

Assay resolution governs the transfer of molecular dependence

Anonymous Authors^{1*}

^{1*}Affiliations withheld in the blinded manuscript.

Abstract

Background: Separate molecular assays retain coexpression even when cross-assay cell pairing is unavailable. We ask which recipient measurements are needed to transfer RNA–protein dependence from a paired reference.

Results: For finite assay profiles and a fixed interaction, we derive the sharp range of binary associations compatible with specified unpaired summaries and its minimax predictor. Retrospective reconstruction in 24 Cambridge donors, 56 Newcastle donors, and 15 patients with pediatric T-cell acute lymphoblastic leukemia (T-ALL) reduced deviance by 16.6%, 29.2%, and 55.6% when both assay-pattern distributions replaced individual frequencies (paired recipient-bootstrap 95% intervals, 13.0–20.3%, 25.0–33.3%, and 47.2–62.7%). Recipient preprocessing and marginal estimates used 256 adaptation cells; 256 untouched cells supplied the endpoint. In an independent test specified publicly before count access, the unchanged blood-reference model reduced deviance by 15.7% in 11 colon-biopsy donors (11.4–21.1%), improving every donor. Both patterns also improved on protein patterns alone and independence after adjustment for the three prespecified comparisons. In a separate fixed-frequency comparison, covariances retained 91–97% of the pattern-level gain and recipient-specific structure outperformed larger matched pools. A covariance response formula improved query prioritization over direct interaction strength in Newcastle and leukemia.

Conclusions: Assay summaries determine both the identifiable range of molecular dependence and the information available for its prediction. Preserving recipient coexpression improves held-cell prediction without supplying scoring-cell frequencies or cross-assay pairings.

Keywords: single-cell multi-omics; molecular dependence; coarsening; statistical matching; optimal transport

1 Background

Joint single-cell assays reveal how molecular measurements covary within cells. When a new cohort supplies the assays separately, a paired reference can support prediction of the missing joint distribution. The recipient still provides two kinds of information: the frequency of each molecular state and the patterns of states observed together within each assay. These are different constraints on the cross-assay prediction.

Association models separate an interaction from its marginal distributions and estimate it through conditional or density-ratio methods [1, 2]. Inverse optimal transport learns an interaction from paired observations [3], and CHAMPOLLION applies this approach to multi-omic bridge integration [4]. We ask how the resolution of recipient observations changes the relationships that such a model can recover. A coarse RNA–protein association can change when the frequencies of finer cellular patterns change, even if

the fine interaction is fixed. This connects cross-cohort prediction to the noncollapsibility of marginal association [5, 6]. Existing methods reconstruct a joint table from specified marginals and interaction parameters [7]. Here the complete marginals can themselves be unknown: observed within-assay summaries constrain them, leaving a set of possible coarse associations.

For finite assay supports and a specified interaction, we derive the sharp range of binary associations compatible with unpaired summaries that determine the query frequencies. The range gives a minimax prediction and quantifies ambiguity remaining after those summaries are known exactly. It includes the case in which every individual marker frequency is supplied; retaining only the two queried frequencies gives a closed form. We then test whether preserving within-assay structure improves prediction on untouched recipient cells. Each comparison holds the source interaction, cells, and estimated recipient marker frequencies fixed, separating the value of retained information from differences between fitted estimators. To identify which structure matters, we preserve recipient covariances with pairwise maximum-entropy models [8], then compare them with models built from source donors or other recipients in the same cohort. A response expansion explains how recipient coexpression changes a transferred association and gives a query-ranking rule computable from unpaired assays.

2 Methods

2.1 Reconstruction at the observed assay resolution

Let x and y denote RNA and protein profiles on finite supports. A finite transferred interaction $S(x, y)$ defines a family of recipient distributions

$$q(x, y) = \exp\{S(x, y) + a(x) + b(y)\}. \quad (1)$$

The adjusting functions a and b enforce the observed recipient assay distributions, including normalization. When both assay distributions are known and strictly positive, they determine a unique reconstruction [1]. Molecular queries are obtained by summing its probabilities over the corresponding profile states.

When only individual marker frequencies are retained, we instead fit the exponential-family distribution whose adjusting functions are linear in the marker indicators. The same interaction therefore supports several reconstructions, depending on the recipient constraints. The means-only model constrains marker frequencies but leaves within-assay dependence unconstrained.

2.2 A sharp limit at the observed resolution

Let the observed RNA and protein summaries be arbitrary finite collections of within-assay means. Assume they define nonempty convex polytopes $\mathcal{R}$ and $\mathcal{C}$ of complete assay distributions consistent with those

summaries. For each $r \in \mathcal{R}$ and $c \in \mathcal{C}$, entropic reconstruction gives the unique joint distribution $Q_S(r, c)$ with interaction S and assay marginals r, c , allowing zero-mass assay states (Additional file 1). For a binary query $U = f(x)$ and $V = g(y)$, define

$$\ell = \min_{r \in \mathcal{R}, c \in \mathcal{C}} \Pr_{Q_S(r, c)}(U = 1, V = 1),$$

$$h = \max_{r \in \mathcal{R}, c \in \mathcal{C}} \Pr_{Q_S(r, c)}(U = 1, V = 1).$$

If the summaries determine the query frequencies $m_U = \Pr(U = 1)$ and $m_V = \Pr(V = 1)$, the set of query tables from these reconstructions is exactly

$$\left\{ \begin{pmatrix} 1 - m_U - m_V + t & m_V - t \\ m_U - t & t \end{pmatrix} : \ell \leq t \leq h \right\}. \quad (2)$$

The result follows because $\mathcal{R} \times \mathcal{C}$ is compact and connected and the matrix-scaled reconstruction is continuous. It is the sharp identified set when all individual marker frequencies are supplied: take the defining means of $\mathcal{R}$ and $\mathcal{C}$ to be the frequencies of all markers. Complete assay-profile histograms collapse both polytopes to points. These are the constraints used by the means-only and full-pattern reconstructions. Additional file 1 gives the formal result and proof.

The means-only exponential-family fit selects one member of this set by information projection. The interval records what the supplied frequencies identify without that additional selection.

Write $P(t)$ for the table in Eq. (2). The unique minimax table P_* equalizes the forward Kullback–Leibler (KL) loss from the endpoint tables $P(\ell)$ and $P(h)$: $\text{KL}(P(\ell) \| P_*) = \text{KL}(P(h) \| P_*)$. Within this fixed-interaction family with exact population summaries, no predictor has smaller worst-case forward KL loss. Computing ℓ and h requires global optimization over the marginal polytopes; these outer problems need not be convex. The interval is an analytic characterization, not a general efficient endpoint algorithm. Writing $\mathcal{H}(P) = -\sum P \log P$, its log odds are

$$\theta_* = \frac{\mathcal{H}(P(\ell)) - \mathcal{H}(P(h))}{h - \ell} \quad (\ell < h). \quad (3)$$

When the recipient retains only the two queried frequencies, the endpoints have a closed form. For a fine rectangle with $f(x_0) = g(y_0) = 0$ and $f(x_1) = g(y_1) = 1$, define its interaction contrast

$$d = S(x_0, y_0) + S(x_1, y_1) - S(x_0, y_1) - S(x_1, y_0). \quad (4)$$

Let L and H be the smallest and largest such contrasts. The compatible coarse log odds fill $[L, H]$ in closure, so ℓ and h are the cell probabilities with margins m_U, m_V and log odds L, H . The information-radius argument uses classical minimax KL geometry [9]; our resolution-specific result is the exact interval induced by the observed within-assay constraints.

To see the interval directly, sum the fine-table identity $q(x_0, y_0)q(x_1, y_1) = e^d q(x_0, y_1)q(x_1, y_0)$ over all rectangles. The coarse odds ratio is a positively weighted average of e^d , placing its log in $[L, H]$. Concentrating each assay margin on an extremizing pair of profiles attains either endpoint in closure. Continuity over the connected set of compatible margins fills the interval between them.

2.3 Individual frequencies can all agree

The ambiguity also occurs when every marker’s individual frequency is known. Consider two binary RNA markers and one protein marker, coded as $Z_1, Z_2, W \in \{-1, 1\}$, each with mean zero. Fix the fine interaction to $S = JZ_2W$, with $J \neq 0$. Let the recipient RNA distribution have correlation $\rho = \mathbb{E}(Z_1Z_2)$. Reconstruction gives

$$\mathbb{E}(Z_1W) = \rho \tanh J. \quad (5)$$

Changing ρ from negative to positive reverses the queried RNA–protein association without changing J or any individual marker frequency. Separate RNA profiles reveal ρ ; their individual frequencies do not. As ρ ranges over its compatible values, the all-marker interval for $\Pr(Z_1 = 1, W = 1)$ is $[(1 - |\tanh J|)/4, (1 + |\tanh J|)/4]$. The example therefore gives a nondegenerate instance of Eq. (2), not a separate ambiguity claim.

2.4 How coexpression changes the prediction

For fixed positive assay distributions on finite supports, scale a bilinear interaction as $S_t(x, y) = tx^\top By$, where B is the source interaction matrix and t sets its strength. Let Σ_x and Σ_y be the two within-assay covariance matrices. Near independence, the reconstructed cross-assay covariance satisfies

$$\text{Cov}_{q_t}(X, Y) = t\Sigma_x B \Sigma_y + O(t^2). \quad (6)$$

To first order, each predicted RNA–protein covariance sums source interactions weighted by recipient covariances in both assays. A source coefficient therefore need not equal the marginal association observed in a new population. This is the multivariate counterpart of Eq. (5).

This local formula is the initial value of an exact finite-strength identity. Let $s_t = \partial_t \log q_t(X, Y)$ be the score along the fixed-marginal reconstruction, and write $\bar{X} = X - \mathbb{E}_r X$ and $\bar{Y} = Y - \mathbb{E}_c Y$. The marginal

constraints give $\mathbb{E}_{q_t}(s_t | X) = \mathbb{E}_{q_t}(s_t | Y) = 0$, and hence

$$\text{Cov}_{q_t}(X, Y) = \int_0^1 \mathbb{E}_{q_t}\{\bar{X}\bar{Y}^\top s_t\} dt. \quad (7)$$

At independence, $s_0 = \bar{X}^\top B \bar{Y}$, which recovers Eq. (6).

The second-order term contracts two copies of B with the third central moments of each assay (Additional file 1). Thus preserving means and covariances preserves the first-order prediction but does not determine the prediction at the fitted interaction strength. The expansion specializes the dependence-model score-projection identities of Sei and Yano [1]; it motivates a controlled test of covariance against complete pattern information.

The response also predicts which queries are sensitive to replacing recipient coexpression by pooled structure. For a binary query with nonconstant marker frequencies, subtract the queried entries of $\Sigma_x B \Sigma_y$ for the recipient and pool, square the difference, and divide by the product of the two marker variances. This is the leading coefficient of twice the KL divergence between their query distributions near independence (Additional file 1). We test whether this score identifies relationships whose empirical prediction improves with recipient-specific reconstruction.

In the fixed-frequency analysis, we evaluated the response rule along $t \in \{0.05, 0.10, 0.25, 0.50, 0.75, 1\}$. At each value, we exactly reconstructed the recipient covariance model and the leave-one-recipient-out cohort-pool model, then compared the response coefficient with twice the forward KL divergence between their 81 query tables. We report pooled and within-recipient Spearman correlations, top-27 overlap, and the ratio of exact discrepancy to its quadratic approximation. Recipient-bootstrap intervals used 20,000 draws with seed 906095.

2.5 Data and recipient information

The source comprised 12 Cambridge calibration donors from the Stephenson CITE-seq (cellular indexing of transcriptomes and epitopes by sequencing) study [10]. Recipients comprised the remaining 24 Cambridge donors, 56 Newcastle donors, and all 15 diagnosis patients in the public pediatric T-ALL study, GSE248287 [11], which sorted and remixed immune and malignant cells before sequencing. The fixed panel was CD4, CD7, CD14, CD19, CD33, CD38, CD44, CD47, and CD52 in both assays. Each person contributed 512 cells selected deterministically without consulting their molecular counts. The first 256 cells supplied adaptation profiles; the remaining 256 supplied evaluation tables. Cross-assay correspondence in the adaptation half was discarded, so reconstruction received two separate assay distributions.

2.6 Prediction from adaptation cells alone

RNA states recorded detection. For each person and protein, the median count among the 256 adaptation cells defined a threshold. Counts strictly above this threshold were high; ties were low. The same threshold was applied to scoring cells without using their values to define it.

We fitted the source interaction once on the 12 Cambridge source donors under these state definitions, using pairing contrasts with penalty 10^{-4} . The contrasts compare observed cell pairs with within-donor swaps and remove additive assay effects from the likelihood [1]. A product-Bernoulli source prior with total pseudocount mass four smoothed each recipient adaptation histogram. The resulting histograms determined every recipient marginal constraint; no scoring-cell frequencies or pairings entered prediction.

The four reconstructions retained individual frequencies, RNA patterns plus protein frequencies, protein patterns plus RNA frequencies, or both pattern distributions. All used the same interaction and the same estimated marker frequencies. An independence control used the product of those frequencies. Predictions were fixed before scoring the remaining 256 cells. Full patterns were compared with individual frequencies, protein patterns, and independence in each cohort, giving nine contrasts. Additional file 1 specifies the source fit, smoothing, and numerical checks.

2.7 Fixed-frequency decomposition of the information gain

To isolate dependence prediction from errors in estimated marker frequencies, we also analyzed the same recipients conditional on their scoring-cell frequencies. This analysis used the original donor-wide protein states and source fit described below. Its losses are reported separately because its states and permitted information differ from adaptation-only prediction.

RNA states recorded detection. Protein states divided each person’s 512 cells into lower and upper rank halves, with deterministic hash-based tie breaking. The first 256 cells supplied separate RNA and protein pattern histograms. A source prior with total pseudocount mass four smoothed each histogram, which was tilted to the individual marker frequencies of the remaining 256 cells. Cross-assay pairings in those remaining cells were used only for scoring.

We estimated B in the binary interaction $S(x, y) = x^\top B y$ by pairing contrasts, with penalty 10^{-4} selected on source donors. These contrasts compare observed pairs with their within-donor swaps, removing additive assay effects from the likelihood [1]. The same fitted B was used in every arm without refitting or recipient tuning.

The four arms retained individual frequencies alone, RNA patterns plus protein frequencies, protein patterns plus RNA frequencies, or both pattern distributions. Each arm used the exponential-family reconstruction satisfying its specified constraints. Complete pattern constraints were fitted by matrix scaling; individual frequency constraints were fitted by moment matching.

For the covariance intervention, we replaced each smoothed, tilted assay distribution by the maximum-entropy distribution matching its first and second marker moments. On a binary panel this distribution has unary and pairwise log-linear terms [8]. Reconstruction used these replacement marginals and the unchanged B . The intervention preserves measured coexpression but replaces pattern structure not determined by those moments.

To test recipient specificity, we constructed two pooled controls from the first 256 cells of either the 12 source donors or all other recipients in the query person’s cohort. The latter pool excluded the query person and contained 23 Cambridge donors, 55 Newcastle donors, or 14 T-ALL patients. Each contributor added prior mass four, preserving the prior fraction. Pooled histograms underwent the same tilting to recipient marker frequencies and maximum-entropy replacement. The cohort control used other recipients’ unpaired profiles; neither control used their scoring-cell patterns.

For query prioritization, the response score used the recipient and cohort-pool covariances. The comparator used the squared direct source coefficient B_{ij}^2 multiplied by the two marker variances. Each rule ranked all 81 pairs within each person, with zero assigned to constant-marker queries and ties resolved by the fixed marker order. The top-third comparison and scores were fixed before calculating pair-specific losses. We compared the mean deviance gain among the response score’s top 27 and bottom 27 pairs, then compared its top 27 with the comparator’s top 27. Both rules used the same recipient predictions; only the ranking changed.

2.8 Endpoint and uncertainty

The outcome was twice the KL divergence from each empirical scoring table to its prediction, averaged over all 81 RNA–protein pairs and then equally over people. Constant-marker tables were retained. Intervals used 20,000 paired recipient-bootstrap draws. The adaptation-only analysis adjusted across its nine contrasts. For the fixed-frequency analyses, Bonferroni percentile intervals were calculated within three comparison families: nine contrasts between full patterns and the three less detailed representations; six contrasts comparing covariance-preserving reconstruction with full patterns or individual frequencies; and six contrasts comparing recipient-specific covariance with the two pools. The six query-ranking contrasts formed a fourth family. Intervals reported as 95% intervals are unadjusted; multiplicity-adjusted conclusions are stated separately. These analyses were specified after earlier outcomes had been examined and are retrospective.

2.9 Independent transfer from blood to colon

We applied the unchanged blood-reference interaction to colon-biopsy CITE-seq from healthy controls and patients with ulcerative colitis [12]. Eleven physical donors supplied at least 512 paired-assay cells: four controls, four patients receiving aminosalicylates, and three receiving vedolizumab. Biopsies from the

same person were pooled; a twelfth donor had 433 eligible cells and did not meet the fixed cell budget. All deposited cell types were retained. The nine-marker panel, cell selection, state definitions, source fit, and prediction protocol were fixed publicly before molecular count access [13].

The first 256 selected cells supplied separate assay distributions and the remaining 256 supplied scoring tables. No source refitting or recipient tuning was performed. The three prespecified comparisons tested full patterns against individual frequencies, protein patterns, and independence. Success required positive Bonferroni percentile lower bounds for all three comparisons and at least 10% mean deviance reduction against individual frequencies. These intervals used 20,000 paired donor-bootstrap draws and were calculated separately from the retrospective cohorts.

3 Results

3.1 Assay patterns improve prediction on untouched cells

Both assay-pattern distributions improved adaptation-only prediction in all three cohorts (Fig. 1). Mean deviance fell from 0.02229 to 0.01858 in Cambridge, from 0.03233 to 0.02289 in Newcastle, and from 0.05847 to 0.02599 in T-ALL. These are reductions of 16.63% (95% interval, 13.05–20.34%), 29.19% (25.00–33.28%), and 55.56% (47.25–62.67%). Full patterns improved 24/24, 55/56, and 15/15 people, respectively.

Retaining RNA patterns in addition to protein patterns reduced loss by 5.36% (95% interval, 3.33–7.49%), 12.27% (10.19–14.31%), and 31.76% (25.91–37.86%). Reductions relative to independence were 46.14%, 47.41%, and 65.16%. All nine multiplicity-adjusted intervals had positive lower endpoints. Thus the pattern-level advantage persisted when recipient preprocessing and marginal estimates used adaptation cells alone, including the uncertainty those estimates introduce into held-cell prediction.

3.2 The blood-reference model transfers to colon biopsies

Full assay patterns reduced mean deviance from 0.02588 to 0.02183 in the 11 independent colon donors, a reduction of 15.65% (95% interval, 11.39–21.09%; Fig. 1). Every donor improved relative to individual frequencies. Loss fell by 5.42% relative to protein patterns alone (2.88–8.66%; 10/11 donors) and by 31.13% relative to independence (25.83–37.11%; 11/11 donors). The three adjusted lower bounds were 10.89%, 2.56%, and 24.85%, respectively, meeting the prespecified criterion. This transfer used recipient measurements made separately within each assay without changing the interaction learned in blood. Additional file 1 reports all five arms and the descriptive clinical-group summaries.

3.3 Within-assay patterns improve dependence prediction

Conditional on the scoring-cell frequencies, retaining both assay patterns reduced deviance in all three cohorts (Table 1). Relative to individual frequencies alone, the reductions were 36.55% in Cambridge,

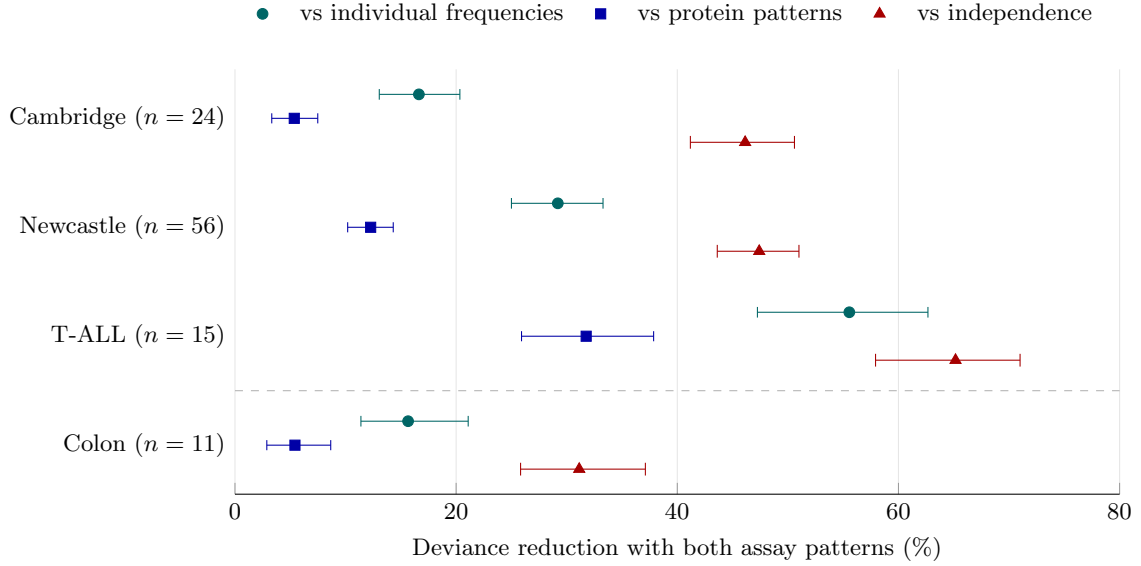


Fig. 1 Separate assay patterns improve prediction without scoring-cell summaries. Full-pattern reconstruction is compared with three controls at the same adaptation-derived marker frequencies. Points show relative mean-deviance reductions; bars are paired recipient-bootstrap 95% intervals. Each cohort uses 256 adaptation and 256 scoring cells per person. Protein thresholds, assay histograms, and marker frequencies use adaptation cells only. The source interaction is fixed across arms. All nine Bonferroni percentile intervals for the three retrospective cohorts have positive lower endpoints. The colon test, separated below, used the unchanged source model and a protocol published before count access; all three of its adjusted intervals are positive. Nominal intervals are shown.

Table 1 Prediction with a fixed source interaction and different assay information.

Recipient information	Cambridge ($n = 24$)	Newcastle ($n = 56$)	T-ALL ($n = 15$)
Individual marker frequencies	0.009606	0.015761	0.038007
RNA patterns and protein frequencies	0.007982	0.012231	0.025008
RNA frequencies and protein patterns	0.007106	0.010869	0.021243
Means and covariances (maximum entropy)	0.006399	0.008572	0.011553
Source pool, matched recipient frequencies	0.007094	0.011927	0.029302
Cohort pool, matched recipient frequencies	0.007043	0.010446	0.021385
Both assay-pattern distributions	0.006095	0.008377	0.010358

Mean donor deviance over all 81 marker pairs; lower is better. The source interaction, scoring cells, and recipient marker frequencies were fixed. Pooled controls replaced recipient adaptation profiles with separate-assay profiles from source donors or other people in the same cohort, keeping the prior fraction unchanged.

46.85% in Newcastle, and 72.75% in T-ALL. The full-pattern arm improved 23/24, 55/56, and 15/15 recipients, respectively.

Protein patterns gave lower mean loss than RNA patterns in each cohort, but retaining RNA patterns in addition improved prediction further. Against the protein-pattern arm, the reductions were 14.23% in Cambridge (95% interval, 9.46–19.07%), 22.92% in Newcastle (15.71–29.28%), and 51.24% in T-ALL (45.29–57.07%). All nine Bonferroni percentile intervals had positive lower endpoints. Retaining RNA patterns therefore improved prediction even when the full protein-pattern distribution was available.

3.4 Recipient covariances recover most of the pattern-level gain

The covariance-preserving replacement retained 91.36% (paired-bootstrap 95% interval, 89.09–94.03%), 97.36% (96.28–98.38%), and 95.68% (93.42–97.25%) of the loss reduction from individual frequencies to full patterns in Cambridge, Newcastle, and T-ALL, respectively. Full patterns improved further by

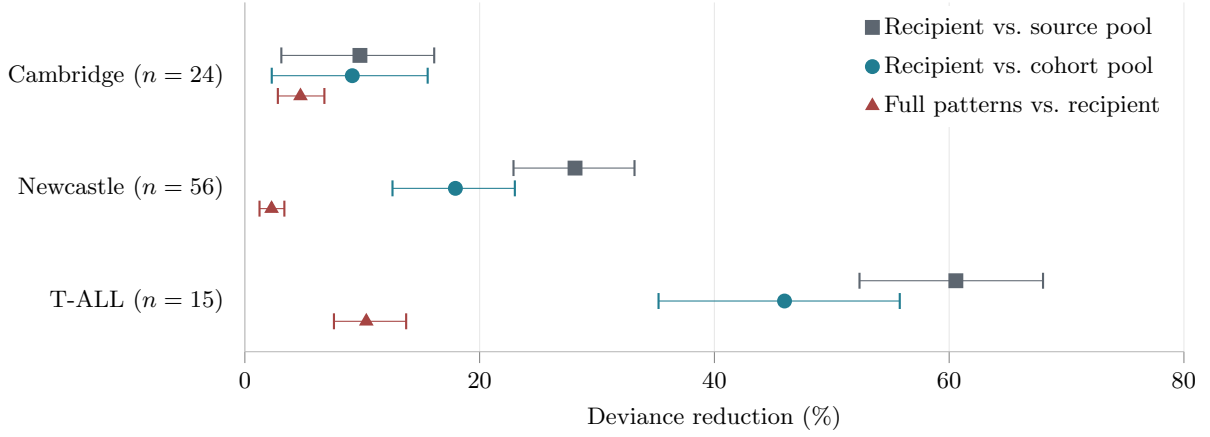


Fig. 2 Recipient-specific structure improves dependence transfer. Squares and circles compare recipient covariance-preserving reconstruction with source and other-recipient pools, respectively. Triangles show the further improvement from full recipient patterns over recipient covariance alone. Points are relative mean-deviance reductions; lines are paired donor-bootstrap 95% intervals conditional on the fitted models and pools. All arms use the same source interaction and recipient marker frequencies. Cambridge and Newcastle belong to the same study.

4.74% (95% interval, 2.81–6.78%), 2.27% (1.25–3.37%), and 10.35% (7.59–13.74%), improving 22/24, 44/56, and 15/15 people. All six contrasts remained positive under the separate Bonferroni adjustment. Pairwise within-assay statistics therefore recovered most of the gain, with a smaller additional benefit from retaining the original patterns.

Recipient-specific covariance also improved on both pooled controls (Fig. 2). Against the source pool, deviance fell by 9.80%, 28.13%, and 60.57% in Cambridge, Newcastle, and T-ALL. Against other recipients in the same cohort, the reductions were 9.15% (95% interval, 2.29–15.57%), 17.94% (12.58–23.00%), and 45.97% (35.24–55.79%), with improvements in 15/24, 42/56, and 14/15 people. All six Bonferroni percentile intervals had positive lower endpoints. Each recipient supplied 256 adaptation cells per assay, compared with 3,072 source cells or 3,584–14,080 cells in the cohort pool. Matching individual frequencies and using more cells from other people did not recover the recipient-specific gain.

3.5 Coexpression prioritizes recipient-specific corrections

The response score concentrated the recipient-specific gain into relatively few relationships (Fig. 3). Its top third accounted for 87.0%, 90.1%, and 85.5% of the net deviance reduction in Cambridge, Newcastle, and T-ALL, compared with 85.0%, 60.9%, and 38.5% under direct source-interaction ranking. In Newcastle, the mean gain per selected pair was 0.00506 under the response score and 0.00342 under direct ranking; the paired difference was 0.00164 (95% interval, 0.00085–0.00256). In T-ALL the gains were 0.02523 and 0.01135, a difference of 0.01388 (0.00906–0.01925). Both contrasts survived the six-comparison adjustment. Cambridge gave similar gains under the two rankings, 0.00168 and 0.00164 (difference, 0.00004; –0.00077 to 0.00085).

The top-to-bottom-third gain differences were 0.00171 in Cambridge (95% interval, 0.00029–0.00330), 0.00498 in Newcastle (0.00318–0.00705), and 0.02426 in T-ALL (0.01648–0.03365). The latter two retained

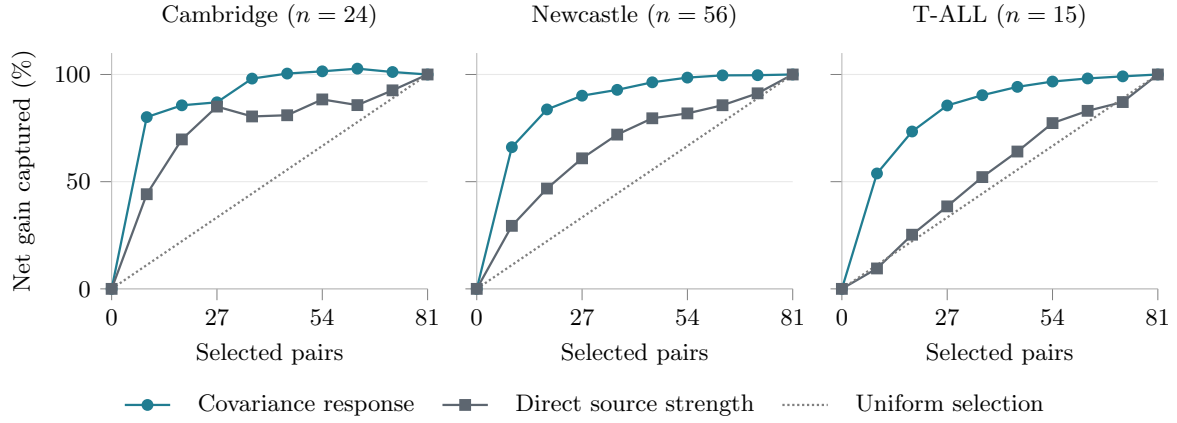


Fig. 3 Unpaired coexpression identifies relationships that benefit from correction. Pairs are ranked separately within each person. Curves show cumulative deviance improvement from recipient covariance over the cohort pool, divided by the net improvement across all 81 pairs. Teal uses the covariance-response score; gray uses direct source-interaction strength. The dotted line is the expected fraction under uniform selection. Curves are descriptive cohort aggregates; donor-bootstrap comparisons at 27 pairs are reported in the text. Net gain can exceed 100% when excluded corrections worsen prediction.

positive Bonferroni intervals. In T-ALL, the three highest cohort-mean scores were for CD7 RNA-CD4 protein, CD4 RNA-CD52 protein, and CD7 RNA-CD52 protein; their mean deviance gains were 0.04513, 0.01245, and 0.04981. The original study used CD4 and CD7 to distinguish T-cell populations and included CD52 in a leukemia expression signature [11]. These corrections concern marker combinations associated with the cohort’s cellular heterogeneity.

3.6 The response ranking persists at fitted strength

The response coefficient remained closely aligned with exact model discrepancy throughout the continuation (Additional file 1). At $t = 1$, mean within-recipient Spearman correlations were 0.941 in Cambridge (95% interval, 0.935–0.948), 0.942 in Newcastle (0.937–0.947), and 0.939 in T-ALL (0.919–0.954). The response and exact rankings shared 90.3%, 89.5%, and 89.1% of their top 27 queries, respectively. The cohort-aggregate exact discrepancy at $t = 1$ was 40.8%, 32.1%, and 28.3% of the quadratic approximation.

4 Discussion

The resolution at which a relationship is reported need not be the resolution at which it should be transferred. A binary RNA-protein table depends on the distribution of finer cellular profiles. Reconstructing those profiles before aggregation allowed one source interaction to predict different recipient associations using measurements made separately within each assay. The adaptation-only analysis establishes this advantage without supplying scoring-cell summaries. The conditional analysis then separates the contribution of coexpression from marginal-frequency estimation error. The independent colon test extends the pattern-level advantage to a new tissue and study without refitting the blood-reference interaction.

The pooled controls distinguish recipient-specific coexpression from generic population structure: more observations from other people did not replace the recipient’s own assay patterns. Equation (6) explains how within-assay covariances alter the association induced by a fixed source interaction. It also explains

why ranking direct source coefficients can miss relationships that change in a recipient. The response score uses the covariances around each query to identify these relationships, improving prioritization in two cohorts. At the other extreme, the coarsening bound describes the range of associations left unresolved when those measurements are unavailable. Together, these results connect the detail retained in separate assays to both the limits of prediction and the choice of queries.

5 Limitations

The blood and leukemia analyses reuse previously examined recipients from two studies. The colon analysis was specified before count access, but uses one existing study and eleven donors; it is not a prospective specimen collection. Its clinical groups are too small to establish disease- or treatment-specific effects. The adaptation-only test changes the protein-state definition as well as the information boundary; its effect sizes cannot be subtracted from the conditional results to measure information leakage. Intervals condition on the source fit and fitted pools, omitting their estimation uncertainty, dependence between overlapping cohort pools, and the adaptive research history.

The panel contains nine RNA and nine protein markers. The conditional analysis uses donor-wide protein ranks and individual scoring-cell marker frequencies; the adaptation-only analysis uses neither. Two Stephenson recipients have almost no counts across the protein panel and were retained. The task is fixed-margin dependence prediction from separately measured assays, not missing-modality imputation without target assay observations. The results do not isolate within-cell-type regulation or establish molecular binding, causal effects, or clinical utility.

The identified interval assumes a known transferable interaction, finite assay supports, and exact recipient summaries. Its all-marker endpoints require global optimization over the corresponding marginal polytopes; only the two-frequency case has the closed form reported here. The examples establish ambiguity, not its biological magnitude. Query-wise minimax predictions need not form a coherent multivariate law. Interaction drift, sampling error, and coarsening ambiguity remain distinct.

We tested one maximum-entropy replacement, which preserves fitted means and covariances but can induce higher-order correlations. The retained gain is descriptive, not causal mediation. The exact response identity holds at finite strength, but the quadratic coefficient overestimated the magnitude of model discrepancy at $t = 1$. Its query score predicts model discrepancy, not error against an arbitrary biological population. The query-ranking advantage over direct source strength was unresolved in Cambridge. Prioritization requires the full measured panel and does not establish savings from measuring fewer markers. Sorting and cell mixtures can contribute to the observed coexpression.

Reconstruction, noncollapsibility, and minimax KL prediction have established foundations. The theory concerns the resolution-dependent identified set; the experiments establish an information advantage,

not a superior interaction learner. The earlier matched CHAMPOLLION-objective comparison did not
establish superiority and remains documented with the historical benchmark.

6 Conclusions

Within-assay cellular patterns improved RNA–protein prediction across cohorts and in an independent
blood-to-colon test with the source interaction held fixed, including prediction on untouched cells without
supplying their marginal summaries. In the conditional analysis, recipient-specific covariances retained
most of the pattern-level gain and outperformed source and cohort pools matched to the same marker
frequencies. Full patterns improved further in each cohort. Separately measured recipient coexpression can
therefore support molecular predictions that individual frequencies and cohort averages do not recover.

Abbreviations

CITE-seq: cellular indexing of transcriptomes and epitopes by sequencing; KL: Kullback–Leibler; RNA:
ribonucleic acid; T-ALL: T-cell acute lymphoblastic leukemia.

Declarations

Ethics approval and consent to participate

This study used only public, de-identified data and involved no new participants, specimens, interventions,
or identifiable records; the original studies report their ethics approvals and consent procedures.

Consent for publication

Not applicable.

Availability of data and materials

Source data are available under E-MTAB-10026 (Stephenson CITE-seq), GSE248287 (T-ALL), and
GSE250490 (colon biopsies). The colon analysis uses the public count matrix in Figshare article 21919356,
version 3 [12]; its frozen protocol and complete results are released separately [13]. The original benchmark’s
analysis plans, software, and study records are available in the v2.0.4 release [14]. The accompanying
analysis artifact contains derived assay distributions, reconstruction code, predictions, and recipient-level
losses for the subsequent resolution analyses, together with executable checks of the reported comparisons.
Additional file 1 specifies the protocols. Raw matrices are not redistributed.

Competing interests

Competing-interest information is withheld in the blinded manuscript.

Funding

Funding information is withheld in the blinded manuscript.

Authors' contributions

Author names and contribution statements are withheld in the blinded manuscript.

Acknowledgements

OpenAI Codex assisted with code development, literature retrieval, manuscript editing, and L^AT_EX preparation.

Additional file

Additional file 1 (PDF). *Mathematical results and analysis protocols*. Identified-set and response proofs, adaptation-only and conditional protocols, cohort-level comparisons, and computational verification.

References

- [1] Sei T, Yano K. Minimum information dependence modeling. *Bernoulli*. 2024;30:2623–2643. <https://doi.org/10.3150/23-BEJ1687>.
- [2] Sei T, Yano K. Estimation of bivariate minimum information dependence models using the density-ratio framework. *Annals of the Institute of Statistical Mathematics*. 2026. <https://doi.org/10.1007/s10463-026-00995-8>.
- [3] Li R, Ye X, Zhou H, Zha H. Learning to match via inverse optimal transport. *Journal of Machine Learning Research*. 2019;20(80):1–37.
- [4] Samaran J, Peyré G, Cantini L. CHAMPOLLION: Robust Multi-Omics Integration via Inverse Optimal Transport Using Paired Cells. *bioRxiv*. 2026. <https://doi.org/10.64898/2026.04.28.721317>.
- [5] Ghosh S, Vellaisamy P. Marginal log-linear parameters and their collapsibility for categorical data. *Statistica Neerlandica*. 2024;78(3):523–543. <https://doi.org/10.1111/stan.12332>.
- [6] Bergsma WP, Rudas T. Marginal models for categorical data. *The Annals of Statistics*. 2002;30(1):140–159. <https://doi.org/10.1214/aos/1015362188>.
- [7] Hammond C, van der Heijden PGM, Smith PA. Generating contingency tables with fixed marginal probabilities and dependence structures described by loglinear models. *Journal of Statistical Computation and Simulation*. 2024;94(12):2797–2812. <https://doi.org/10.1080/00949655.2024.2353760>.
- [8] Schneidman E, Berry MJ II, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*. 2006;440:1007–1012. <https://doi.org/10.1038/nature04701>.
- [9] van Erven T, Harremoës P. Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*. 2014;60(7):3797–3820. <https://doi.org/10.1109/TIT.2014.2320500>.

- [10] Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*. 2021;27:904–916. <https://doi.org/10.1038/s41591-021-01329-2>.
- [11] Wiggers CRM, Cho EY, Ozdemir M, et al. Remodeling of the immune microenvironment is linked to adverse outcome in pediatric T cell acute lymphoblastic leukemia. *Nature Communications*. 2025;16:10263. <https://doi.org/10.1038/s41467-025-65134-y>.
- [12] Mennillo E, Kim YJ, Lee G, Rusu I, Patel RK, et al. Single-cell and spatial multi-omics highlight effects of anti-integrin therapy across cellular compartments in ulcerative colitis. *Nature Communications*. 2024;15:1493. <https://doi.org/10.1038/s41467-024-45665-6>.
- [13] Assay-resolution study contributors.: Blood-to-colon dependence transfer: frozen analysis protocol and results. Protocol frozen before count access in commit b6e6cab, 13 September 2026. https://github.com/sushaan-k/coupling-fields-benchmark/tree/main/analysis/colon_generalization.
- [14] sushaan-k. Coupling Fields: public benchmark, version 2.0.4 [Internet]. GitHub; 2026. Available from: <https://github.com/sushaan-k/coupling-fields-benchmark/releases/tag/coupling-fields-v2.0.4-public-benchmark>. Accessed 5 September 2026.