galen Xing, Chenling Xu, Martin Kim, Michael Jayasuriya, Erin McGeever, Angela Oliveira Pisco, Aaron Streets, and Nir Yosef. Consensus prediction of cell type labels in single-cell data with popv. *Nature Genetics*, 56:2731–2738, 2024. doi: 10.1038/s41588-024-01993-3.

Christopher A Field and Alan H Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3):369–390, 2007.

Felix Fischer, David S Fischer, Roman Mukhin, Andrey Isaev, Evan Biederstedt, Alexandra-Chloé Villani, and Fabian J Theis. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611, 2024.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods*, 18(3):272–282, 2021.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.62. URL https://aclanthology.org/2020.emnlp-main.62/.

Pengbo Hu, Xingyu Li, and Yi Zhou. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. *arXiv preprint arXiv:2205.00302*, 2022.

Christopher Lance, Malte D. Luecken, Daniel B. Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, et al. Multimodal single cell data integration challenge: Results and lessons learned. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pp. 162–176, 2022. URL https://proceedings.mlr.press/v176/lance22a.html.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Guoyan Luo, Jia Song, Yongbo Fu, Yuanjie Jiang, Yuan Gao, Zhixing Zhong, Ling Li, Yong Wei, Hao-Ran Jia, Lu Guo, et al. Spark-seq: a high-throughput platform for aptamer discovery and kinetic profiling. *Science*, 391(6780):eadv6127, 2026.

Christopher S McGinnis, Lyndsay M Murrow, and Zev J Gartner. Doubletfinder: Doublet detection in single-cell rna sequencing data using artificial nearest neighbors. *Cell Systems*, 8(4):329–337.e4, 2019. doi: 10.1016/j.cels.2019.03.003.

Lieke Michielsen, Marcel JT Reinders, and Ahmed Mahfouz. Hierarchical progressive learning of cell identities in single-cell data. *Nature Communications*, 12(1):2799, 2021.

Jordan W Squair, Matthieu Gautier, Claudia Kathe, Mark A Anderson, Nicholas D James, Thomas H Hutson, Rémi Hudelle, Taha Qaiser, Kaya JE Matson, Quentin Barraud, et al. Confronting false discoveries in single-cell differential expression. *Nature communications*, 12(1):5692, 2021.

Emily Stephenson, Gary Reynolds, Rachel A Botting, Fernando J Calero-Nieto, Michael D Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B Worlock, Masahiro Yoshida, et al. Single-cell multi-omics analysis of the immune response in covid-19. *Nature medicine*, 27 (5):904–916, 2021.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias, and Russ B Altman. Leveraging the cell ontology to classify unseen cell types. *Nature communications*, 12(1):5556, 2021.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21501–21509, 2025. doi: 10.1609/aaai.v39i20.35452.

Xiaoqiu Wu, Xinrui Lin, Xiangqi Ma, Huidong Huang, Dengwei Zhang, Yuqing Liu, Rui Cheng, Jia Song, Lifei Zhang, Yamin Tan, et al. High-throughput multiplexed quantification of molecules by aptamer sequencing (apt-seq) in single cells. *Journal of the American Chemical Society*, 147 (24):20394–20405, 2025.

Yu Yang, Xueyan Chen, Jieying Pan, Huiheng Ning, Yaojun Zhang, Yufei Bo, Xianwen Ren, Jiesheng Li, Shishang Qin, Dongfang Wang, et al. Pan-cancer single-cell dissection reveals phenotypically distinct b cell subtypes. *Cell*, 187(17):4790–4811, 2024.

Yuanhang Zhang, Xiaorui Zhou, Yu Zhong, Xi Chen, Zeyu Li, Rui Li, Pengfei Qin, Shanshan Wang, Jianhua Yin, Shang Liu, et al. Pan-cancer scrna-seq analysis reveals immunological and diagnostic significance of the peripheral blood mononuclear cells. *Human Molecular Genetics*, 33(4):342–354, 2024.

Kip D Zimmerman, Mark A Espeland, and Carl D Langefeld. A practical solution to pseudoreplication bias in single-cell studies. *Nature communications*, 12(1):738, 2021.

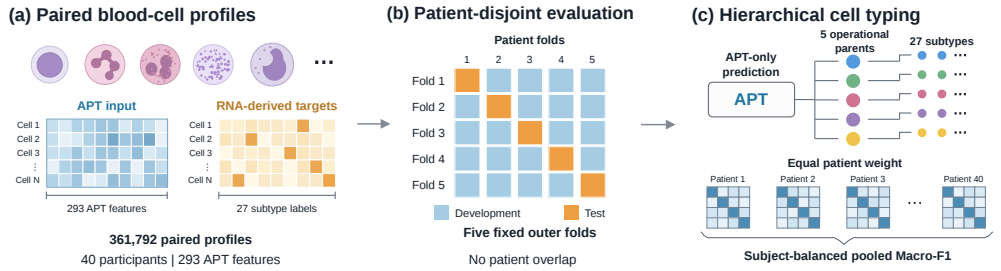


Figure 3: Overview of the AptamerBench cohort and evaluation structure. **a.** Paired RNA and APT measurements with RNA-derived subtype targets. **b.** Five patient-disjoint outer folds, with cyclic validation omitted. **c.** The H_0 hierarchy and equal patient weighting used for subject-balanced pooled Macro-F1.

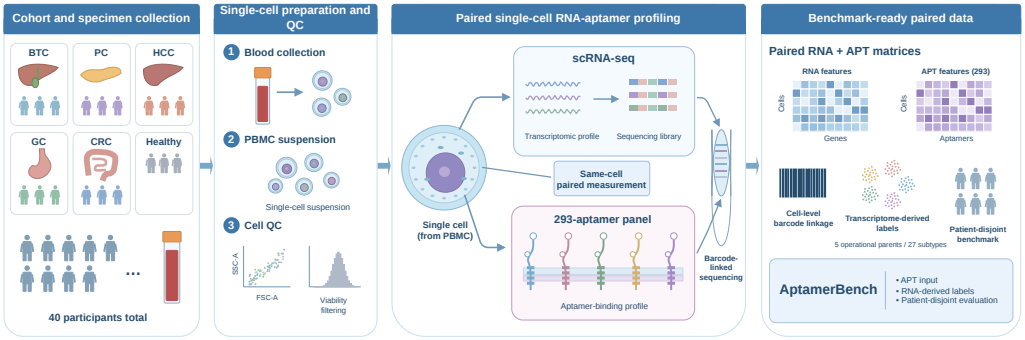


Figure 4: Construction of the paired AptamerBench cohort. Blood samples from 40 participants underwent PBMC isolation, single-cell quality control, and paired profiling of RNA and aptamer binding. Cell-level barcodes link transcriptomic measurements with a 293-aptamer panel. RNA-derived annotations define five operational parents and 27 terminal subtypes, yielding paired RNA and APT matrices for patient-disjoint evaluation. BTC, PC, HCC, GC, and CRC denote biliary tract, pancreatic, hepatocellular, gastric, and colorectal cancers, respectively. Participant icons are schematic and are not proportional to cohort size.

A Dataset Documentation and Evaluation Definitions

A.1 Cohort Processing and Reference Annotation

The cohort contains peripheral blood specimens from 40 participants across colorectal cancer, gastric cancer, hepatocellular carcinoma, pancreatic cancer, biliary tract cancer, and healthy controls. Each disease or control group contains between six and eight participants. RNA and aptamer measurements were obtained from the same cells. Figure 3 summarizes the paired measurements, fixed patient-disjoint partitions, operational hierarchy, and subject-balanced scoring. Figure 4 provides the complementary specimen-processing and benchmark-construction workflow.

RNA counts were processed with Seurat (Hao et al., 2021). Predicted doublets and low-quality cells were removed using the cohort processing workflow, which included DoubletFinder (McGinnis et al., 2019). After quality control, cell linkage, and exclusion of the operational Unknown cluster, the benchmark contains 361,792 cells with paired RNA profiles and 293 raw APT count features.

Prediction targets were constructed from RNA-derived clusters rather than from the APT measurements used for evaluation. Two annotators independently assigned initial cluster labels using canonical marker expression informed by peripheral blood and pan-cancer references (Zhang et al., 2024; Yang et al., 2024). A senior reviewer resolved ambiguous or discordant assignments and established the final cluster-level annotations.

The resulting labels are treated as operational reference targets. They are not assumed to constitute error-free biological ground truth. Cell-type annotation can depend on marker availability, clustering resolution, and reference label granularity (Fischer et al., 2024; Ergen et al., 2024). The complete annotation record includes the RNA-derived cluster identifier, both initial annotations, the senior adjudicated label, the operational parent, and positive marker summaries.

All specimens were collected at one hospital and processed fresh under a shared laboratory workflow. Disease or control groups were collected, prepared, and sequenced within distinct acquisition windows. Disease-associated and acquisition-associated variation therefore cannot be separated in the available cohort.

A.2 Operational Hierarchy and Label-Granularity

The primary hierarchy H0 assigns the 27 terminal labels to five operational parents named B, T, Myeloid, CD8T/NK, and Other. Myeloid combines monocyte and dendritic cell subtypes. Other contains Platelet and Erythrocyte. Under H0, CD8T/NK contains exactly Memory CD8T-1 and NK cell. Memory CD8T-2, Memory CD8T-3, Naive CD8T, MAIT-1, MAIT-2, and Proliferating T cell are assigned to T.

The alternative hierarchy H1 contains six parents. It retains the B and Myeloid definitions and separates CD4T from a broader cytotoxic CD8T/NK parent. The H1 CD8T/NK parent contains Memory CD8T-1, Memory CD8T-2, Memory CD8T-3, Naive CD8T, MAIT-1, MAIT-2, Proliferating T cell, and NK cell. Platelet and Erythrocyte form separate singleton parents. H1 preserves all 27 terminal labels.

The coarsened label space L1 uses the H1 parent structure. It merges terminal clusters only when their senior adjudicated names differ by a terminal numerical identifier. For example, Memory CD8T-1, Memory CD8T-2, and Memory CD8T-3 map to Memory CD8T. L1 preserves distinctions between naive and memory cells, CD4 and CD8 T cells, and CD14 and CD16 monocytes. It contains 15 terminal classes.

H0 defines the primary benchmark targets, candidate sets, and shuffle strata. H1 and L1 are exploratory sensitivity settings. Shared parent membership under H1 does not imply biological equivalence among its children. Label merging under L1 defines a coarser prediction target rather than a claim that the merged clusters are biologically identical.

A.3 Evaluation Metrics and Support-Aware Scoring

Macro-F1 can refer to different aggregation procedures. We therefore define the averaging unit, supported label set, and order of aggregation for every reported metric.

Subject-balanced pooled Macro-F1. For patient p, let $C^{(p)} \in \mathbb{R}^{K\times K}$ be the confusion matrix over K evaluated classes, where rows denote reference labels and columns denote predictions. Let

$$n_p = \sum_{i=1}^{K}\sum_{j=1}^{K} C_{ij}^{(p)} \tag{7}$$

be the evaluated cell count for patient p. With P patients, subject-balanced pooled Macro-F1 is

$$\overline{C} = \frac{1}{P}\sum_{p=1}^{P}\frac{C^{(p)}}{n_p}, \qquad \mathrm{SBMacroF1} = \frac{1}{K}\sum_{k=1}^{K}\frac{2\overline{C}_{kk}}{\sum_j \overline{C}_{kj} + \sum_i \overline{C}_{ik}}. \tag{8}$$

A zero denominator contributes zero. Each patient contributes equal total mass before classwise F1 is computed. This differs from cell-weighted pooled Macro-F1, which pools cell counts without patient normalization, and mean within-patient Macro-F1, which computes F1 before averaging patients.

Mean within-patient Macro-F1. Let $\mathcal{P}_{\text{eligible}}$ denote the patients eligible for a specified diagnostic label set. Mean within-patient Macro-F1 is

$$\text{WithinPatientMacroF1} = \frac{1}{|\mathcal{P}_{\text{eligible}}|} \sum_{p \in \mathcal{P}_{\text{eligible}}} \text{MacroF1}\Big(C^{(p)}\Big). \tag{9}$$

Macro-F1 is computed separately within each patient before patients are averaged. Correct predictions in one patient therefore cannot compensate for a class-specific failure in another patient.

Balanced accuracy and minority recall. For the Memory CD8T-1 versus NK cell diagnostic, a patient is eligible when both reference classes are present. Within-patient balanced accuracy is the mean of Memory CD8T-1 recall and NK cell recall within that patient. The reported score averages the resulting values across eligible patients.

Minority recall identifies the less frequent reference class within each patient and computes recall for that class. A patient without both classes is ineligible. A tied class count does not define a minority class and is excluded only from minority recall. Within-patient Macro-F1, balanced accuracy, and minority recall measure complementary aspects of within-patient discrimination.

Multichild true-parent Macro-F1. True-parent decoding restricts the terminal prediction for cell i to the child set $\mathcal{C}(g_i)$ of its reference parent g_i,

$$\widehat{y}_i^{\text{oracle}} = \arg\max_{k \in \mathcal{C}(g_i)} p_\theta(k \mid x_i). \tag{10}$$

The reference parent is unavailable during ordinary deployment, so this score is a privileged diagnostic. If a parent has only one terminal child, supplying the parent deterministically reveals the terminal label. The primary multichild score therefore includes only leaves belonging to parents with at least two children. Full hierarchy oracle scores that include singleton parents are reported separately.

Support-aware within-patient-parent Macro-F1. For patient p and parent g, let $C^{(p,g)}$ denote the confusion matrix after true-parent decoding within g. Define the supported child set as

$$\mathcal{K}_{pg} = \left\{ k \in \mathcal{C}(g) : \sum_j C_{kj}^{(p,g)} > 0 \right\}. \tag{11}$$

A patient and parent unit is eligible when

$$|\mathcal{K}_{pg}| \geq 2. \tag{12}$$

For an eligible unit, support-aware Macro-F1 is

$$\text{SAF1}_{pg} = \frac{1}{|\mathcal{K}_{pg}|} \sum_{k \in \mathcal{K}_{pg}} \frac{2C_{kk}^{(p,g)}}{\sum_j C_{kj}^{(p,g)} + \sum_i C_{ik}^{(p,g)}}. \tag{13}$$

The final score averages SAF1_{pg} with equal weight across eligible patient and parent units.

A child absent from a patient has no local recall to estimate. Assigning that class an F1 value of zero would convert missing support into an apparent classification failure. The support-aware definition therefore omits unsupported children from the class average. A prediction of an observed cell into an unsupported child remains an error for the observed reference class and continues to reduce its recall and F1.

Support-aware balanced accuracy uses the same eligibility rule and averages recall over the supported children. The Memory CD8T-1 versus NK cell analysis is a two-class special case in which eligible patients contain both children.

Table 3: Information available at test time under each evaluation condition. Patient and donor identifiers are used only to define partitions, context sets, shuffle groups, or within-patient retrieval candidate sets. They are not provided as model inputs.

Question	Condition	Test information	Access type
Q1	Ordinary APT	Target-cell APT	Ordinary
Q2	Real conditional APT	Target-cell APT with true-parent decoding	Privileged
Q2	Matched shuffled APT	Patient-by-parent shuffled APT with true-parent decoding	Privileged
Q3	RNA only	Target-cell RNA	Ordinary
Q3	Paired RNA and APT	Target-cell RNA and its paired target-cell APT	Ordinary
Q3	Group-shuffled RNA and APT	Target-cell RNA and an APT vector shuffled within the specified group	Diagnostic
Q3	RNA with patient context	Target-cell RNA and APT context from other cells of the held-out patient, excluding the target cell	Transductive
Pairing audit	Label-independent same-cell retrieval	Target-cell RNA or APT query and candidate profiles from the opposite modality drawn only from the same held-out patient. No subtype, parent, or disease labels are used	Diagnostic

A.4 Information Access and Statistical Inference

Table 3 separates ordinary target-cell inputs from privileged true-parent decoding and transductive patient context. This distinction is part of the estimand. In particular, true-parent results do not represent deployable single-cell annotation, and patient-context conditions require other cells from the held-out patient.

Paired bootstrap intervals. Cells are not treated as independent resampling units. Each bootstrap replicate samples patients with replacement and includes all evaluated cells belonging to each sampled patient. The same patient draws are used for both conditions in a contrast.

When conditions contain matched training seeds, the same seed slots are sampled across conditions. Shuffle realization slots are matched or resampled according to the corresponding experimental protocol. Each fitted run is scored separately before scores are averaged over shuffle realizations and training seeds. Model probabilities are not averaged before nonlinear metrics are computed.

Unless otherwise stated, intervals use 5,000 paired bootstrap replicates and report the 2.5th and 97.5th percentiles. These intervals quantify uncertainty over the observed patients and completed repeated runs. They condition on the fixed cohort, reference labels, patient partitions, completed model selection, tested model family, and available training and shuffle repetitions.

B Detailed Experimental Protocols

B.1 Shared Patient Partitions and Training Procedure

Patient partitions. The primary evaluation uses 5 frozen disease-stratified folds with eight patients per fold. All cells from a patient remain in the same fold. For outer test fold f, fold $(f + 1) \bmod 5$ supplies eight validation patients, and the remaining 24 patients form the inner fit set. The test fold is excluded from preprocessing estimation, feature selection, model selection, early stopping, and epoch selection.

After validation selection, preprocessing estimators and model parameters are reinitialized and fitted on all 32 development patients. The resulting model is evaluated once on the eight held-out patients. This separation prevents test performance from influencing model selection.

Preprocessing. All learned transformations are estimated from the current fit or refit partition. Raw APT counts are transformed using $\log(1 + x)$ and standardized with fit partition means and standard deviations. RNA profiles are normalized to a total count of 10,000 per cell and log transformed. Highly variable genes and RNA scaling parameters are selected using only the relevant fit partition.

For shuffled conditions, complete raw APT vectors are permuted before APT transformations are estimated or applied. For patient context conditions, context vectors are constructed separately

within each fit, validation, refit, and test partition. No validation or test statistics are used to fit preprocessing transformations.

Selection and refitting. Neural model selection uses training seed 17. Candidate configurations are selected by validation subject-balanced fine Macro-F1. Training stops after at most 50 epochs with patience eight. The selected configuration and epoch count are freshly refitted with seeds 17, 29, and 43.

Each experimental condition performs its own validation selection and fresh refit. This includes real APT, shuffled APT, RNA only, paired RNA and APT, shuffled RNA and APT, and RNA with patient context. A condition does not inherit fitted preprocessing parameters, selected hyperparameters, or stopping decisions from another condition.

Run-level scoring. Every fitted run is scored separately. Model probabilities are not averaged across training seeds or shuffle realizations before nonlinear metrics are computed. Scores are first averaged across shuffle realizations associated with the same training seed and are then averaged across training seeds.

Paired uncertainty. Unless otherwise stated, confidence intervals use 5,000 paired bootstrap replicates. Patients are sampled with replacement, and all evaluated cells belonging to a sampled patient are retained. The same patient draws and training seed slots are used across the conditions in each contrast. Shuffle slots are additionally resampled when the protocol contains crossed shuffle realizations.

The reported intervals are two-sided percentile intervals based on the 2.5th and 97.5th percentiles. They condition on the observed cohort, fixed partitions, completed model selection, finite training seeds, and finite shuffle realizations.

Group-holdout scope. The original folds hold out individual patients while retaining development patients from each disease and acquisition group. The leave-one-group-out analysis instead holds out an entire joint disease and acquisition group. Disease and acquisition cannot be separated because they are aligned in the cohort.

Results from the two regimes are reported separately. They are not pooled, and their difference is not interpreted as a formal test between evaluation regimes.

B.2 Q1 Reference Models

All Q1 models receive the same 293 target-cell APT features and use the shared patient partitions and preprocessing protocol described in Appendix B.1.

Logistic regression. Multinomial logistic regression uses the LBFGS solver with at most 1,000 iterations and no class weighting. The regularization parameter is selected from

$$C \in \{0.01, 0.1, 1, 10\}. \tag{14}$$

Parent and terminal subtype classifiers select C independently.

XGBoost. XGBoost (Chen & Guestrin, 2016) uses histogram-based tree construction. Tree depth is selected from $\{3, 6\}$, learning rate from $\{0.03, 0.1\}$, and tree count from $\{100, 300, 500\}$. Row and feature subsampling are both set to 0.8. Parent and terminal subtype models are selected independently.

Multilayer perceptron. The multilayer perceptron contains hidden layers of width 256 and 128 with ReLU activation and dropout 0.1. Separate heads predict the terminal subtype and operational parent. Training minimizes terminal cross-entropy plus 0.5 times parent cross-entropy.

Transformer references. The Flat Transformer, hierarchical Transformer, and Soft Cascade use an APT identity and value tokenizer followed by two attention layers. The hidden dimension is 128, the feedforward dimension is 256, and each layer uses four attention heads and dropout 0.1.

The Flat Transformer predicts terminal and parent labels using separate output heads. The hierarchical Transformer allocates probability mass over parent and terminal nodes and derives parent predictions from subtree mass. The Soft Cascade combines a global terminal classifier with parent-specific experts weighted by predicted parent probabilities. None of these models receives the reference parent as an ordinary input.

Neural optimization. Neural models use AdamW with learning rate selected from

$$\left\{3 \times 10^{-4}, 10^{-3}\right\} \tag{15}$$

and weight decay selected from

$$\left\{10^{-5}, 10^{-3}\right\}. \tag{16}$$

The multilayer perceptron uses batch size 1,024. Transformer models use batch size 128. Complete selection records and selected configurations are retained with the released evaluation artifacts.

Random-cell split sensitivity. After completing and inspecting the canonical patient-disjoint Q1 analysis, we conducted an exploratory split-sensitivity analysis using the same MLP pipeline. We assigned cells to five folds using stratified five-fold splitting over the terminal subtype labels with split seed 2027. For outer test fold f, fold $(f+1) \bmod 5$ supplied validation cells and the remaining three folds supplied the inner fitting set. After model selection, the selected configuration was freshly refitted on all four non-test folds with seeds 17, 29, and 43.

APT preprocessing, model architecture, loss, optimization, hyperparameter search, early stopping, and scoring were unchanged from the canonical Q1 protocol. Preprocessing statistics were estimated only from the applicable fitting or refitting cells. Unlike the patient-disjoint evaluation, all 40 patients were represented in the fitting, validation, and test portions of each random-cell outer fold. Patient identifiers were used only for grouped evaluation and were not classifier inputs.

B.3 Q2 Matched Conditional Comparison

Real and shuffled conditions. The primary Q2 analysis uses the 256 and 128 unit multilayer perceptron with separate terminal and parent heads. The real condition uses each target-cell's original raw APT vector.

For the matched condition, let p_i denote the patient of target-cell i, and let g_i denote its configured parent. The shuffled input is

$$\widetilde{x}_i^{\mathrm{APT}} = x_{\pi(i)}^{\mathrm{APT}}, \qquad p_{\pi(i)} = p_i, \qquad g_{\pi(i)} = g_i, \tag{17}$$

where π is a random permutation within each patient and parent stratum.

Complete APT vectors are permuted rather than individual features. Target identifiers and reference labels remain fixed. Shuffling is performed independently within the inner fit, validation, outer refit, and test partitions. It therefore does not move APT vectors across patients, parents, or data partitions.

Independent model development. Real and shuffled conditions undergo independent preprocessing estimation, hyperparameter selection, early stopping, and fresh refitting. Shuffle realizations 101, 211, and 307 are crossed with final training seeds 17, 29, and 43. Every crossed run is scored separately.

Both conditions use the same reference parent only during diagnostic decoding. For cell i, the predicted terminal subtype is

$$\widehat{y}_i^{\text{oracle}} = \arg\max_{k \in \mathcal{C}(g_i)} p_\theta(k \mid x_i), \tag{18}$$

where $\mathcal{C}(g_i)$ is the child set of reference parent g_i. The reference parent is not provided as an input feature.

H1 and L1 sensitivities. H1 repeats the complete training procedure after reconstructing parent targets, shuffle strata, and oracle candidate sets from the H1 hierarchy. Its primary conditional score excludes the singleton Platelet and Erythrocyte parents.

L1 uses the H1 parent structure and directly trains the model on the 15 class coarsened label space. L1 predictions are not obtained by merging probabilities from a model trained on the original 27 labels. Real and shuffled L1 conditions perform independent selection and refitting.

These analyses were conducted after the primary H0 setting had been established. They are exploratory sensitivity analyses rather than independent confirmation or evidence of invariance across possible hierarchy definitions.

B.4 Q3 Original-Fold RNA and APT Protocol

RNA and APT preprocessing. The original-fold multimodal analysis uses 10,000 highly variable genes selected from the current fit partition. RNA counts are normalized to 10,000 counts per cell and transformed using $\log(1+x)$. Selected RNA features are scaled without centering. APT counts are log transformed and standardized using fit partition statistics. RNA and APT representations are concatenated after their separate transformations.

All input conditions use a multilayer perceptron with hidden widths 256 and 128, ReLU activation, dropout 0.1, and separate terminal and parent heads. Each condition independently selects the AdamW learning rate and weight decay using the grid defined in Appendix B.2.

Multimodal conditions. The RNA only model receives the target-cell RNA representation. The paired model receives the target-cell RNA representation and its original APT vector. Shuffled multimodal models receive the target-cell RNA representation together with a complete APT vector permuted within-patient or patient and parent groups.

RNA profiles, target identifiers, and reference labels remain attached to the original target cells. APT shuffling is performed separately within the fit, validation, refit, and test partitions. Patient and parent shuffling uses the reference parent only to construct permutation groups.

The original-fold multimodal controls use fixed shuffle realizations paired with final training seeds 17, 29, and 43. These paired mappings are retained during bootstrap resampling rather than treated as fully crossed repetitions.

Patient context. For target cell i from patient p in partition d, patient-wide APT context is

$$c_i^{\text{APT}} = \frac{1}{|\mathcal{I}_{pd}| - 1} \sum_{\substack{j \in \mathcal{I}_{pd} \\ j \neq i}} \log\left(1 + x_j^{\text{APT}}\right), \tag{19}$$

where $\mathcal{I}_{pd}$ denotes the cells from patient p contained in partition d. The target cell is excluded from its own context. Context construction does not use terminal subtype or parent labels.

Context scaling parameters are fitted only on the corresponding fit or refit partition. At evaluation time, the context condition accesses other cells belonging to the held-out patient. It is therefore transductive and is not an ordinary single-cell deployment condition.

scVI/totalVI architecture sensitivity. As an exploratory architecture sensitivity, we repeated the Q3 modality and correspondence contrasts using scVI and totalVI on the same frozen patient-disjoint folds. The protocol was fixed before any new outer-test scores were computed.

For each fold, the analysis reused the corresponding 10,000 fit-only HVGs, while scVI and totalVI received the selected genes as raw RNA counts. totalVI additionally received the 293 raw APT counts. Patient identity was not registered as a batch key or covariate.

We compared four independently selected and refitted conditions: scVI with RNA only, totalVI with paired RNA and APT, totalVI with APT shuffled within patient, and totalVI with APT shuffled within patient and reference parent. The latter is privileged because the reference parent defines the permutation strata. Complete APT vectors were permuted within each fit, validation, refit, and test partition.

The latent dimension was selected from $\{20, 40\}$ using validation SB Macro-F1. Posterior-mean latent representations were decoded with a fixed unweighted multinomial logistic regression ($C = 1$). Each condition performed independent selection and fresh refitting with training seeds 17, 29, and 43. We report both subject-balanced pooled Macro-F1 and mean within-patient support-aware Macro-F1 with paired patient-level bootstrap intervals.

B.5 Joint Disease and Acquisition-Group Holdout

The leave-one-group-out analysis defines six joint disease and acquisition groups. Each outer iteration withholds every patient from one group. Model selection uses only the remaining groups, after which the selected condition is refitted on all development groups and evaluated on the held-out group. The conditional real and patient-by-parent-shuffled APT conditions repeat the Q2 training and diagnostic decoding under these partitions. The conditions using RNA alone, paired RNA and APT, shuffled RNA and APT, or RNA with context likewise repeat their own selection and refitting. The held-out group contributes neither preprocessing estimates nor feature selection decisions.

This protocol tests transfer across the observed joint groups. Because each disease group was collected in its own acquisition window, it cannot identify disease and acquisition effects separately. The six outer results are also not treated as six independent cohort replications.

B.6 Earlier Donor-Disjoint RNA and ADT Evaluation

The external analysis uses 90,261 paired RNA and ADT profiles from nine donors. All cells from a donor remain in the same outer partition, with a cyclic validation donor and the remaining donors used for fitting. The evaluation retains the external 45-label ontology and jointly informed reference annotations. It is therefore a protocol transfer study rather than biological validation of the APT cohort.

Conditions using RNA alone, paired RNA and ADT, or shuffled RNA and ADT are selected and refitted independently at 2,000 and 5,000 RNA features. The most restricted correspondence control permutes complete ADT vectors within donor, sample, and reference-parent groups. The parent used to form these strata and for conditional decoding makes the contrast privileged.

After the fixed-configuration analysis, a bounded exploratory sensitivity allowed each condition to select among the anchor MLP, a lower learning rate, stronger weight decay, and a wider MLP. ADT conditions also selected between log1p and the specified cell-local centered representation. Selection used only the validation donor, followed by fresh development-donor refits with seeds 17, 29, and 43. The shuffled condition crossed three independently generated whole-vector permutations with all training seeds.

C Detailed Results and Sensitivity Analyses

This appendix reports the detailed estimates and sensitivity analyses underlying Section 4. Confidence intervals use the bootstrap procedure defined in Appendix A.3. For contrasts between matched conditions, resampling is paired across patients and matched run indices. Metrics are computed separately for each fitted run before averaging across training seeds and shuffle realizations.

Table 4: Ordinary APT-only prediction in unseen patients. Brackets denote paired 95% confidence intervals.

Model	Fine SB Macro-F1	Parent SB Macro-F1
Logistic regression	$0.134\ [0.114, 0.146]$	$0.296\ [0.276, 0.315]$
XGBoost	$0.134\ [0.118, 0.148]$	$0.326\ [0.304, 0.348]$
MLP	$0.137\ [0.121, 0.152]$	$0.333\ [0.307, 0.356]$
Flat Transformer	$0.147\ [0.127, 0.162]$	$0.342\ [0.317, 0.365]$
Hierarchical Transformer	$0.146\ [0.126, 0.160]$	$0.339\ [0.316, 0.362]$
Soft Cascade	$0.148\ [0.129, 0.161]$	$0.336\ [0.311, 0.359]$

Table 5: Q1 split-sensitivity analysis for the MLP. Random-cell evaluation permits cells from the same patient to appear in model development and testing, whereas the canonical patient-disjoint evaluation holds out complete patients. Brackets denote 95% confidence intervals from patient-level bootstrap resampling. Differences use paired patient and training-seed resampling across the two protocols.

Evaluation protocol	Test-patient overlap	Fine SB Macro-F1	Parent SB Macro-F1
Random-cell	40/40	0.2252 [0.1792, 0.2356]	0.3974 [0.3651, 0.4236]
Patient-disjoint	0/8	0.1370 [0.1215, 0.1521]	0.3328 [0.3076, 0.3563]
Random-cell minus patient-disjoint	Not applicable	**+0.0882 [0.0341, 0.1010]**	**+0.0646 [0.0288, 0.1013]**

C.1 Q1 Reference-Model Results

Table 4 reports ordinary APT-only performance for all evaluated model families. All models use the same patient partitions, information access, and fit-only preprocessing protocol.

Random-cell split sensitivity. Table 5 compares the canonical patient-disjoint Q1 MLP with the exploratory random cell-level split analysis. Random-cell splitting increases fine SB Macro-F1 by $+0.0882$ and parent SB Macro-F1 by $+0.0646$. All test patients are represented during model development under the random-cell split, whereas the patient-disjoint folds have no overlap between training and test patients. The comparison measures sensitivity to evaluation partitioning rather than the mechanism underlying the performance gap.

For fine subtype prediction, the Flat Transformer exceeds the MLP by $+0.0100\ [-0.0047, 0.0234]$. The Soft Cascade exceeds the Flat Transformer by $+0.0011\ [-0.0093, 0.0093]$. Both intervals include zero, so these comparisons do not establish a superior architecture among the tested neural models.

C.2 Q2 Parent-Specific Results

Table 6 reports the H0 conditional contrast by operational parent. Pooled differences use true-parent restricted decoding. Within-patient differences use support-aware within-patient-parent Macro-F1. The parent-specific results show that the aggregate H0 contrast is not a uniform effect. In particular, CD8T/NK has an unresolved pooled contrast but a small positive within-patient contrast. The estimate for Other is based on only six eligible patients and consequently has substantially greater uncertainty.

C.3 Hierarchy and Label-Granularity Sensitivities

Table 7 reports matched conditional contrasts under H0, H1, and L1. Each setting repeats model selection, fitting, shuffling, oracle decoding, and scoring using its own parent and terminal label definitions.

For H0, there are no globally singleton operational parents, so the full hierarchy oracle and multichild oracle endpoints coincide. The support-aware within-patient-parent analyses additionally exclude patient-by-parent units in which fewer than two child classes are represented. Thus, the

Table 6: H0 conditional subtype association by operational parent. Eligible patients contain at least two supported children of the corresponding parent.

Parent	Real	Shuffle	Pooled difference	Within-patient difference	Eligible patients
B	0.469	0.182	$+0.288$ $[0.229, 0.321]$	$+0.205$ $[0.176, 0.236]$	40
Myeloid	0.265	0.111	$+0.155$ $[0.125, 0.182]$	$+0.101$ $[0.086, 0.118]$	40
CD8T/NK	0.869	0.872	-0.003 $[-0.011, 0.004]$	$+0.018$ $[0.010, 0.027]$	40
Other	0.844	0.496	$+0.349$ $[0.036, 0.424]$	$+0.287$ $[0.125, 0.442]$	6
T	0.225	0.136	$+0.089$ $[0.072, 0.102]$	$+0.048$ $[0.041, 0.055]$	40

Table 7: Conditional subtype association under the primary hierarchy H0, the alternative hierarchy H1, and the coarsened label space L1.

Setting	Multichild pooled difference	Support-aware within-patient-parent difference	Balanced accuracy difference
H0	$+0.1596$ $[0.1315, 0.1727]$	$+0.1000$ $[0.0863, 0.1145]$	$+0.0809$ $[0.0681, 0.0942]$
H1	$+0.1517$ $[0.1244, 0.1660]$	$+0.1018$ $[0.0897, 0.1149]$	$+0.0813$ $[0.0700, 0.0927]$
L1	$+0.1411$ $[0.1214, 0.1564]$	$+0.1361$ $[0.1232, 0.1490]$	$+0.1057$ $[0.0925, 0.1187]$

reported H0 conditional contrast is not driven by deterministic decoding of singleton parents or by treating locally absent child classes as classification errors.

Table 8 provides the absolute real and shuffled scores for H1 and L1. Full hierarchy oracle scores include singleton parents. Multichild oracle scores exclude them.

Absolute scores across H0, H1, and L1 should not be interpreted as direct model comparisons because the candidate sets, shuffle strata, and terminal label spaces differ.

C.4 Support-Aware and Patient-Heterogeneity Analyses

Table 9 reports the detailed Memory CD8T-1 versus NK cell analysis. The patient context condition uses other cells from the same held-out patient and excludes the target cell.

The context model predicts one hard class within every patient and run, with mean within-patient minority recall equal to zero. Its high pooled score therefore reflects patient-level composition rather than target-cell discrimination. Across patients, the majority-class fraction for the Memory CD8T-1 versus NK task ranges from 0.684 to 0.997.

Real APT nevertheless retains limited patient-internal signal: it improves within-patient Macro-F1 and AUROC relative to the matched shuffle, although its hard minority recall remains low. The reference composition is also strongly structured across patients and joint disease and acquisition groups (Figure 5). Memory CD8T-1 is the majority class in 32 patients, whereas NK cell is the majority in eight patients, including all six healthy-control patients and two colorectal-cancer patients.

C.5 Original-Fold RNA and APT Results

Table 11 reports the primary original-fold multimodal comparisons using the early-concatenation MLP. Contrast direction is stated explicitly because modality and correspondence comparisons use different reference conditions.

The 10,000 feature representation shows a positive paired modality increment. The corresponding shuffled APT and patient context comparisons do not establish an advantage from preserving the exact target-cell pairing. The 5,000 feature analysis does not reproduce the positive modality increment, which further limits claims of robustness across RNA representations.

scVI/totalVI architecture sensitivity. Because the primary Q3 analysis uses early-concatenation MLP fusion, we additionally repeated the modality and correspondence comparisons using scVI and

Table 8: Absolute conditional scores under H1 and L1. Ordinary scores do not use true parent decoding. Oracle and patient parent scores are privileged diagnostics.

Setting	Readout	Real	Shuffle	Difference
H1	Ordinary fine SB Macro-F1	0.1390	0.0755	$+0.0635$
H1	Ordinary parent SB Macro-F1	0.2881	0.2796	$+0.0085$
H1	Full hierarchy oracle SB Macro-F1	0.3902	0.2498	$+0.1404$ $[0.1151, 0.1536]$
H1	Multichild oracle SB Macro-F1	0.3416	0.1899	$+0.1517$ $[0.1244, 0.1660]$
H1	Support-aware within-patient-parent Macro-F1	0.2592	0.1574	$+0.1018$ $[0.0897, 0.1149]$
H1	Support-aware balanced accuracy	0.2727	0.1914	$+0.0813$ $[0.0700, 0.0927]$
L1	Ordinary fine SB Macro-F1	0.2016	0.1343	$+0.0673$
L1	Ordinary parent SB Macro-F1	0.2841	0.2815	$+0.0026$
L1	Full hierarchy oracle SB Macro-F1	0.5415	0.4193	$+0.1222$ $[0.1052, 0.1354]$
L1	Multichild oracle SB Macro-F1	0.4711	0.3300	$+0.1411$ $[0.1214, 0.1564]$
L1	Support-aware within-patient-parent Macro-F1	0.4589	0.3228	$+0.1361$ $[0.1232, 0.1490]$
L1	Support-aware balanced accuracy	0.4680	0.3623	$+0.1057$ $[0.0925, 0.1187]$

Table 9: Pooled and within-patient results for Memory CD8T-1 versus NK cell. One-class runs count fitted runs that predict a single hard class within a patient.

Condition	Pooled SB Macro-F1	Mean within-patient AUROC	Minority recall	One-class runs
Real target-cell APT	0.8692	0.6259	0.0266	38 of 120
Patient-by-parent shuffled APT	0.8721	0.4970	0.0007	222 of 360
Patient context	0.8730	Not applicable	0.0000	120 of 120

Real minus shuffle pooled SB Macro-F1 is -0.0029 $[-0.0116, 0.0036]$. The corresponding mean within-patient Macro-F1 difference is $+0.0178$ $[0.0096, 0.0270]$. The mean within-patient AUROC difference is $+0.1290$ $[0.1033, 0.1546]$.

totalVI on the same patient-disjoint folds. This analysis was exploratory and its protocol was fixed before any new scVI/totalVI outer-test scores were computed. Table 12 reports the results. Paired totalVI improves mean within-patient Macro-F1 over the scVI RNA reference by $+0.0103$ (95% CI $[0.0037, 0.0173]$), whereas the pooled SB Macro-F1 interval for the modality increment includes zero ($+0.0054$, 95% CI $[-0.0026, 0.0122]$). In contrast, neither the patient-shuffle nor the privileged patient-by-parent-shuffle comparison establishes an additional pairing-specific classification advantage: the pooled and within-patient confidence intervals all include zero. Thus, the qualitative distinction between modality contribution and pairing-specific classification value is also observed with scVI/totalVI, although the magnitude and statistical evidence for the modality increment depend on the model and evaluation metric.

C.6 Q3 Uncertainty and Patient-Level Heterogeneity

The pooled Q3 pairing-control intervals include zero, which does not establish an additional classification benefit from preserving exact pairing between RNA and APT but also does not establish

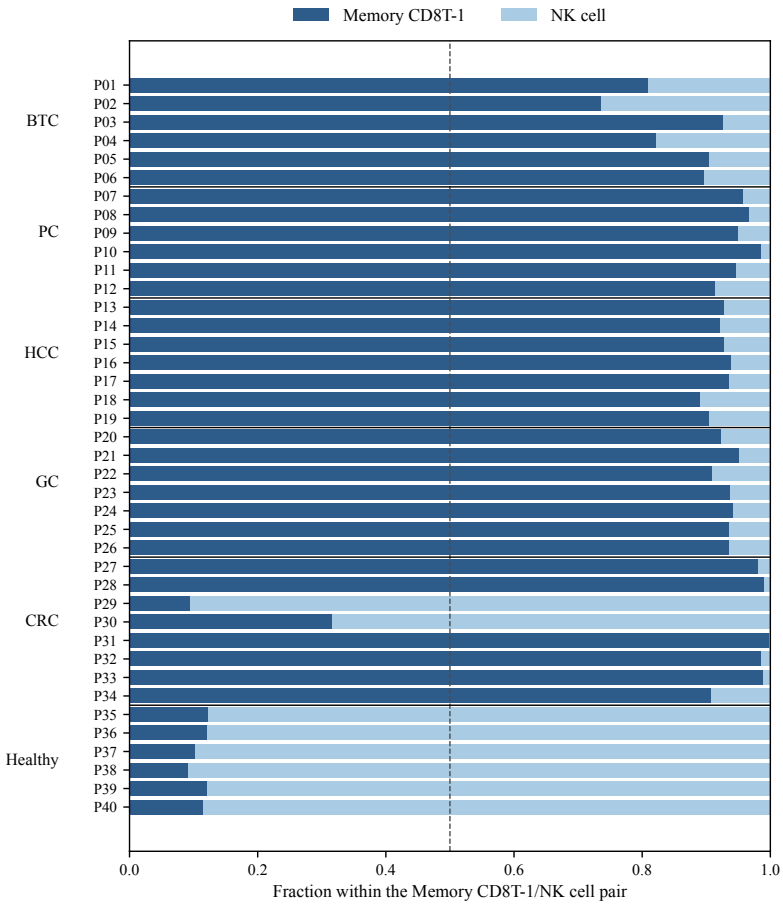


Figure 5: Memory CD8T-1 and NK-cell composition within the operational CD8T/NK group for all 40 anonymized patients, grouped by joint disease and acquisition stratum. Bar lengths show each subtype's fraction among cells in the evaluated pair, and the dashed line marks equal composition. The 32-to-8 majority split and the concentration of NK-cell-majority patients are descriptive and do not separate disease-associated from acquisition-associated variation.

equivalence. To make the remaining uncertainty explicit, Table 13 reports the positive effect sizes still contained within the corresponding 95% confidence intervals. These values are descriptive compatibility limits rather than prespecified practical-equivalence margins. No applicable smallest effect size of interest was specified before this analysis.

Across all three controls, patient-level differences span both sides of zero rather than concentrating uniformly at zero (Figure 6). For patient-shuffled APT, 20 of 40 patients have positive differences and 20 have negative differences. The corresponding proportions are 47.5% versus 52.5% for the patient-by-parent shuffle and 55% versus 45% for patient context.

The mean within-patient differences remain close to zero: -0.0035 $[-0.0086, 0.0020]$ for patient shuffle, -0.0007 $[-0.0063, 0.0045]$ for patient-by-parent shuffle, and -0.0005 $[-0.0069, 0.0061]$ for patient context. Thus, the unresolved pooled contrasts coexist with modest patient-level heterogeneity rather than demonstrating uniformly absent pairing-related differences.

Table 10: Confidence intervals for the Q3 contrasts summarized in Table 2. Panel A reports paired RNA+APT minus RNA. Panel B reports paired RNA+APT minus the stated control under the primary 10,000-HVG MLP setting. All entries use pooled SB Macro-F1. Point estimates in the main table are rounded for display. Intervals that include zero are unresolved and do not establish equivalence.

Panel	Setting or control	Contrast	Paired 95% confidence interval
A	MLP, 10,000 HVGs	$+0.0204$	$[0.0004, 0.0374]$
A	MLP, 5,000 HVGs	-0.0037	$[-0.0146, 0.0165]$
A	scVI / totalVI	$+0.0054$	$[-0.0026, 0.0122]$
A	Joint disease/acquisition holdout	-0.00001	$[-0.0084, 0.0060]$
B	Patient context	-0.00014	$[-0.01210, 0.00800]$
B	Patient-shuffled APT	-0.0036	$[-0.0131, 0.0061]$
B	Patient-by-parent-shuffled APT	-0.0017	$[-0.0059, 0.0039]$

Table 11: Original-fold RNA and APT results. Brackets denote paired 95% confidence intervals.

RNA features	Condition	SB Macro-F1	Difference	Contrast direction
5,000	RNA	0.727	Reference	Not applicable
5,000	Paired RNA and APT	0.724	-0.0037 $[-0.0146, 0.0165]$	Condition minus RNA
10,000	RNA	0.72837	Reference	Not applicable
10,000	Paired RNA and APT	0.74879	$+0.02042$ $[0.0004, 0.0374]$	Condition minus RNA
10,000	RNA with patient context	0.74893	$+0.02056$ $[-0.00073, 0.03675]$	Condition minus RNA
10,000	RNA with matched noise	0.727	$+0.0220$ $[0.0008, 0.0400]$	Paired minus condition
10,000	RNA with patient shuffled APT	0.752	-0.0036 $[-0.0131, 0.0061]$	Paired minus condition
10,000	RNA with patient-by-parent shuffled APT	0.751	-0.0017 $[-0.0059, 0.0039]$	Paired minus condition
10,000	Paired RNA and APT versus patient context	Not applicable	-0.00014 $[-0.01210, 0.00800]$	Paired minus context

Table 12: Exploratory scVI/totalVI architecture sensitivity for Q3. All conditions use the same patient-disjoint folds and independently selected and freshly refitted models. Brackets denote paired 95% confidence intervals. The patient-by-parent shuffle is privileged because its permutation strata use the reference parent.

Condition / contrast	Pooled SB Macro-F1	Within-patient Macro-F1
scVI RNA	0.7453	0.5926
totalVI paired RNA+APT	0.7507	0.6029
totalVI patient-shuffled APT	0.7478	0.6022
totalVI patient×parent-shuffled APT	0.7486	0.6010
Paired − RNA	+0.0054 [-0.0026, 0.0122]	**+0.0103 [0.0037, 0.0173]**
Paired − patient shuffle	+0.0029 [-0.0039, 0.0099]	+0.0008 [-0.0037, 0.0051]
Paired − patient×parent shuffle	+0.0021 [-0.0016, 0.0065]	+0.0020 [-0.0025, 0.0070]

Table 13: Uncertainty in the original-fold Q3 contrasts. The final column reports the upper limit of the corresponding 95% confidence interval to indicate the magnitude of positive gain still compatible with the completed analysis. These values are not prespecified equivalence margins.

Contrast	Estimate	95% CI	Upper positive compatibility limit
Modality contrast			
Paired − RNA	$+0.0204$	$[0.0004, 0.0374]$	$+0.0374$
Pairing/control contrasts			
Paired − patient context	-0.00014	$[-0.01210, 0.00800]$	$+0.0080$
Paired − patient shuffle	-0.0036	$[-0.0131, 0.0061]$	$+0.0061$
Paired − patient×parent shuffle	-0.0017	$[-0.0059, 0.0039]$	$+0.0039$

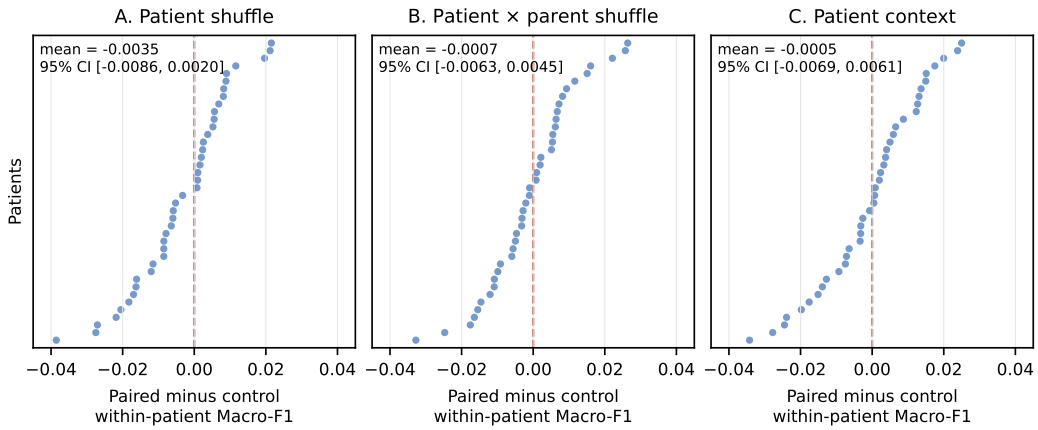


Figure 6: Patient-level heterogeneity in the original-fold Q3 pairing controls. Each point shows the within-patient Macro-F1 difference between paired RNA+APT and the indicated control for one held-out patient, after the frozen run-level aggregation procedure. Patients are ordered independently within each panel for visualization. The dashed vertical line denotes zero difference. Panel means and paired patient-bootstrap 95% confidence intervals are shown above each distribution. Positive and negative patient-level differences coexist for all three controls. These patient-level endpoints are complementary heterogeneity diagnostics and do not algebraically decompose the pooled subject-balanced Macro-F1 contrasts.

Table 14: Training-weight and APT-representation sensitivities. Oracle and matched-shuffle columns use privileged true-parent decoding. AUROC is computed within each eligible patient for the two-child CD8T/NK group and then averaged. The raw-log1p, unweighted row reuses the primary Q2 predictions and point estimate, but its interval follows this table's targeted-follow-up bootstrap: each sampled training-seed slot draws one shuffle realization. The primary Q2 interval instead averages three shuffle realizations within each sampled training-seed slot. The two intervals therefore differ despite sharing predictions and a point estimate.

Input/training	Ordinary fine	Oracle fine	Real minus shuffle Δ	CD8T/NK AUROC
Raw log1p, unweighted	0.137	0.376	0.160 [0.130, 0.172]	0.626
Raw log1p, patient-balanced	0.136	0.376	0.156 [0.128, 0.169]	0.625
Raw log1p, class-balanced	0.126	0.364	0.116 [0.088, 0.146]	0.610
Median-depth normalized	0.141	0.389	0.171 [0.136, 0.184]	0.613
Two count summaries	0.037	0.192	0.005 [0.003, 0.009]	0.554
Cell-local CLR	0.141	0.383	0.165 [0.134, 0.181]	0.612

Table 15: Summary of group-holdout and external scope checks. Δ_{cond}, Δ_{mod}, and Δ_{corr} denote conditional, modality, and correspondence contrasts, respectively. The external correspondence contrast uses donor-by-sample-by-parent shuffling with multichild true-parent scoring. Brackets denote paired 95% confidence intervals.

Setting	Estimand	Difference (95% CI)
Joint disease and acquisition-group holdout		
H0 conditional	Δ_{cond}	$+0.086$ [0.062, 0.099]
Multimodal	Δ_{mod}	0.000 [-0.008, 0.006]
External RNA/ADT		
2,000 RNA features	Δ_{corr}	$+0.051$ [0.035, 0.067]
5,000 RNA features	Δ_{corr}	$+0.036$ [0.021, 0.051]

C.7 Representation and Weighting Sensitivities

The primary raw-log1p representation was compared with patient-balanced and class-balanced training, median-depth normalization, a mathematically defined cell-local centered log representation, and two count summaries. For raw counts x_i, the centered representation is

$$\mathrm{CLR}_{ij} = \log(x_{ij} + 1) - \frac{1}{293} \sum_{k} \log(x_{ik} + 1). \tag{20}$$

The median-depth target is estimated only from the applicable fit partition. The two technical summaries are $\log(1 + \text{total APT count})$ and the number of detected APT features. Table 14 shows that the matched conditional contrast remains positive across the tested full-vector representations and weighting schemes. The count summaries produce a much smaller contrast. These checks limit simple count-depth explanations but do not isolate a biological mechanism or establish robustness to untested transformations.

C.8 Scope Summary: Group Holdout and External Data

Table 15 summarizes the principal scope checks reported in Sections C.9 and C.10. The group-holdout analyses test robustness to the observed joint disease and acquisition-group shift, whereas the external RNA/ADT analysis evaluates whether the same diagnostic framework can recover correspondence signal in a second paired single-cell dataset. These analyses target different generalization questions and should not be interpreted as direct cross-dataset effect-size comparisons.

Under joint disease and acquisition-group holdout, the H0 conditional contrast remains positive, whereas the original-fold multimodal increment is not reproduced. In the external RNA/ADT analysis, the paired-versus-shuffled correspondence contrast is positive under both tested RNA feature

Table 16: Results under joint disease and acquisition-group holdout. Brackets denote paired 95% confidence intervals.

Analysis	SB Macro-F1	Difference	Contrast direction
H0 conditional subtype association	Not applicable	$+0.086$ $[0.062, 0.099]$	Real minus shuffle
RNA	0.67948	Reference	Not applicable
Paired RNA and APT	0.67947	-0.00001 $[-0.0084, 0.0060]$	Condition minus RNA
RNA with patient-shuffled APT	Not reported	-0.0046 $[-0.0180, 0.0065]$	Paired minus condition
RNA with patient-by-parent-shuffled APT	Not reported	-0.0052 $[-0.0174, 0.0077]$	Paired minus condition
RNA with patient context	0.65099	-0.02849 $[-0.05047, -0.01232]$	Condition minus RNA
Paired RNA and APT versus patient context	Not applicable	$+0.02848$ $[0.01387, 0.04533]$	Paired minus context

budgets. The detailed absolute scores, control comparisons, and scope limitations are reported below.

C.9 Joint Disease and Acquisition-Group Holdout Results

The leave-one-group-out protocol holds out each joint disease and acquisition group in turn. It therefore evaluates a different generalization regime from the original patient-disjoint folds, and estimates from the two protocols are reported separately rather than pooled.

The H0 conditional contrast remains positive under this group shift, supporting the direction of the conditional APT association beyond the original patient folds.

The multimodal result is less stable. RNA and paired RNA and APT obtain nearly identical pooled scores, with a paired-minus-RNA contrast of -0.00001 $[-0.0084, 0.0060]$. Thus, the positive modality increment observed under the original folds is not reproduced under joint disease and acquisition-group holdout. Neither grouped shuffle establishes a pairing-specific disadvantage relative to the paired condition.

Patient context performs below RNA under this shift (-0.02849 $[-0.05047, -0.01232]$), and paired RNA and APT correspondingly exceeds patient context by $+0.02848$ $[0.01387, 0.04533]$. This differs from the original-fold setting, in which patient context attained a score similar to paired RNA and APT.

In an exploratory within-patient analysis using a fixed ontology, RNA, paired RNA and APT, and patient context obtain scores of 0.52023, 0.52135, and 0.50310, respectively. The context-minus-RNA difference is -0.01712 $[-0.02400, -0.00996]$. Across held-out groups, the paired-minus-RNA difference ranges from -0.0287 for healthy controls to $+0.0184$ for hepatocellular carcinoma, indicating heterogeneity across the observed joint groups.

These results should be interpreted as a robustness limitation for the original-fold modality increment rather than evidence that APT has no multimodal value. Because disease group and acquisition window coincide in this cohort, the analysis cannot identify which source of distribution shift accounts for the difference.

C.10 Earlier Donor-Disjoint RNA and ADT Results

The external analysis uses paired RNA and ADT measurements from nine donors. Its measured proteins, label ontology, and annotation procedure differ from those of AptamerBench. The analysis therefore evaluates transfer of the diagnostic framework rather than biological replication of the AptamerBench finding.

Adding ADT improves prediction over RNA alone under both bounded feature settings. Under the privileged multichild endpoint, the paired RNA and ADT condition also exceeds RNA with donor-by-sample-by-parent shuffled ADT at both 2,000 and 5,000 RNA features. Thus, the classification-based correspondence diagnostic can recover a positive signal in a second paired single-cell dataset.

Table 17: External RNA and ADT results under bounded optimization. Ordinary modality increments use unrestricted decoding over the fixed 45-label ontology and compare paired RNA and ADT with RNA alone. Privileged conditional contrasts compare paired RNA and ADT with RNA plus ADT shuffled within donor-by-sample-by-parent strata. Both conditional models use the reference parent to restrict decoding to its children, and Macro-F1 is computed over the 34 labels belonging to the eight multichild parents. All rows use full-development-cell MLPs with 2,000 or 5,000 training-selected RNA features. Each condition is independently selected on the ordinary 45-label validation endpoint and freshly refitted. Brackets denote paired 95% confidence intervals.

RNA budget	Contrast	First condition	Second condition	Difference
A. Ordinary 45-label endpoint, unrestricted decoding				
2,000	Paired−RNA	Paired: 0.596	RNA: 0.549	$+0.048$ $[0.031, 0.059]$
5,000	Paired−RNA	Paired: 0.584	RNA: 0.541	$+0.042$ $[0.029, 0.054]$
B. Conditional multichild endpoint, true-parent restricted decoding				
2,000	Paired−grouped shuffle	Paired RNA+ADT	RNA+shuffled ADT	$+0.0507$ $[0.0345, 0.0673]$
5,000	Paired−grouped shuffle	Paired RNA+ADT	RNA+shuffled ADT	$+0.0364$ $[0.0206, 0.0507]$

Table 18: Bounded external optimization sensitivity for ordinary donor-balanced Macro-F1 over the fixed 45-label ontology. Predictions use unrestricted 45-class argmax, with no true-parent decoding. The parent label is used only to construct donor-by-sample-by-parent ADT shuffle strata. Full-development-cell MLPs use 2,000 or 5,000 training-selected RNA features, and every condition receives independent bounded validation selection and fresh refitting. In contrast, the conditional rows of Table 17 restrict predictions to children of the reference parent and score only the 34 labels in multichild parents. The $+0.015$ and $+0.007$ ordinary contrasts here are therefore not repeated estimates of the $+0.0507$ and $+0.0364$ conditional contrasts in Table 17.

Ordinary donor-balanced Macro-F1 (45-label fixed ontology, no true-parent decoding)

RNA features	RNA	Paired RNA+ADT	RNA+grouped-shuffle ADT	Paired−RNA	Paired−donor×sample×parent shuffle
2,000	0.549	0.596	0.581	+0.048	+0.015
5,000	0.541	0.584	0.577	+0.042	+0.007

This result should not be interpreted as external validation of the AptamerBench biological finding. The external analysis does not use the H0 hierarchy, does not perform source-to-external transfer, and does not support quantitative comparison of APT and ADT effect sizes. Its released labels were informed by both RNA and ADT, and the cohort contains only nine donors.

Bounded optimization sensitivity. The conditional rows of Table 17 use a privileged multichild true-parent endpoint. Table 18 instead uses ordinary unrestricted decoding over all 45 labels, with the parent used only to construct the donor-by-sample-by-parent shuffle strata. It therefore asks whether original RNA and ADT pairing improves ordinary fixed-ontology prediction after bounded validation selection, rather than repeating the conditional endpoint above. Every condition receives independent validation selection and fresh refitting. The direct 5,000-minus-2,000 difference in the correspondence contrasts is -0.009 [-0.023, +0.003]. Therefore, the smaller 5,000-feature point estimate does not by itself establish a feature-budget effect.

D LABEL-INDEPENDENT SAME-CELL RETRIEVAL

Question and scope. The Q3 classification controls test whether preserving the original RNA and APT pairing improves prediction of the operational RNA-derived labels. We separately ask whether exact same-cell correspondence is detectable without using those labels. This exploratory diagnostic tests correspondence in held-out patients. It does not measure a cell-typing gain or identify a biological mechanism.

Patient-disjoint design. The experiment uses the frozen five-fold patient partition. For outer fold f, fold f is the test set, fold $(f + 1) \bmod 5$ is the validation set, and the remaining three folds form the inner fit set. After model selection, both encoders and all fitted preprocessing steps are reinitialized and fitted on the four development folds. Evaluation is then performed once on the held-out patients. All 40 patients receive out-of-fold predictions, and no patient is shared between development and test partitions.

Representations and objective. RNA uses the 10,000-HVG Q3 representation with library-size normalization to 10,000, log1p transformation, fit-only mean-binned dispersion selection, and fit-only standard-deviation scaling without centering. APT uses all 293 barcode-aligned raw-count features with log1p transformation and fit-only feature standardization. The RNA encoder has widths 512 and 128. The APT encoder has widths 256 and 128. Both use ReLU, dropout 0.1, and unit-normalized 128-dimensional outputs. Training minimizes a symmetric InfoNCE objective with temperature 0.07. Negatives are drawn only from the same patient, which prevents patient identity alone from solving the retrieval task.

Matched control and candidate construction. The real condition learns the observed pairing between RNA and APT from the same cell. The control condition reassigns complete APT vectors within each training patient using a fixed-point-free permutation before fitting. It does not shuffle individual APT features. Validation and test queries always target the true same-cell profile. Each query is evaluated against one true profile and distinct distractors from the same held-out patient. Candidate maps are shared across conditions, training seeds, shuffle realizations, and retrieval directions. Candidate-set sizes are 16, 32, 64, and 128.

Selection and uncertainty. The selection grid contains learning rates $\{3\times10^{-4}, 10^{-3}\}$ and weight decays $\{10^{-5}, 10^{-3}\}$. Selection uses the mean validation Recall@1 across both retrieval directions at $K = 64$. Final real-pair refits use training seeds 17, 29, and 43. Shuffled refits cross these training seeds with shuffle seeds 101, 211, and 307. Metrics are computed separately for each patient and then averaged with equal patient weight. Paired 95% intervals use 5,000 shared patient bootstrap draws. Shuffle realizations are averaged within a training seed before the training seeds are averaged. Seed-level values remain available in the machine-readable results and are not treated as independent patient observations.

Table 19: Label-independent same-cell retrieval with $K = 64$ same-patient candidates. Real-pair and shuffled columns report patient-balanced means. Differences are real minus shuffled with paired patient-bootstrap 95% confidence intervals. Chance is the analytical expectation under random ranking.

Direction	Metric	Real pair	Shuffled	Difference (95% CI)	Chance
RNA→APT	R@1	0.0761	0.0155	+0.0606 **[0.0517, 0.0701]**	0.0156
RNA→APT	R@5	0.2765	0.0778	+0.1987 **[0.1723, 0.2260]**	0.0781
RNA→APT	MRR	0.1909	0.0739	+0.1171 **[0.1018, 0.1328]**	0.0741
APT→RNA	R@1	0.0676	0.0155	+0.0521 **[0.0431, 0.0615]**	0.0156
APT→RNA	R@5	0.2489	0.0778	+0.1711 **[0.1446, 0.1981]**	0.0781
APT→RNA	MRR	0.1750	0.0739	+0.1011 **[0.0856, 0.1168]**	0.0741

Results. At $K = 64$, real-pair training improves all three retrieval metrics in both directions relative to matched within-patient shuffling (Table 19). For RNA to APT, the real and shuffled Recall@1 values are 0.0761 and 0.0155. The paired difference is $+0.0606$ ($[0.0517, 0.0701]$). For

APT to RNA, the corresponding values are 0.0676 and 0.0155. The paired difference is $+0.0521$ ($[0.0431, 0.0615]$). The Recall@5 and MRR contrasts are also positive. Shuffled performance remains close to analytical random-ranking expectations under these original random candidate sets. The same qualitative pattern is present for $K \in \{16, 32, 64, 128\}$ (Figure 7).

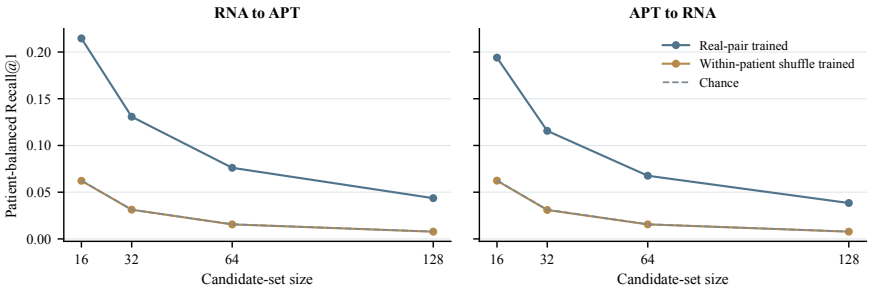


Figure 7: Candidate-set sensitivity for label-independent same-cell retrieval. Curves show patient-balanced Recall@1 for retrieval of APT from RNA and RNA from APT under the original random candidate construction. Models are trained using observed pairs or complete APT vectors reassigned within training patients. The dashed line is analytical random-ranking chance for these random candidate sets. Candidate maps are shared across conditions, training seeds, shuffle realizations, and retrieval directions.

Audit and interpretation. The primary retrieval audit covers 60 fitted retrieval models, 40 held-out patients, and 3,840 patient-level metric rows. Across the 45 shuffled refits, the training permutations contain no fixed points. The primary training and evaluation code does not access subtype, parent, disease, or cluster labels. The observed retrieval contrasts support detectable same-cell correspondence between RNA and APT under this task. They do not show that preserving correspondence improves cell-type classification, establish a molecular mechanism, or demonstrate external-cohort generalization. The analysis was designed after inspection of the Q3 classification results, so its intervals are conditional on the frozen folds, the selected dual-encoder recipe, and the finite training and shuffle seeds.

D.1 Post-hoc QC Sensitivity Analysis

Design. We conducted this post-hoc exploratory analysis after inspecting the original retrieval results. For each cell, we computed total counts and the number of detected features separately for RNA and APT from the raw count matrices. Total counts were log1p-transformed, and the two QC coordinates for each candidate modality were standardized using parameters fitted on the corresponding development patients. At $K = 64$, we retained the original held-out queries and selected 63 distinct distractors from the same patient by Euclidean proximity to the true partner in this two-dimensional QC space. RNA-to-APT candidate matching used APT total counts and detected aptamers. APT-to-RNA matching used RNA total counts and detected genes. The true partner was excluded from the distractor search. Candidate maps were fixed across real and shuffled models and training seeds. All 40 patients and 361,792 queries remained eligible.

This construction uses the true partner's QC measurements to choose distractors. It is therefore a privileged, truth-conditioned stress test, not a deployable candidate-selection procedure. The trained full-profile retrieval models were kept fixed. Only their test candidate sets changed. We additionally trained small-capacity QC-only retrieval models using the two specified QC features from each modality under the same patient-disjoint folds and real-pair versus within-patient-shuffled training conditions. QC preprocessing was fitted on development patients, and model selection used validation patients. Neither candidate construction nor QC-only training used subtype, parent, disease, or cluster labels.

Table 20: Post-hoc QC sensitivity analysis at $K = 64$. Values are patient-balanced Recall@1. Differences are real-pair minus shuffled-pair training, with 5,000-draw paired patient-bootstrap 95% intervals. The original candidate sets use random same-patient distractors. QC-matched sets use truth-centered, same-patient distractors. The analytical $1/64$ random-ranking reference applies to the original random candidate construction and should not be used as a null reference for the truth-conditioned QC-matched sets.

Direction and model	Candidates	Real / Shuffle	Difference [95% CI]
RNA→APT, full	Random	0.0761 / 0.0155	$+0.0606$ [0.0517, 0.0701]
APT→RNA, full	Random	0.0676 / 0.0155	$+0.0521$ [0.0431, 0.0615]
RNA→APT, full	QC-matched	0.0565 / 0.0154	$+0.0411$ [0.0348, 0.0481]
APT→RNA, full	QC-matched	0.0443 / 0.0157	$+0.0287$ [0.0241, 0.0334]
RNA→APT, QC-only	Random	0.021304 / 0.016498	$+0.004806$ [0.003442, 0.006231]
APT→RNA, QC-only	Random	0.020949 / 0.016081	$+0.004869$ [0.003205, 0.006554]
RNA→APT, QC-only	QC-matched	0.000899 / 0.001011	-0.000112 [-0.000293, 0.000052]
APT→RNA, QC-only	QC-matched	0.000688 / 0.000763	-0.000076 [-0.000229, 0.000086]

Full-profile results. The frozen full-profile models retain positive real-minus-shuffle contrasts under QC-matched candidate selection: $+0.0411$ ($[0.0348, 0.0481]$) for RNA to APT and $+0.0287$ ($[0.0241, 0.0334]$) for APT to RNA (Table 20). Both intervals exclude zero. The corresponding contrasts under the original random candidates are $+0.0606$ and $+0.0521$. QC matching also reduces the patient-balanced mean true-partner minus distractor distance in the standardized candidate-modality QC space from 0.903 to 0.043 for RNA-to-APT and from 0.731 to 0.035 for APT-to-RNA. Thus, the selected distractors are closer on the specified QC features, while a real-pair training advantage remains under this stress test.

QC-only ranking diagnostic. The QC-only models show small positive real-minus-shuffle contrasts on the original random candidates. On the truth-conditioned QC-matched sets, their Recall@1 values are approximately 0.0007 to 0.0010. We reloaded all 60 QC-only final checkpoints and audited the same queries and candidate maps. The mean strict rank of the true partner is approximately 32 among 64 candidates in both directions and training conditions. Exact score ties do not account for the low Recall@1: the strict-rank rule counts only candidates with scores strictly above the true partner, so ties favor the true partner.

The QC-matched sets select distractors around the true partner's QC measurements. The true partner therefore has a special position in candidate construction and is not exchangeable in rank with the distractors. Accordingly, $1/64$ is not an appropriate random-ranking null for the QC-only result on these sets. We interpret its low Recall@1 as behavior under truth-centered candidate selection, not as evidence that QC features contain no pairing information.

Interpretation and limits. The QC-matched full-profile results support detectable correspondence under a more difficult, specified retrieval diagnostic. They do not isolate a purely biological component: total counts and detected features can reflect both measurement conditions and genuine cell state. The difference between random-candidate and QC-matched performance cannot be interpreted as the fraction of retrieval signal caused by technical variation. QC-only and full-profile model scores likewise do not provide an additive decomposition of the signal. This exploratory analysis fixes one two-feature-per-modality distance, $K = 64$, and the completed models. Its intervals are conditional on those choices and the observed cohort.

E AUDIT AND REPRODUCIBILITY

E.1 FROZEN PARTITIONS, SEEDS, AND RESAMPLING

The primary benchmark uses the frozen five-fold patient assignment described in Appendix B.1. The joint disease and acquisition-group analysis uses separate leave-one-group-out partitions. The earlier donor-disjoint RNA/ADT analysis in Appendix C.10 uses donor-level partitions, while the Stephenson multicenter analysis in Appendix F uses three leave-one-center-out folds with all samples from the same subject kept in one partition. Results from these evaluation regimes are reported separately and are never pooled.

For the primary AptamerBench analyses, neural model selection uses seed 17, and selected configurations are refitted with seeds 17, 29, and 43. The primary matched APT shuffle uses realizations 101, 211, and 307. Paired contrasts use shared resampling draws and matched training-run slots whenever defined by the corresponding protocol. The primary benchmark uses patient-level resampling, while the Stephenson multicenter analysis uses subject-level resampling stratified by held-out center.

All preprocessing estimators, RNA feature selection, context construction, and shuffle mappings are partition specific. Whole-vector shuffles preserve the grouping variables specified by each protocol and do not cross fit, validation, refit, or test boundaries. Saved mapping audits verify patient or subject membership, parent membership, partition assignment, and target-identifier alignment for the reported runs.

E.2 MODEL SELECTION AND OUTPUT PROVENANCE

Each experimental condition performs its own validation selection and fresh refit. Reported nonlinear metrics are computed separately for each fitted run before aggregation. Model probabilities are therefore not averaged across training seeds or shuffle realizations before nonlinear metrics are computed.

Tables in this appendix are generated from saved aggregate results and, where available, out-of-fold predictions or compact run-level confusion matrices.

The H_0, H_1, and L_1 analyses use separately trained models together with their respective parent definitions, shuffle strata, and candidate sets. The earlier donor-disjoint and Stephenson multicenter RNA/ADT analyses are separate experimental pipelines.

E.3 SOFTWARE, COMPUTE, AND HYPERPARAMETER GRIDS

The implementation uses Python with scikit-learn for logistic regression, XGBoost for boosted trees, and PyTorch for neural models. Data processing and annotation use the documented single-cell workflow and cited packages. Classical reference models run on CPU, while neural reference models run on NVIDIA RTX 6000 Ada GPUs. Table 21 records realized search and refit effort for the reported Q1 reference-model suite. These wall-time records are provided for reproducibility and are not intended as a cross-hardware efficiency comparison.

Table 21: Realized search/refit records summed across folds (and across tasks for LR/XGBoost). Neural search uses four candidates per fold. LR uses four and XGBoost twelve per task/fold. Epoch ranges are selected stopping epochs, not the epochs actually consumed by all search trials. Times are accumulated timed-block wall hours, not elapsed project time or GPU-hours: parallel jobs overlap, hosts differ, and input/scaler preparation is outside these blocks. Neural runs used RTX 6000 Ada GPUs. Classical runs used CPU. This is a reproducibility record, not a cross-hardware efficiency ranking. HCE correction inference is excluded.

Recipe	Candidate fits	Refit epochs	Search h	Refit h
Logistic Regression	40	n/a	0.53	0.20
XGBoost	120	n/a	4.32	2.31
Plain MLP	20	15 to 26	0.08	0.04
Flat Transformer	20	14 to 45	4.73	3.79
HCE Transformer	20	14 to 37	4.10	3.34
Soft-cascade (no disease)	20	11 to 39	5.07	3.84
Flat + contrastive	20	11 to 47	5.85	6.19
Cascade + contrastive	20	19 to 46	6.23	7.03

Hyperparameter grids for Q1 are listed in Appendix B.2. Q2 uses the MLP family while independently selecting and refitting the real and shuffled conditions. Q3 independently selects each input condition. The bounded optimization follow-up in the earlier donor-disjoint RNA/ADT analysis is exploratory because its search space was specified after inspection of the fixed-configuration results. The Stephenson multicenter analysis is also an exploratory follow-up because it was designed after inspection of the primary AptamerBench and earlier external results. Its protocol was frozen before any leave-one-center-out test score was computed.

E.4 Verification Scope and Known Limitations

The audit checks partition disjointness, fit-only transformations, whole-vector shuffle boundaries, label and target alignment, probability outputs, and finite model outputs for the completed analyses. These checks establish implementation consistency with the stated protocols. They do not establish that the operational reference labels are error-free, that every plausible model or preprocessing choice has been exhausted, or that the reported conditional contrasts identify a molecular mechanism.

The primary cohort contains 40 participants but only six aligned disease and acquisition groups. Joint-group holdout therefore evaluates transfer beyond the original patient folds without separating disease-associated from acquisition-associated shift.

The external RNA/ADT analyses use different assays, cohorts, label ontologies, and annotation procedures from AptamerBench and should be interpreted as tests of framework transfer rather than biological replication of the APT findings. In the Stephenson cohort, the released labels were informed by both RNA and surface-protein measurements. Clinical composition also differs across only three collection centers, so the leave-one-center-out analysis does not separately identify center and disease effects or establish generalization to arbitrary future centers.

True-parent analyses remain privileged diagnostics, and patient-context analyses remain transductive. Label-independent retrieval establishes task-specific detectable correspondence but does not identify whether that correspondence arises from biological state, technical variation, or another shared cell-level signal. Confidence intervals condition on the observed datasets, reference labels, partitions, completed selection procedures, tested model families, and finite training and shuffle repetitions.

F External Multicenter Protocol Transfer in the Stephenson RNA/ADT Cohort

F.1 Cohort, Released Labels, and Operational Ontology

We evaluated whether the AptamerBench estimand decomposition transfers to the independently collected Stephenson et al. paired RNA/ADT cohort (Stephenson et al., 2021). The analysis used the released processed object. The original object contains 647,366 cells from 130 subjects and 143 samples collected at three centers. After excluding the IV-LPS condition, the primary analysis contains 639,482 cells from 118 subjects and 131 samples. Thirteen subjects contribute multiple samples. All samples from the same subject remain in the same partition.

Each retained cell has 24,737 RNA features and 192 ADT features, and every ADT feature has nonzero observations at each center. We used the unique cell identifiers and paired rows in the released H5AD object. We did not infer pairing from row positions in separately processed matrices or silently intersect modality-specific cell sets.

The prediction target is the released 51-class `full_clustering` annotation. The accompanying documentation describes these labels as based primarily on RNA markers with support from surface-protein measurements. The classification experiments therefore test transfer of the diagnostic protocol rather than validation against an ADT-independent biological ground truth.

The released `initial_clustering` variable does not define a deterministic parent hierarchy: 40 of the 51 fine labels occur under more than one `initial_clustering` value. We therefore froze an operational five-parent mapping comprising B, T/NK, Myeloid/DC, HSPC, and Other before any LOCO test score was computed. The mapping covers all 51 fine labels and assigns each to exactly one parent. It defines conditional decoding, shuffle strata, and support-aware scoring. It is not a new cell-level biological annotation.

Clinical composition differs across centers. Sanger contains only COVID-19 samples, whereas non-COVID respiratory controls occur only at Ncl. LOCO thus measures transfer across the observed combinations of center, acquisition, and clinical composition. It does not separately identify center and disease effects.

F.2 Leave-One-Center-Out Evaluation Protocol

The external evaluation uses three LOCO folds. In each fold, every subject and sample from one center is reserved for testing. The held-out center contributes no observations to normalization, feature selection, ADT scaling, validation, early stopping, hyperparameter selection, or model refitting. Within the two development centers, 20% of subjects per center are assigned to validation. After selection, preprocessing and model parameters are reinitialized and fitted on all development subjects. The primary analysis uses all eligible cells without a training-cell cap.

RNA counts are normalized to a library size of 10,000 per cell and log transformed. The 2,000 highest-variance RNA features are selected only on the applicable fit or refit partition. Fit-only means and standard deviations then scale these features, which are clipped to $[-10, 10]$. ADT-containing conditions select between cell-local $\log(1+x)$ and CLR, $\log(1+x) - \mathrm{mean}_{\mathrm{features}} \log(1+x)$, followed by fit-only feature standardization.

Each classification condition independently selects its ADT representation, MLP configuration, and stopping epoch. Models have two ReLU hidden layers, dropout 0.1, cross-entropy loss, AdamW optimization, batch size 1,024, at most 100 epochs, and patience 10. The frozen configurations are widths $(256, 128)$ with learning rate/weight decay $(10^{-3}, 10^{-4})$, $(3{\times}10^{-4}, 10^{-4})$, or $(10^{-3}, 10^{-3})$, and widths $(512, 256)$ with $(10^{-3}, 10^{-4})$. Final refits use training seeds 17, 29, and 43. Shuffled conditions cross these seeds with shuffle seeds 73001, 73002, and 73003. Selection uses shuffle seed 91001.

All shuffles reassign complete ADT vectors while retaining the target cell identifier, RNA profile, and reference label. Shuffles are generated separately within fit, validation, refit, and test. Subject shuffling permutes within subject. Subject-by-parent shuffling additionally restricts source cells to the same operational parent. Fixed-point-free derangements are used whenever a stratum contains at

Table 22: Composition of the Stephenson cohort after exclusion of the IV-LPS condition. Subjects with multiple samples remain within one partition.

Center	Subjects	Samples	Cells
Cambridge	47	60	130,644
Ncl	60	60	433,925
Sanger	11	11	74,913
Total	118	131	639,482

Table 23: Subject-level partitions for Stephenson LOCO evaluation. Entries report subjects/cells. Model selection uses fit and validation. Final models are freshly trained on the complete refit partition.

Held-out center	Fit	Validation	Refit	Test
Cambridge	57 / 413,673	14 / 95,165	71 / 508,838	47 / 130,644
Ncl	47 / 179,223	11 / 26,334	58 / 205,557	60 / 433,925
Sanger	86 / 456,459	21 / 108,110	107 / 564,569	11 / 74,913

least two cells. Singleton strata remain unchanged and are recorded. Because the subject-by-parent test shuffle uses reference parent to define its strata, it is a privileged diagnostic control.

The protocol was frozen before any Stephenson LOCO test score was computed. Because the AptamerBench and earlier external RNA/ADT results had already been inspected, this experiment is an exploratory follow-up rather than a preregistered confirmation.

F.3 Conditional ADT Association

The conditional analysis compares real target-cell ADT with an independently selected and refitted subject-by-parent shuffled condition. Both conditions receive the same true operational parent only during decoding, restricting each terminal prediction to the children of that parent. The contrast is therefore a privileged diagnostic of terminal-label-associated ADT information under the specified hierarchy.

The pooled real-minus-shuffle difference is $+0.257$ (95% CI $[0.244, 0.269]$). The corresponding support-aware within-subject-parent difference is $+0.284$ ($[0.267, 0.300]$). Center-specific pooled contrasts are positive in all three held-out centers. These results establish terminal-subtype-associated predictive information under the specified ontology, model family, shuffle, and privileged decoding. They do not estimate a causal contribution of ADT or ordinary deployment performance.

Table 24: Conditional ADT association under LOCO evaluation. Real and shuffled conditions use matched true-parent decoding. Differences are real minus subject-by-parent shuffled ADT, with paired 95% confidence intervals.

Center	Subjects	Real	Shuffle	Δ [95% CI]
Cambridge	47	0.377	0.075	$+0.302\ [0.285, 0.317]$
Ncl	60	0.291	0.082	$+0.209\ [0.191, 0.233]$
Sanger	11	0.372	0.084	$+0.287\ [0.259, 0.303]$
Pooled LOCO	118	0.340	0.083	$+0.257\ [0.244, 0.269]$

F.4 Modality and Classification Correspondence

We independently selected and refitted RNA alone, paired RNA+ADT, RNA with subject-shuffled ADT, and RNA with subject-by-parent shuffled ADT. Paired minus RNA measures the modality increment under the specified classifier. The two paired-minus-shuffle contrasts ask whether preserving target-cell correspondence improves classification relative to each grouped control.

Table 25: Stephenson multimodal classification contrasts. Pooled columns report subject-balanced Macro-F1 differences. Within-subject columns compute support-aware scores within each subject before averaging. Intervals containing zero do not establish equivalence.

	Pooled Δ	Within-subject Δ
Contrast	CI	CI
Paired − RNA	-0.001 $[-0.008, 0.006]$	$+0.002$ $[-0.004, 0.008]$
Paired − subject shuffle	$+0.004$ $[-0.007, 0.013]$	$+0.010$ $[0.003, 0.017]$
Paired − subject-by-parent shuffle	$+0.005$ $[-0.001, 0.011]$	$+0.004$ $[-0.001, 0.008]$

The pooled scores are 0.510 for RNA alone, 0.508 for paired RNA+ADT, 0.505 for subject-shuffled ADT, and 0.503 for subject-by-parent shuffled ADT. All three pooled contrast intervals include zero. The pooled LOCO analysis therefore does not establish a modality increment or an additional classification benefit from preserving same-cell pairing.

The within-subject endpoint gives a different answer for one control: paired RNA+ADT exceeds subject-shuffled ADT by $+0.010$ (95% CI $[0.003, 0.017]$), whereas paired-minus-RNA and paired-minus-subject-by-parent intervals include zero. Pooled and within-subject Macro-F1 are distinct nonlinear endpoints. Their differences are not an additive information decomposition.

Table 26: Center-specific pooled SB Macro-F1 contrasts for the Stephenson multimodal classification analysis. Entries are paired differences with 95% confidence intervals. All nine center-specific intervals include zero.

Center	Paired − RNA	Paired − subject shuffle	Paired − subject-by-parent shuffle
Cambridge	$+0.0035$ $[-0.0216, 0.0256]$	$+0.0102$ $[-0.0063, 0.0272]$	$+0.0034$ $[-0.0167, 0.0186]$
Ncl	-0.0088 $[-0.0185, 0.0006]$	-0.0003 $[-0.0168, 0.0135]$	-0.0001 $[-0.0065, 0.0058]$
Sanger	-0.0027 $[-0.0190, 0.0195]$	-0.0035 $[-0.0185, 0.0183]$	-0.0033 $[-0.0199, 0.0156]$

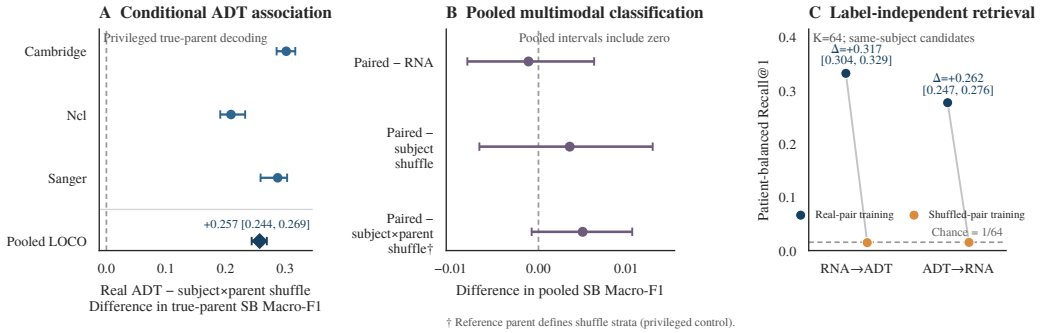


Figure 8: External multicenter diagnostics in the Stephenson paired RNA/ADT cohort. **A.** Center-specific and pooled real-minus-subject-by-parent shuffle contrasts under matched true-parent decoding. **B.** Pooled subject-balanced classification contrasts for paired RNA+ADT relative to RNA and two grouped-shuffle controls. All three pooled intervals include zero. **C.** Patient-balanced Recall@1 for bidirectional label-independent retrieval with $K = 64$ same-subject candidates. Models are trained with observed pairs or within-subject shuffled pairs. The dashed line denotes analytical chance. Error bars in A and B denote paired 95% confidence intervals.

F.5 Label-Independent Same-Cell Retrieval

To test correspondence without released cell-type labels, we trained separate RNA and ADT encoders whose 128-dimensional outputs are normalized and optimized with a symmetric contrastive objective. The RNA encoder uses one 512-unit hidden layer and the ADT encoder one 256-unit hidden layer, each with ReLU and dropout 0.1. Selection considers four learning-rate/weight-decay configurations, uses $K = 64$ validation Recall@1, and trains for at most 50 epochs with patience 8. Real-pair models use observed same-cell pairs. Controls use fixed-point-free ADT reassignment within each training subject. Test candidates come only from the same held-out subject and contain exactly one true partner. Cell-type, parent, disease, and center labels are neither inputs nor prediction targets.

At $K = 64$, RNA-to-ADT Recall@1 is 0.332 after real-pair training and 0.015 after shuffled-pair training, a difference of $+0.317$ (95% CI $[0.304, 0.329]$). ADT-to-RNA Recall@1 is 0.277 versus 0.015, a difference of $+0.262$ ($[0.247, 0.276]$). Shuffled performance is close to the analytical chance level $1/64$. Both directional contrasts are positive at every held-out center.

Table 27: Center-specific label-independent retrieval contrasts at $K = 64$. Entries are real-pair minus shuffled-pair patient-balanced Recall@1 with paired 95% confidence intervals computed within each held-out center.

Center	RNA→ADT	ADT→RNA
Cambridge	$+0.385$ $[0.363, 0.403]$	$+0.345$ $[0.324, 0.363]$
Ncl	$+0.257$ $[0.237, 0.276]$	$+0.186$ $[0.164, 0.209]$
Sanger	$+0.357$ $[0.330, 0.381]$	$+0.321$ $[0.295, 0.342]$

Candidate-set sensitivity for $K \in \{16, 32, 64, 128\}$ is shown in Figure 9. The real-pair advantage is positive in both directions across all tested candidate-set sizes.

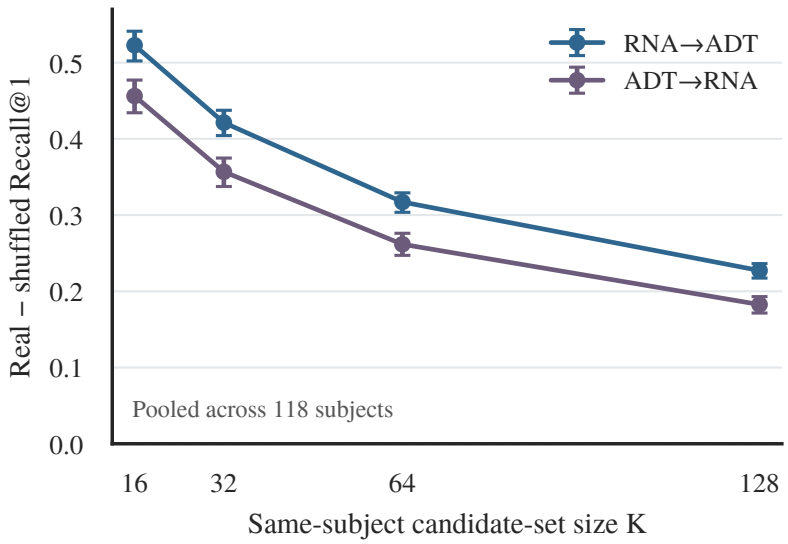


Figure 9: Stephenson label-independent retrieval sensitivity. Curves show pooled real-minus-shuffled patient-balanced Recall@1 with paired 95% confidence intervals. Candidate sets contain exactly one same-cell partner and otherwise contain cells from the same held-out subject.

F.6 Statistical Inference, Audit, and Limitations

Confidence intervals use 5,000 percentile bootstrap replicates with subjects as the resampling unit. Pooled LOCO resampling is stratified by held-out center. Compared conditions share subject draws, training-seed slots, and applicable shuffle draws. Each fitted run is scored before aggregation. Probabilities are not averaged before nonlinear metrics are computed.

The final audit verifies code-freeze integrity, subject-disjoint partitions, exclusion of held-out centers from fitted preprocessing, alignment of target identifiers and predictions, finite normalized probabilities, shuffle boundaries, and all 108 expected classification runs. Retrieval checks verify same-subject candidate sets, exactly one positive partner, and no access to classification labels.

Several limitations bound the experiment. This is an exploratory follow-up, and released labels were partly informed by surface protein. The conditional result is therefore not validation against ADT-independent truth. The five-parent hierarchy is operational. Center and clinical composition are associated across only three centers, so the subject bootstrap describes the observed cohort rather than arbitrary future centers. Retrieval establishes task-specific correspondence but does not separate biological state, technical variation, or other shared cell-level signals.

The external RNA/ADT outcomes are complementary. GSE194122 has a positive conditional paired-versus-shuffled classification contrast. Stephenson has strong label-independent correspondence but unresolved pooled modality and classification-correspondence contrasts. Because assays, ontologies, cohorts, partitions, and endpoints differ, they do not estimate one biological effect. They show why correspondence and target-specific predictive value require separate evaluation.