ROW SELECTION IN CONTEXT-LABEL ATTACKS ON TABULAR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A tabular foundation model uses labeled context rows to predict new rows. Changing a small number of context labels can therefore affect many predictions. We study how the choice of rows changes this damage. We compare neighbor-degree, mean-distance, and query-coverage selectors across TabPFNv2-12L, TabICLv2, and Mitra. Class-matched random controls separate total corruption damage from the additional damage due to row selection. Hub attacks span five count budgets and four percentage budgets. Their total loss grows with corruption, while additional damage depends on geometry and budget. On eight external datasets, representation hubs and random selection have equivalent normalized loss at 25 labels within a margin of 0.005. At 15%, input hubs and mean-distance selection add 4.16 and 2.60 percentage points of accuracy loss relative to their respective controls. Input anti-hubs, which appear in few neighbor sets, add 1.00 percentage point at 5% and retain additional loss under coverage matching, repeated tie choices, and row permutations. In an exploratory single-label analysis, query coverage has a clearer association with influence than neighbor degree. Selecting high-coverage rows adds 1.38 percentage points of accuracy loss at 5%. Together, these results identify how selection geometry, attack size, and the prediction measure shape context-label attacks.

1 INTRODUCTION

Tabular foundation models (TFMs) predict new rows from a table of labeled examples supplied at inference time. These examples form the *context*, and the new rows are the *queries*. Models such as TabPFNv2, TabICLv2, and Mitra use this context without updating their weights (Hollmann et al., 2025; Qu et al., 2026; Zhang et al., 2025). The labels in the context are therefore part of the prediction input. An attacker who can edit a label store can change this input while leaving the features and model weights fixed.

The effect of an attack may depend on which rows receive incorrect labels. A *geometric selector* is a rule that chooses rows using distances or neighborhoods among feature vectors or model representations. A hub is a row that appears in many other rows' nearest-neighbor sets; an anti-hub appears in few such sets. A third choice is to select rows with small average distance to their neighbors. Prior work used this mean-distance rule to attack TFMs by changing context labels (Biloš et al., 2026). We examine these choices alongside a query-aware rule that selects context rows close to many queries.

To measure what row selection adds, we compare each attack with a *matched-random counterfactual*: a random attack that changes the same number of labels from each original class and uses the same replacement labels. For example, if an attack changes ten class-A labels to B and fifteen class-B labels to A, its random control makes the same changes on randomly selected rows from those classes. Both attacks use the same features, queries, and model. Their difference measures the additional damage associated with choosing particular rows. We call this difference *selector specificity* and define it formally in Section 3.3.

We make four contributions to the study of context-label attacks:

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.

1. We measure input- and representation-hub attacks across budgets of 1–50 labels and 0.5%–15% of the context. The curves separate growing total corruption damage from the additional effect of hub selection (Figure 1).
2. We compare reverse-neighbor degree with mean-distance selection at 25 labels, 5%, and 15%. Their effects depend on both the budget and the prediction measure. At 15%, input hubs and mean-distance selection add 4.16 and 2.60 percentage points of accuracy loss relative to their own matched controls (Table 1).
3. We show that input anti-hubs cause additional accuracy loss at both 25 labels and 5%, including a direct advantage over input hubs (Table 2). Their additional normalized loss at 5% persists under coverage matching, repeated tie choices, and row permutations (Table 6).
4. We connect single-label influence with query-aware attacks. Comparing seven row properties shows a clearer exploratory association for query coverage than for neighbor degree (Table 5). Selecting rows near many queries adds 1.38 percentage points of accuracy loss at 5% (Table 6).

Together, these comparisons explain when different row-selection rules add damage beyond class-matched label corruption.

2 Background and related work

In-context learning predicts from examples included with the input, without changing model weights (Brown et al., 2020). Prior-fitted networks learn this behavior by training on synthetic prediction tasks (Müller et al., 2022; Hollmann et al., 2023). TabPFNv2 applies this approach to tabular classification (Hollmann et al., 2025). TabICL processes columns and rows before combining context information for prediction (QU et al., 2025); we will evaluate its later TabICLv2 version (Qu et al., 2026). Mitra learns from a mixture of synthetic data distributions (Zhang et al., 2025). In each model, the labeled context is supplied when a prediction is requested, making context quality part of reliable deployment.

Data poisoning studies how altered training examples affect learned models (Biggio et al., 2012; Koh & Liang, 2017). Related work on language models examines malicious prompts and corrupted in-context examples (Perez et al., 2022; Greshake et al., 2023; He et al., 2025). For TFMs, Djilani et al. (2025) study test-time feature attacks and defenses that modify the context. Our setting gives the attacker control over context labels while holding all features and model parameters fixed. Biloš et al. (2026) connect TFM predictions to similarity-based aggregation and attack context rows with small mean distance in model representations. We evaluate that mean-distance selector with random controls matched to its label changes, alongside hub, anti-hub, and query-coverage selectors.

Nearest-neighbor geometry provides several ways to choose influential rows. The k nearest neighbors of a row are the k other rows with the smallest distances to it. A row's reverse-k-nearest-neighbor degree counts how often it occurs in those neighbor sets. High-degree rows are called hubs, a common feature of high-dimensional data (Radovanović et al., 2010; Tomasev et al., 2011). Mean neighbor distance measures local closeness, whereas reverse degree measures how often other rows select a point as a neighbor. Query coverage applies the same neighbor idea to the queries being predicted. These quantities give concrete, distinct rules for selecting labels to change.

3 Attack methods

3.1 Attack setting and class-matched label changes

Let $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^{n}$ be a context of n rows. The vector x_j contains row j's features and $y_j \in \{1, \ldots, C\}$ is its label, where C is the number of classes. The model predicts labels for m query feature vectors $u_1, \ldots, u_m$. An attacker selects a set $S \subseteq \{1, \ldots, n\}$ of b context rows and replaces their labels. Thus, $b = |S|$ is the attack budget, and $|S|$ denotes the number of elements in the set. Features and model weights remain fixed. The attacker observes context labels but never query labels; the query-aware selector and coverage-matched controls also use query features (Table 8).

We define the *source-class count* $b_c = |\{j \in S : y_j = c\}|$ as the number of selected rows whose original label is class c. The counts satisfy $\sum_{c=1}^{C} b_c = b$. A class permutation π assigns each original class a different replacement class, so $\pi(c) \neq c$. Every selected row from class c receives label $\pi(c)$. The *source-class flow* is the collection of these counts and destinations: b_c labels move from c to $\pi(c)$. For each target attack we draw three random controls, sampling b_c rows uniformly without replacement from each original class and applying the same π. This preserves both the number of changed labels and the resulting class counts.

3.2 GEOMETRIC SELECTORS

We first define a neighbor graph, then specify the distance used by each attack. Let d_{ij} be the distance from context row i to context row j, and fix $k = 10$. For each i, let τ_i be the kth smallest distance among rows $j \neq i$. To handle nearly equal distances, define the tolerance $\epsilon_{ij} = 10^{-12} + 10^{-10} \max(|d_{ij}|, |\tau_i|)$. The sets of strictly closer and tied neighbors are

$$L_i = \{j \neq i : d_{ij} < \tau_i - \epsilon_{ij}\}, \qquad T_i = \{j \neq i : |d_{ij} - \tau_i| \leq \epsilon_{ij}\}.$$

The adjacency weight A_{ij} measures how much row j belongs to row i's neighbor set:

$$A_{ij} = \begin{cases} 1, & j \in L_i, \\ (k - |L_i|)/|T_i|, & j \in T_i, \\ 0, & \text{otherwise}, \end{cases} \qquad h_j = \sum_{i=1}^{n} A_{ij}. \tag{1}$$

Tied neighbors share the remaining weight, so each row contributes exactly k and $A_{ii} = 0$. The scalar h_j is row j's reverse-neighbor degree: the total weight it receives from other rows. Large h_j identifies a hub, and small h_j identifies an anti-hub. The graph is described by A_{ij}; h_j is a score for a single row.

The *representation hub* selector, abbreviated Rep-RkNN, uses cosine distance between the model's context-row representations. Here RkNN abbreviates reverse k nearest neighbors. If r_i and r_j are nonzero representation vectors, their distance is $d_{ij} = 1 - r_i^\top r_j / (\|r_i\|_2 \|r_j\|_2)$, where $r_i^\top r_j$ is their inner product and $\|\cdot\|_2$ is Euclidean length. We take representations from the ninth layer of TabPFNv2-12L, the eleventh context-processing layer of TabICLv2, and the twelfth layer of Mitra. The selector takes the b largest degrees while leaving at least two unchanged rows in each original class.

The *input hub* selector, Input-RkNN, applies the same graph to a distance computed directly from the features. We use Gower distance, which averages feature-wise differences across numeric and categorical columns. For numeric column ℓ, let $x_{j\ell}$ be its value in row j, μ_ℓ its median, and s_ℓ a positive scale. Its transformed value is $\widetilde{x}_{j\ell} = \max(-5, \min(5, (x_{j\ell} - \mu_\ell)/s_\ell))$. The scale is the first positive value among the interquartile range (IQR), 1.4826 times the median absolute deviation (MAD), the 95th-minus-5th percentile range, and the full observed range. IQR is the 75th percentile minus the 25th percentile; MAD is the median of $|x_{j\ell} - \mu_\ell|$. A constant column is transformed to zero. Numeric feature ℓ contributes $\min(|\widetilde{x}_{i\ell} - \widetilde{x}_{j\ell}|/10, 1)$ to a pair's distance. Categorical features contribute zero for agreement and one for disagreement. Missing numeric values are omitted from that pair's average, and missing categorical values form a separate category. Scaling uses context rows for this selector.

The *input anti-hub* selector uses the input hub's source-class counts b_c, choosing the b_c lowest-degree rows within each class. The input hub, input anti-hub, and their random controls thus share the same class flow. When equal scores span a class's selection boundary, we randomly choose the required number of rows. The *mean-distance* selector follows Biloš et al. (2026). Let N_j contain row j's k closest representation neighbors. Its mean distance is $\bar{d}_j = k^{-1} \sum_{i \in N_j} d_{ji}$, and the selector chooses the b smallest values. Distance centrality is $-\bar{d}_j$, so a larger value means a more central row. This measures local closeness, whereas h_j counts how often other rows choose j as a neighbor. Its distance and selection ties use the original row order. Its random controls match the class counts produced by its own selected set.

The *query-coverage* selector uses Gower distance between context and query features, with scaling computed from their combined feature rows. Let W_{ij} be the neighbor weight that query i assigns to context row j, using the same cutoff and tie rule as (1). The coverage $v_j = m^{-1} \sum_{i=1}^{m} W_{ij}$ is the

average share of queries covered by row j. Each query distributes a total weight of k among context rows. We select the b_c largest coverages within each class, using the input hub's class flow. To test whether coverage explains the anti-hub effect, we also construct *coverage-matched controls*. Each anti-hub row is matched to an unused row of the same class with the closest coverage; target rows are excluded when enough alternatives exist. Random choices resolve equally close matches, and we average three controls.

3.3 Target estimands

An estimand is the quantity a comparison aims to measure. Let M be an evaluation measure oriented so that larger values mean more harm, such as prediction loss or error rate. Write M_0 for its value on the original context and $M(S)$ for its value after changing the labels in set S. We measure both the attack's damage and its additional damage over random:

$$D_M(S) = M(S) - M_0, \qquad S_M(S) = D_M(S) - \frac{1}{3}\sum_{r=1}^{3} D_M(R_r). \tag{2}$$

Here R_r is the selected set for matched-random draw r. The second quantity is selector specificity. For example, positive accuracy specificity means that the chosen rows cause a larger accuracy drop than random rows with the same class flow. Input hubs and anti-hubs share the same controls, so subtracting their specificities directly compares their attack damage.

3.4 Prediction measures and statistical analysis

For query i, let t_i be its true class and p_{ic} the model's predicted probability of class c after reading the given context. Thus $\sum_{c=1}^{C} p_{ic} = 1$. We use normalized negative log-likelihood and the multiclass Brier score:

$$\mathrm{nNLL} = -\frac{1}{m \log C}\sum_{i=1}^{m} \log p_{i,t_i}, \qquad \mathrm{Brier} = \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{C}(p_{ic} - \mathbf{1}\{t_i = c\})^2. \tag{3}$$

The indicator $\mathbf{1}\{t_i = c\}$ equals one when the true class is c and zero otherwise; $\log$ denotes the natural logarithm. Normalized loss equals one for uniform class probabilities. Accuracy drop is the clean accuracy minus the attacked accuracy, reported in percentage points (pp). Prediction flip is the fraction of queries whose most probable class changes after the attack.

For each dataset, we average the five split results within each model and then average the three models. We form confidence intervals (CIs) by sampling datasets with replacement $10{,}000$ times. For a positive-effect test, the bootstrap p-value is $(1 + \sum_{r=1}^{B} \mathbf{1}\{\Delta_r^* \le 0\})/(B+1)$, where $B = 10{,}000$ and Δ_r^* is the mean effect in resample r. Normalized loss is the primary outcome. For each contrast and budget, Holm adjustment jointly accounts for tests of accuracy, Brier, and flip. We report the additional damage with its 95% CI. Practical equivalence means that the 90% CI lies inside ± 0.005 for normalized loss or ± 0.25 pp for accuracy. For the eight external datasets, we also report exact sign-flip tests, which enumerate all $2^8 = 256$ assignments of positive and negative signs to the dataset effects (Appendix D).

3.5 Influence of an individual context label

To measure how much one row can affect predictions, we change its label to each of the other $C-1$ classes in turn. Let L_0 be the original context's normalized query loss and $L_{j\to c}$ the loss when only row j is relabeled as class c. The row's mean influence is

$$I_{\mathrm{mean}}(j) = \frac{1}{C-1}\sum_{\substack{c=1 \\ c \neq y_j}}^{C} \big(L_{j\to c} - L_0\big). \tag{4}$$

Thus, $I_{\mathrm{mean}}(j)$ averages the increase in normalized loss over every incorrect replacement label. The features, all other labels, and the query set stay fixed. We sample up to twenty rows from each of ten representation-degree groups, covering the degree distribution on eight datasets. Within each

dataset, split, and model, we compute Spearman correlations, which compare the ranks of a row property and its influence. We then average splits and models at the dataset level. Query coverage in this exploratory analysis uses membership in a set of exactly k neighbors per query. The attack experiments use the fractional weights defined in Section 3.2. We compare input and representation degree, negative input degree (the anti-hub score), distance centrality, query coverage, and two neighbor-label properties. Same-label support is the share of neighbor weight assigned to the row's own class. Label disagreement is the share assigned to other classes after giving closer neighbors more weight. Their detailed definitions and the full exploratory comparisons appear in Appendix F.

4 EXPERIMENT DESIGN

We evaluate TabPFNv2-12L, TabICLv2, and Mitra 1.1 with fixed weights. Our eight-dataset test panel contains anneal, analcatdata_authorship, cmc, Fitness_Club, Is-this-a-good-customer, kc2, maternal_health_risk, and MIC. A broader panel adds sixteen datasets (Tables 10 and 11). For each dataset, seeds 0 through 4 produce five splits with about 80% context rows and 20% queries, balancing class proportions. Identical feature rows remain together; MiceProtein also groups measurements from the same mouse. Contexts in the test panel contain 417–$1{,}378$ rows. Attacks within a dataset, split, and model share the same queries and original predictions.

Input- and representation-hub attacks use count budgets $b \in \{1, 5, 10, 25, 50\}$ and percentage budgets $b = \lceil \rho n \rceil$ for $\rho \in \{0.005, 0.01, 0.05, 0.15\}$. Here ρ is the fraction of context labels changed, and the ceiling rounds upward to a whole label. Each target has three random controls with the same source-class flow. The complete comparison contains $24 \times 5 \times 3 \times 2 \times 9 = 6{,}480$ paired settings. To compare count budgets on a fixed dataset panel, we retain contexts where even 50 changes affect at most 10% of labels. This gives seven test datasets and sixteen broader-panel datasets at every count budget. Percentage curves use all eight and all twenty-four datasets, respectively. We display 15% separately from the lower fractions. The curve intervals describe individual budget estimates, with no joint test across budgets.

The centrality comparison uses all eight test datasets at 25 labels, 5%, and 15%. Mean-distance selection uses the same budgets, including the 15% setting of Biloš et al. (2026). Hub–anti-hub and query-aware comparisons use 25 labels and 5%, giving 240 settings each. The latter budget changes 21–69 labels. Anti-hub robustness uses five alternative tie choices, five row permutations, and three coverage-matched controls per setting. The separate exploratory influence study uses banknote-authentication, blood-transfusion-service-center, breast-w, breast_cancer, climate-model-simulation-crashes, cylinder-bands, dresses-sales, and iris.

5 RESULTS

Results average three models on the eight test datasets unless another panel is specified. Positive specificity means additional damage relative to the stated control. Intervals are 95% CIs, except for the 90% intervals used to assess equivalence.

5.1 TOTAL CORRUPTION DAMAGE AND ADDITIONAL SELECTION DAMAGE

Figure 1 separates total normalized-loss damage from hub specificity across both budget rules. On the seven-dataset count panel, total damage rises from approximately zero at one label to 0.0190 for input hubs and 0.0142 for representation hubs at fifty labels. Their matched-random controls reach 0.0159 in both cases. The resulting additional damage is 0.0031 and -0.0017, respectively. The rising total loss therefore includes a substantial contribution from label corruption shared with random selection.

On the eight-dataset percentage panel, total damage increases from 0.0157 to 0.0907 for input hubs between 5% and 15%. Its matched-random damage increases from 0.0139 to 0.0645, leaving specificity of 0.0018 and 0.0262. Representation-hub specificity changes from -0.0019 to 0.0055 over the same budgets. Thus, the extra effect of hub selection depends on geometry and attack size, alongside the increase in overall corruption damage. The broader-panel loss curves and both panels' accuracy curves appear in Appendix C.

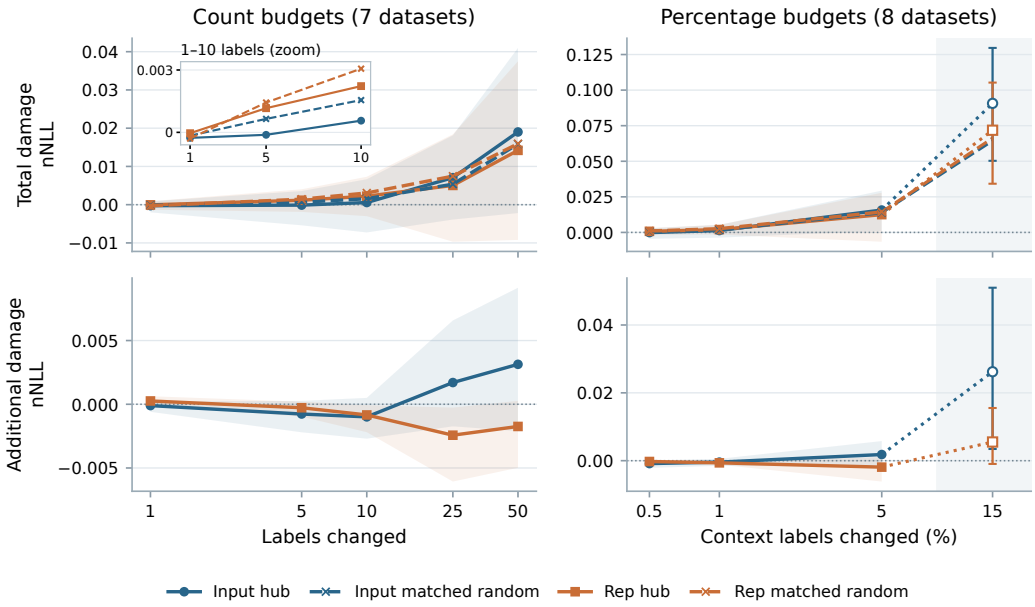

Figure 1: Hub attacks across count and percentage budgets. Top: total normalized-loss damage for hub attacks (solid lines) and their matched-random controls (dashed lines). Bottom: hub minus matched-random damage within each color (selector specificity). Shading and error bars show pointwise 95% CIs for selected-row damage and specificity. Count curves use the same seven datasets throughout; percentage curves use all eight test datasets. Dotted extensions and open markers identify the separate 15% comparison. Main axes use logarithmic spacing, the inset magnifies the one- to ten-label curves.

5.2 Centrality definitions yield different attack effects

Table 1 compares input degree, representation degree, and mean distance on the same eight datasets. Mean-distance accuracy specificity grows from 0.08 pp at 25 labels to 0.60 pp at 5% and 2.60 pp at 15%. Its normalized-loss specificity stays near zero, with CIs spanning zero at all three budgets. The 15% accuracy comparison has exact Holm-adjusted $p = 0.0469$ (Table 14). Accuracy measures whether the predicted class is correct, whereas normalized loss measures the probability assigned to the true class; they capture different aspects of the attack.

Table 1: Additional damage from three centrality-based selectors on all eight test datasets. Each attack is compared with its own class-matched random control. Entries are means with pointwise 95% CIs; accuracy is in percentage points. The 15% budget is a separate higher-fraction comparison.

Budget	Selector	nNLL	Accuracy drop (pp)
25 labels	Input-RkNN	+0.0024 [-0.0011,+0.0067]	-0.006 [-0.242,+0.300]
25 labels	Rep-RkNN	-0.0016 [-0.0052,+0.0011]	+0.048 [-0.236,+0.382]
25 labels	Mean distance	-0.0032 [-0.0083,+0.0011]	+0.075 [-0.098,+0.240]
5%	Input-RkNN	+0.0018 [-0.0018,+0.0058]	+0.373 [+0.005,+0.911]
5%	Rep-RkNN	-0.0019 [-0.0062,+0.0010]	-0.111 [-0.272,+0.032]
5%	Mean distance	-0.0018 [-0.0115,+0.0058]	+0.603 [+0.138,+1.140]
15%	Input-RkNN	+0.0262 [+0.0034,+0.0510]	+4.158 [+1.884,+6.525]
15%	Rep-RkNN	+0.0055 [-0.0010,+0.0155]	+0.265 [+0.004,+0.562]
15%	Mean distance	+0.0010 [-0.0271,+0.0239]	+2.603 [+0.815,+4.482]

The degree selectors also differ. At 15%, input hubs add 0.0262 normalized loss and 4.16 pp accuracy drop, compared with 0.0055 and 0.27 pp for representation hubs, each relative to its own random control. At 25 labels, representation-hub normalized loss is practically equivalent to matched random: its difference is -0.00156, with 90% CI $[-0.00457, 0.00080]$. The corresponding estimate on the

eligible twenty-two-dataset subset is $+0.00075$, with CI $[-0.00183, 0.00383]$. Both intervals lie within ± 0.005 (Table 13). This equivalence concerns representation degree, normalized loss, and the fixed budget. Mean distance and reverse degree also select different rows: among sampled rows in the influence study, their representation-space rank correlation is 0.550 and their 25-row sets share 19.3% of their union (Table 17). These comparisons distinguish local closeness from frequent membership in other rows' neighborhoods.

5.3 Anti-hub selection adds damage at both budget rules

At 5%, changing input anti-hub labels raises normalized query loss by 0.0217 relative to the clean context, with CI $[0.0039, 0.0367]$. Seven of eight datasets have a positive mean change (Table 9). Relative to matched random, anti-hubs add 0.64 pp accuracy loss at 25 labels and 1.00 pp at 5% (Table 2). The 5% accuracy difference is positive on all eight datasets. The direct comparison with input hubs also favors anti-hubs in accuracy damage, by 0.65 and 0.63 pp at the two budgets.

Normalized-loss evidence is clearer at 5%: the additional loss is 0.0078, with CI $[0.0014, 0.0158]$ and exact sign-flip $p = 0.0273$. At 25 labels, its CI spans zero (Table 2). Brier and prediction flip increase relative to matched random at both budgets (Figure 2). Their additional damage is 0.0054 and 0.0100 at 25 labels, rising to 0.0103 and 0.0152 at 5%. For 5% accuracy, Brier, and flip, the exact Holm-adjusted p-values are each 0.0117; complete estimates and tests remain in Tables 12 and 14.

Table 2: Input anti-hub attacks at both budgets on eight datasets. Entries are additional damage with 95% CIs. Accuracy differences are in percentage points. Anti-hub and input-hub attacks share class counts and replacement labels, allowing a direct comparison.

Budget	nNLL: anti–random	Accuracy: anti–random	Accuracy: anti–hub
25 labels	+0.00303 [-0.00002,+0.00645]	+0.64 [+0.11,+1.26]	+0.65 [+0.21,+1.19]
5%	+0.00775 [+0.00140,+0.01575]	+1.00 [+0.43,+1.65]	+0.63 [+0.22,+1.13]

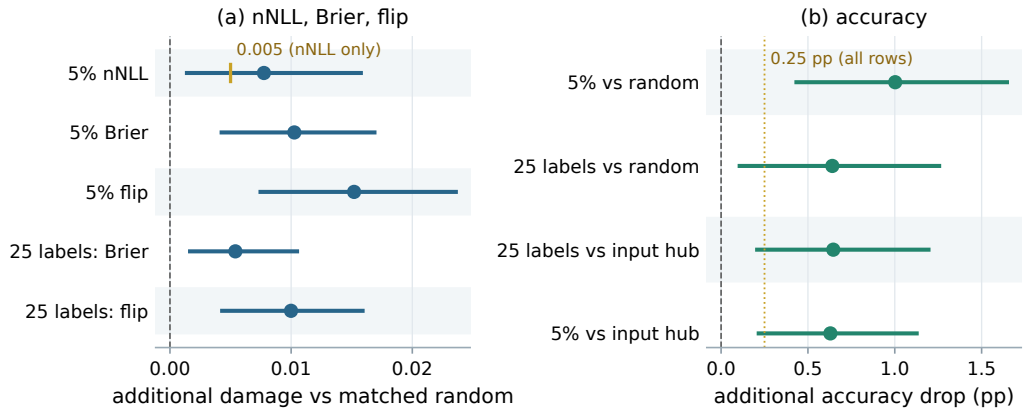

Figure 2: Additional damage from input anti-hub selection. Left: normalized loss, Brier, and prediction flip relative to matched random. Right: accuracy loss relative to the named control. Points average models and datasets; bars are 95% CIs. Gray dashed lines mark zero additional damage. Gold marks show reference effect sizes: 0.005 for the nNLL row only (left), and 0.25 pp for all accuracy rows (right).

Table 3 separates the 5% results by model. All three models have positive additional accuracy loss, with each model's CI above zero. Normalized-loss differences vary more across models: all three point estimates are positive, while only TabPFNv2-12L's interval excludes zero. At 25 labels, mean accuracy, Brier, and flip differences against random remain positive when any one dataset is omitted. The same holds for the mean accuracy difference against input hubs.

Table 3: Additional damage from 5% input anti-hub selection relative to matched random. Each row summarizes eight datasets for one model, with 95% dataset-bootstrap CIs; intervals are unadjusted for comparisons across models.

Model	nNLL	Accuracy drop (pp)
Mitra 1.1	+0.0030 [-0.0019,+0.0085]	+0.68 [+0.18,+1.23]
TabICLv2	+0.0058 [-0.0058,+0.0183]	+1.23 [+0.57,+1.96]
TabPFNv2-12L	+0.0145 [+0.0035,+0.0273]	+1.10 [+0.33,+1.93]

The two budget rules change both attack size and the selected rows (Table 4). The percentage rule changes a median of 50 labels, compared with the fixed 25. For selected sets S and T, the Jaccard index $J(S,T) = |S \cap T|/|S \cup T|$ is the fraction of their union shared by both. Its mean is 0.544. The 25-row set is fully contained in the 5% set on five datasets, partly contained on two, and larger on kc2. This set comparison helps interpret the change in damage across the two budgets.

Table 4: Attack size and selected-row overlap for the two budget rules. Context size is averaged over splits. Jaccard measures the shared rows divided by the union of the two anti-hub sets. Containment is the fraction of settings in which the 25-row set is fully contained in the 5% set.

Dataset	Context rows	Fixed	5%	Jaccard	Containment
Fitness_Club	1200	25	60	0.417	1.00
Is-this-a-good-customer	1378	25	69	0.362	1.00
MIC	1359	25	68	0.368	1.00
analcatdata_authorship	672	25	34	0.688	0.60
anneal	719	25	36	0.694	1.00
cmc	1178	25	59	0.424	1.00
kc2	418	25	21	0.798	0.00
maternal_health_risk	811	25	41	0.602	0.80

5.4 Single-label influence and query-aware selection

Table 5 compares geometric and neighbor-label properties as predictors of single-label influence. Representation and input degree have mean rank correlations of 0.005 and 0.029, with CIs spanning zero. Anti-hub score, distance centrality, label disagreement, and same-label support also have intervals spanning zero. These estimates provide no stable positive association between degree and individual-label influence in this study. This question differs from the joint effect of changing a selected set of labels measured by the attack experiments.

Table 5: Row properties and mean single-label influence I_{mean} on eight influence datasets. Spearman correlations are averaged over splits and models before summarizing datasets. Intervals are 95% CIs. These associations are exploratory; all 16 comparisons and multiple-test adjustments appear in Table 16.

Row property	Spearman	95% CI
Representation degree	+0.005	[-0.009,+0.019]
Input degree	+0.029	[-0.007,+0.071]
Anti-hub score	-0.029	[-0.070,+0.008]
Mean-distance centrality	+0.026	[-0.024,+0.075]
Representation label disagreement	-0.041	[-0.109,+0.031]
Representation same-label support	+0.016	[-0.045,+0.058]
Query coverage	+0.104	[+0.053,+0.165]

Query coverage has a clearer association: 0.104, with CI $[0.053, 0.165]$ and positive correlations on all eight influence datasets. Bootstrap-based Holm and Benjamini–Hochberg adjustments both give 0.0015, while exact Holm adjustment gives 0.0586 across the fifteen estimable comparisons

(Table 16). We therefore treat this pattern as exploratory evidence for query-aware selection. The direct query-coverage attack on the separate eight-dataset test panel adds 1.38 pp accuracy loss and 0.0205 prediction flip at 5% (Table 6); both exact Holm-adjusted p-values are 0.0469. These interventions establish effects on those two outcomes, while the influence correlations describe associations on a different dataset panel.

Table 6: Repeated anti-hub selections and query-aware attacks at 5%. Each estimate is additional damage relative to the stated control, with a 95% CI. Exact p is unadjusted for normalized loss and Holm-adjusted across accuracy, Brier, and flip.

Comparison	Measure	Estimate	95% CI	Exact p
Repeated ties–random	nNLL	+0.0086	[+0.0019,+0.0167]	0.03125
Row permutation–random	nNLL	+0.0086	[+0.0015,+0.0169]	0.0469
Anti-hub–coverage match	nNLL	+0.0082	[+0.0014,+0.0167]	0.0195
Anti-hub–coverage match	Accuracy (pp)	+0.91	[+0.35,+1.61]	0.0117
Query coverage–random	Accuracy (pp)	+1.38	[+0.27,+2.79]	0.0469
Query coverage–random	Flip rate	+0.0205	[+0.0057,+0.0381]	0.0469

5.5 Alternative explanations and robustness

Matching random controls on both original class and query coverage leaves 0.0082 additional normalized loss for anti-hubs at 5%, with CI $[0.0014, 0.0167]$ (Table 6). Additional accuracy loss is 0.91 pp, with CI $[0.35, 1.61]$ and exact Holm-adjusted $p = 0.0117$. All 720 control sets have distinct rows and complete matches; the mean coverage-distribution gap is below 10^{-5} (Table 7). The adopted query-coverage matching thus leaves additional anti-hub damage on both normalized loss and accuracy.

Table 7: Query coverage of anti-hub sets and their matched controls. Coverage gap averages absolute differences after sorting each set's row coverages. All values are averaged over settings. All 720 controls contain distinct rows and complete matches.

Budget	Control sets	Anti-hub coverage	Control coverage	Coverage gap
25 labels	360	0.00292	0.00293	7.67×10^{-6}
5%	360	0.00317	0.00318	4.75×10^{-6}

Across settings, a mean 52.6% of selected rows at 25 labels and 45.8% at 5% lie in class-specific boundary ties. A boundary tie is a group of equal-score rows at the last selected score within a class. Repeated choices among these rows and permutations of the context row order give additional normalized loss of 0.0086 and 0.0086, respectively, at 5% (Table 6). Both CIs have positive lower bounds. The permutation experiment restores the original row order before prediction, isolating the selector's response to row ordering. Together, these comparisons show that anti-hub damage persists under coverage matching and alternative ways to resolve equal-score candidates.

6 Conclusion

Context-label attacks depend on the row-selection rule, attack budget, and prediction measure. Hub budget curves show how total corruption damage and additional selection damage evolve separately (Figure 1). Reverse degree and mean distance yield different effects, including stronger accuracy damage at larger budgets (Table 1). Input anti-hubs add accuracy loss at both budget rules, and query coverage supplies another effective selector (Tables 2 and 6). The influence comparisons connect query proximity with single-label effects while distinguishing exploratory associations from multi-label attacks. Future work can study interactions among changed labels, query-aware selection at additional budgets, and the transfer of geometric selectors across datasets and model representations.

AI USE STATEMENT

Generative AI tools assisted with experimental implementation, figures, literature retrieval, and language editing. The authors verified the experimental procedures, numerical results, and citations, and take responsibility for the paper.

ETHICS STATEMENT

This work studies inference-time label corruption on public tabular classification benchmarks. It involves no human subjects, personal identifiers, or deployed systems. The attacker's access and the permitted label changes are specified in Section 3.1.

REPRODUCIBILITY STATEMENT

The experiments use fixed model weights, shared data splits, and class-matched random controls. Section 3 defines the selectors and statistical measures. Section 4 describes the experiment design, and the appendices provide dataset details and complete numerical results.

REFERENCES

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, 2012.

Marin Biloš, James T. Wilson, Anderson Schneider, and Yuriy Nevmyvaka. A mechanistic study of tabular foundation models. *arXiv preprint arXiv:2605.21288*, 2026.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, 2020.

Mohamed Djilani, Thibault Simonetto, Karim Tit, Florian Tambon, Salah Ghamizi, Maxime Cordy, and Mike Papadakis. On the robustness of tabular foundation models: Test-time attacks and in-context defenses. *arXiv preprint arXiv:2506.02978*, 2025.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, 2023. doi: 10.1145/3605764.3623985.

Pengfei He, Han Xu, Yue Xing, Hui Liu, Makoto Yamada, and Jiliang Tang. Data poisoning for in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. doi: 10.18653/v1/2025.findings-naacl.91.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 2017.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.225.

Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0VvD1PmNzM.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Milos Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 2010.

Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic. The role of hubness in clustering high-dimensional data. In *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, PAKDD'11, 2011.

Xiyuan Zhang, Danielle C. Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W. Mahoney, Cuixiong Hu, Huzefa Rangwala, George Karypis, and Bernie Wang. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=t8YRsWY6HM.

A Models and additional attack results

We use TabPFNv2-12L, TabICLv2, and Mitra 1.1 with fixed classification weights. The representation-based selectors use context-row vectors from the ninth layer of TabPFNv2-12L, the eleventh context-processing layer of TabICLv2, and the twelfth layer of Mitra, as described in Section 3.2.

Figure 3 summarizes the attack setting, and Table 8 lists the information available to each selector.

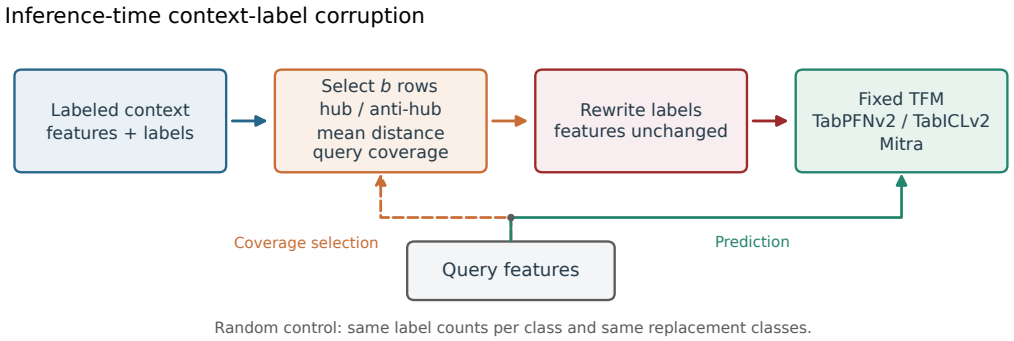

Figure 3: Context-label attacks. A selector chooses b rows, and the attacker changes their labels before the model predicts the queries. The matched-random control changes the same number of labels from each original class to the same replacement class.

Table 9 reports all six comparisons between an attacked and an original context. Table 3 in the main text separates the 5% anti-hub results by model. Each interval in that table summarizes the eight datasets for one model; the aggregate comparisons average all three models before estimating uncertainty. Every input hub–anti-hub pair uses the same context, queries, budget, original-class counts, and replacement mapping across the 240 evaluated settings.

Table 8: Information used by each selector and control. Context features and labels are (x_j, y_j); u_i denotes query features. Model representations are row vectors produced by the model. Query labels are unavailable to every attack.

Selector	Context (x_j, y_j)	Query u_i	Representations
Representation hub	yes	no	yes
Mean distance	yes	no	yes
Input hub / anti-hub	yes	no	no
Query coverage / matching	yes	yes	no
Matched random	yes	no	no

Table 9: Normalized-loss changes relative to the original context for all six clean-to-attack comparisons. Intervals are 95% CIs. Positive/negative counts refer to dataset mean effects; p is the unadjusted exact sign-flip value.

Attack	Budget	Estimate	95% CI	$+/-$ datasets	Exact p
Matched random	25 labels	+0.0073	[-0.0067,+0.0196]	6/2	0.1641
Input hub	25 labels	+0.0098	[-0.0011,+0.0205]	6/2	0.0703
Input anti-hub	25 labels	+0.0104	[-0.0047,+0.0240]	7/1	0.0977
Input anti-hub	5%	+0.0217	[+0.0039,+0.0367]	7/1	0.03125
Matched random	5%	+0.0139	[-0.0037,+0.0285]	7/1	0.0781
Rep-RkNN hub	25 labels	+0.0074	[-0.0071,+0.0201]	6/2	0.1719

B Dataset details and complete attack comparisons

Table 10 describes the eight-dataset panel. Table 11 lists the broader twenty-four-dataset panel, which includes those eight datasets. Five-fold stratified group splitting divides rows into five approximately class-balanced groups. For each random seed, the first group supplies queries and the other four supply context. Identical feature rows are grouped together before splitting. MiceProtein groups all measurements from one biological mouse together, using the mouse identifier before its measurement suffix.

For a percentage budget, the number of changed labels is $b = \lceil \rho n \rceil$, where ρ is the attack fraction and n is the context size. Hub curves use $\rho \in \{0.005, 0.01, 0.05, 0.15\}$; the anti-hub and query-aware comparisons use 0.05, and mean distance uses 0.05 and 0.15. A fixed 25-label attack is included in the small-budget panel when $25/n \leq 0.10$. Table 4 in the main text reports the actual selected-set sizes and overlap. For Gower distance, a pair with no comparable features receives distance one; missing categorical values count as a category.

Table 10: Details of the eight-dataset panel. OpenML is the dataset identifier in the public data repository; C is the number of classes. Numeric and categorical columns are counted separately. Context, query, and budget ranges cover the five splits.

Dataset	OpenML	C	Numeric	Categorical	Context	Queries	5% labels
Fitness_Club	46927	2	3	3	1200	300	60
Is-this-a-good-customer	46938	2	5	8	1378	345	69
MIC	46980	8	17	94	1359	340	68
analcatdata_authorship	458	4	70	0	672–673	168–169	34
anneal	46906	5	6	25	719	179	36
cmc	23	3	2	7	1177–1180	293–296	59
kc2	1063	2	21	0	417–418	104–105	21
maternal_health_risk	46941	3	6	0	811	203	41

Table 11: Details of the broader twenty-four-dataset panel. Target names the prediction label, and C is the number of classes. The final column indicates whether changing 25 labels affects at most 10% of the context. The corresponding comparison uses 22 datasets.

Dataset	OpenML	Target	C	Num.	Cat.	Context	$25/n \leq 10\%$
LED-display-domain-7digit	40496	label	10	7	0	393–394	yes
MiceProtein	40966	class	8	77	0	855–870	yes
balance-scale	11	label	3	4	0	500	yes
banknote-authentication	1462	label	2	4	0	1097	yes
blood-transfusion-service-center	1464	label	2	4	0	598–599	yes
breast-w	15	label	2	9	0	559–560	yes
breast_cancer	sklearn	sklearn target	2	30	0	455	yes
car	40975	label	4	0	6	1382–1383	yes
climate-model-simulation-crashes	40994	label	2	18	0	432	yes
cylinder-bands	6332	label	2	18	17	432	yes
dresses-sales	23381	label	2	1	11	400	yes
eucalyptus	188	label	5	14	5	588–589	yes
iris	sklearn	sklearn target	3	4	0	120	no
mfeat-pixel	40979	label	10	240	0	1600	yes
steel-plates-fault	40982	label	7	27	0	1552–1553	yes
wine	sklearn	sklearn target	3	13	0	142	no
Fitness_Club	46927	label	2	3	3	1200	yes
Is-this-a-good-customer	46938	label	2	5	8	1378	yes
MIC	46980	label	8	17	94	1359	yes
analcatdata_authorship	458	label	4	70	0	672–673	yes
anneal	46906	label	5	6	25	719	yes
cmc	23	label	3	2	7	1177–1180	yes
kc2	1063	label	2	21	0	417–418	yes
maternal_health_risk	46941	label	3	6	0	811	yes

Table 12 gives all 36 aggregate comparisons of anti-hub, hub, mean-distance, and matched-random attacks. The 95% intervals summarize uncertainty in additional damage; the 90% intervals support the equivalence comparisons described in Section 3.4.

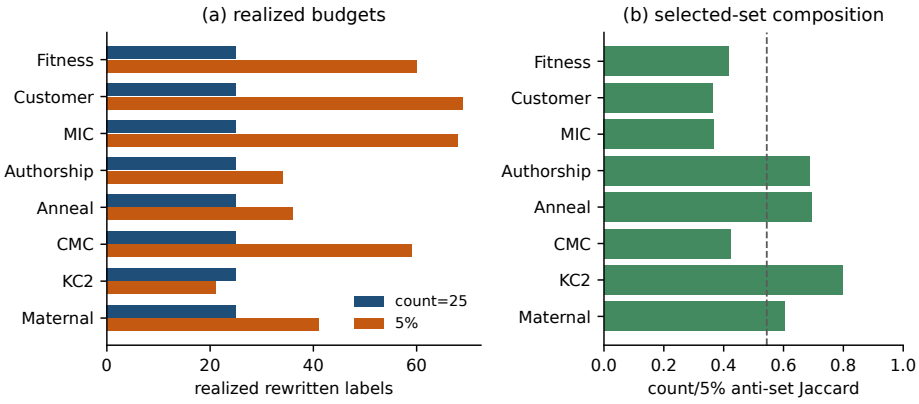

Figure 4: Attack size and selected-set overlap under fixed and percentage budgets. The right panel reports the Jaccard index; the dashed line marks its mean across eight datasets.

C HUB ATTACKS ACROSS COUNT AND PERCENTAGE BUDGETS

The count curves hold dataset membership fixed across all five budgets. Every included context has at least 500 rows, so changing 50 labels affects at most 10% of its labels. The test-panel curves exclude kc2 at every count budget and retain the other seven datasets. The broader-panel curves retain Fitness_Club, Is-this-a-good-customer, MIC, MiceProtein, analcatdata_authorship, anneal, balance-scale, banknote-authentication, blood-transfusion-service-center, breast-w, car, cmc, eucalyptus, maternal_health_risk, mfeat-pixel, and steel-plates-fault. Holding membership fixed lets changes along each curve reflect the budget on the same datasets. The separate 25-label comparisons use all

Table 12: All 36 model-averaged attack comparisons. Accuracy is additional accuracy drop in percentage points. Each estimate subtracts the stated control. Comparisons use eight datasets unless marked as the 22-dataset subset of the broader panel. Both 95% and 90% CIs are shown.

Comparison	Budget	Measure	Estimate	95% CI	90% CI
Anti-hub–random	25 labels	nNLL	+0.0030	[-0.0000,+0.0065]	[+0.0004,+0.0059]
Anti-hub–random	25 labels	Accuracy (pp)	+0.64	[+0.11,+1.26]	[+0.16,+1.15]
Anti-hub–random	25 labels	Brier	+0.0054	[+0.0017,+0.0105]	[+0.0020,+0.0095]
Anti-hub–random	25 labels	Flip rate	+0.0100	[+0.0043,+0.0159]	[+0.0051,+0.0150]
Anti-hub–random	5%	nNLL	+0.0078	[+0.0014,+0.0158]	[+0.0021,+0.0144]
Anti-hub–random	5%	Accuracy (pp)	+1.00	[+0.43,+1.65]	[+0.51,+1.54]
Anti-hub–random	5%	Brier	+0.0103	[+0.0043,+0.0169]	[+0.0051,+0.0158]
Anti-hub–random	5%	Flip rate	+0.0152	[+0.0075,+0.0236]	[+0.0087,+0.0221]
Anti-hub–hub	25 labels	nNLL	+0.0006	[-0.0052,+0.0055]	[-0.0042,+0.0048]
Anti-hub–hub	25 labels	Accuracy (pp)	+0.65	[+0.21,+1.19]	[+0.26,+1.10]
Anti-hub–hub	25 labels	Brier	+0.0020	[-0.0003,+0.0045]	[+0.0001,+0.0040]
Anti-hub–hub	25 labels	Flip rate	+0.0081	[+0.0035,+0.0142]	[+0.0040,+0.0131]
Anti-hub–hub	5%	nNLL	+0.0059	[-0.0011,+0.0148]	[-0.0003,+0.0134]
Anti-hub–hub	5%	Accuracy (pp)	+0.63	[+0.22,+1.13]	[+0.28,+1.03]
Anti-hub–hub	5%	Brier	+0.0047	[+0.0016,+0.0086]	[+0.0020,+0.0079]
Anti-hub–hub	5%	Flip rate	+0.0116	[+0.0027,+0.0225]	[+0.0038,+0.0203]
Mean distance–random	25 labels	nNLL	-0.0032	[-0.0083,+0.0011]	[-0.0074,+0.0006]
Mean distance–random	25 labels	Accuracy (pp)	+0.08	[-0.10,+0.24]	[-0.07,+0.21]
Mean distance–random	25 labels	Brier	+0.0007	[-0.0013,+0.0031]	[-0.0010,+0.0026]
Mean distance–random	25 labels	Flip rate	+0.0015	[-0.0048,+0.0081]	[-0.0040,+0.0072]
Mean distance–random	5%	nNLL	-0.0018	[-0.0115,+0.0058]	[-0.0096,+0.0049]
Mean distance–random	5%	Accuracy (pp)	+0.60	[+0.14,+1.14]	[+0.21,+1.06]
Mean distance–random	5%	Brier	+0.0034	[-0.0013,+0.0078]	[-0.0006,+0.0072]
Mean distance–random	5%	Flip rate	+0.0100	[-0.0025,+0.0239]	[-0.0008,+0.0213]
Mean distance–random	15%	nNLL	+0.0010	[-0.0271,+0.0239]	[-0.0218,+0.0209]
Mean distance–random	15%	Accuracy (pp)	+2.60	[+0.82,+4.48]	[+1.07,+4.18]
Mean distance–random	15%	Brier	+0.0134	[-0.0047,+0.0296]	[-0.0020,+0.0272]
Mean distance–random	15%	Flip rate	+0.0333	[-0.0002,+0.0642]	[+0.0056,+0.0605]
Mean distance–Rep-RkNN	25 labels	nNLL	-0.0016	[-0.0069,+0.0018]	[-0.0061,+0.0017]
Mean distance–Rep-RkNN	25 labels	Accuracy (pp)	+0.03	[-0.24,+0.25]	[-0.19,+0.22]
Mean distance–Rep-RkNN	25 labels	Brier	+0.0014	[-0.0009,+0.0036]	[-0.0006,+0.0033]
Mean distance–Rep-RkNN	25 labels	Flip rate	+0.0013	[-0.0066,+0.0097]	[-0.0058,+0.0089]
Mean distance–random (22 datasets)	25 labels	nNLL	-0.0016	[-0.0065,+0.0030]	[-0.0056,+0.0024]
Mean distance–random (22 datasets)	25 labels	Accuracy (pp)	+0.62	[+0.27,+1.07]	[+0.30,+0.98]
Mean distance–random (22 datasets)	25 labels	Brier	+0.0035	[+0.0012,+0.0060]	[+0.0016,+0.0056]
Mean distance–random (22 datasets)	25 labels	Flip rate	+0.0034	[-0.0036,+0.0113]	[-0.0024,+0.0098]

Table 13: Representation-hub minus matched-random normalized loss at 25 labels. Both 90% CIs lie within ± 0.005. The larger comparison includes all eight datasets in the smaller panel.

Dataset panel	Datasets	Estimate	90% CI
Eight-dataset panel	8	-0.00156	[-0.00457,+0.00080]
22-dataset subset	22	+0.00075	[-0.00183,+0.00383]

eight test datasets and the twenty-two broader-panel datasets with at least 250 context rows. This explains the different dataset counts in Figure 1 and Table 13.

Percentage curves use all datasets in each panel because the relative attack size is controlled directly. The 15% points are shown separately from 0.5%, 1%, and 5%; results are summarized at each budget without pooling across fractions. All curves first average five splits within each model, then the three models within each dataset. Confidence intervals resample datasets, preserving the paired target and random effects. They describe each budget individually.

D Statistical analysis with eight datasets

The exact sign-flip test assigns every combination of plus and minus signs to the eight dataset effects and compares their means with the observed mean. The resulting p-value is the fraction of assignments at least as large as the observed statistic. A studentized sign-flip test uses the mean divided by its estimated standard error. The Wilcoxon signed-rank test instead ranks the absolute

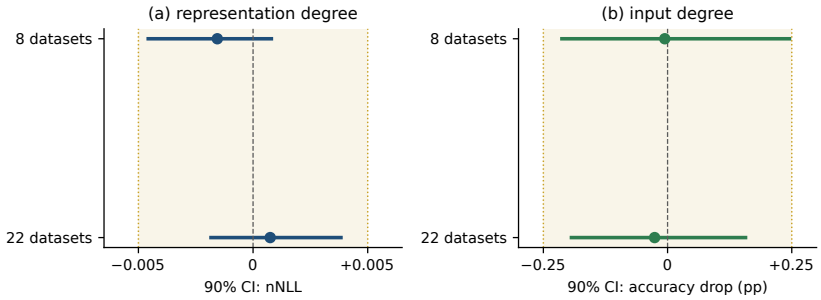

Figure 5: Hub attacks relative to matched random at 25 labels. Points average models and datasets, and bars are 90% CIs. Shaded bands mark equivalence margins of 0.005 normalized loss and 0.25 percentage points of accuracy drop. The twenty-two-dataset comparison includes the eight datasets in the smaller panel.

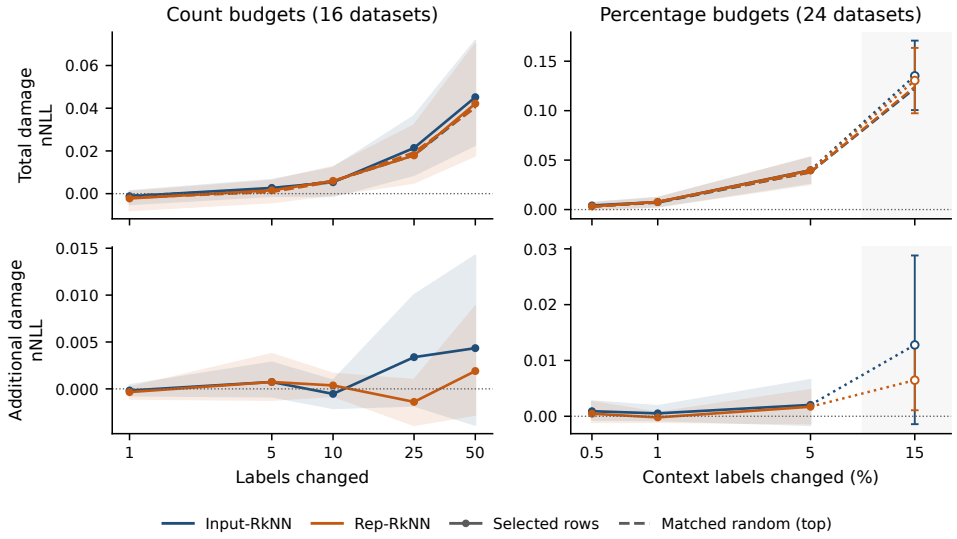

Figure 6: Normalized-loss damage and hub specificity on the broader dataset panel. Count curves use sixteen datasets throughout, and percentage curves use all twenty-four. The plotting conventions and pointwise 95% CIs match Figure 1.

nonzero effects and sums the ranks associated with positive effects. These analyses use the same eight dataset effects as the bootstrap intervals.

Table 14 reports these tests for the original attack comparisons. Holm adjustment is applied separately to accuracy, Brier, and flip within each contrast and budget. For each group, it orders the three p-values and adjusts them for the number of comparisons remaining. The normalized-loss test is reported separately. The equivalence comparisons use margins of 0.005 normalized loss and 0.25 pp accuracy drop. These correspond to 0.5% of the normalized-loss reference for uniform predictions and one quarter of a percentage point of accuracy.

E Repeated row selection and query-aware attacks

We repeat the anti-hub selector with five alternative random choices among equal-score rows and with five permutations of context row order. After a permutation, selected rows are mapped back to their original positions before labels are changed. The model receives the original row order, so this comparison measures the selection rule's dependence on ordering. At 25 labels, 52.6% of selected rows lie in a class-specific boundary tie; at 5%, the fraction is 45.8%. A boundary tie

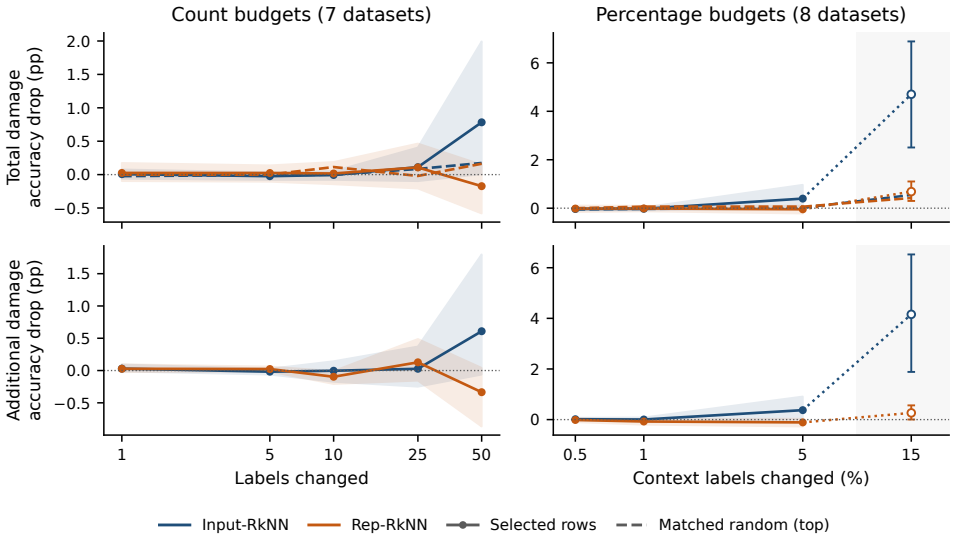

Figure 7: Accuracy damage and hub specificity on the test panel, in percentage points. The count and percentage curves use seven and eight datasets, respectively. The plotting conventions and pointwise 95% CIs match Figure 1.

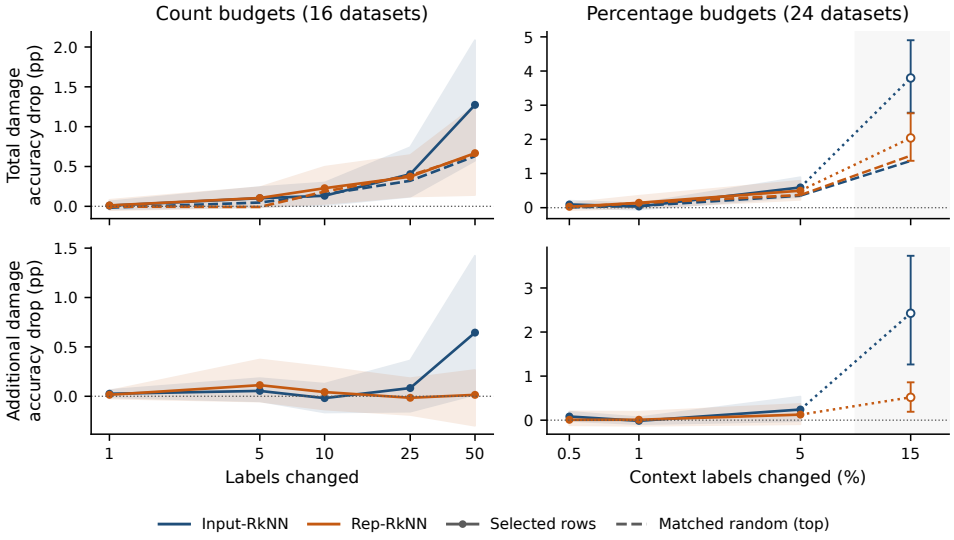

Figure 8: Accuracy damage and hub specificity on the broader dataset panel, in percentage points. Count curves use sixteen datasets and percentage curves use twenty-four. The plotting conventions and pointwise 95% CIs match Figure 1.

contains equal-score candidates at the final selected score. The corresponding fractions with zero reverse-neighbor degree are 46.3% and 36.7%.

The query-coverage attack uses the largest fractional coverage scores within each source class. Coverage-matched controls pair each anti-hub row with an unused row of the same class whose coverage is closest. Targets are processed in random order, and random choices resolve equally close candidates. All 720 control sets contain distinct rows and have complete matches. Their coverage distributions are summarized in Table 7 in the main text. Table 15 reports all 40 comparisons at both budgets, including direct comparisons between query-coverage and anti-hub selection.

Table 14: Exact tests on the eight dataset effects. SF denotes sign-flip, TSF studentized sign-flip, and W Wilcoxon signed-rank. Holm adjustment groups accuracy, Brier, and flip within each comparison and budget; normalized loss is tested separately.

Comparison	Budget	Measure	+/−	SF	TSF	W	Holm
Anti-hub–hub	25 labels	Accuracy (pp)	7/1	0.0156	0.0117	0.0117	0.03125
Anti-hub–hub	25 labels	Brier	7/1	0.0859	0.0859	0.0742	0.0859
Anti-hub–hub	25 labels	Flip rate	8/0	0.0039	0.0039	0.0039	0.0117
Anti-hub–hub	25 labels	nNLL	4/4	0.4258	0.4258	0.3203	—
Anti-hub–hub	5%	Accuracy (pp)	7/1	0.0078	0.0078	0.0078	0.0234
Anti-hub–hub	5%	Brier	6/2	0.0156	0.0156	0.0195	0.03125
Anti-hub–hub	5%	Flip rate	7/1	0.0234	0.0234	0.0273	0.03125
Anti-hub–hub	5%	nNLL	5/3	0.1133	0.1133	0.1914	—
Anti-hub–random	25 labels	Accuracy (pp)	6/2	0.0273	0.0273	0.0391	0.0273
Anti-hub–random	25 labels	Brier	8/0	0.0039	0.0039	0.0039	0.0117
Anti-hub–random	25 labels	Flip rate	8/0	0.0039	0.0039	0.0039	0.0117
Anti-hub–random	25 labels	nNLL	5/3	0.0703	0.0703	0.0742	—
Anti-hub–random	5%	Accuracy (pp)	8/0	0.0039	0.0039	0.0039	0.0117
Anti-hub–random	5%	Brier	7/1	0.0078	0.0078	0.0078	0.0117
Anti-hub–random	5%	Flip rate	8/0	0.0039	0.0039	0.0039	0.0117
Anti-hub–random	5%	nNLL	6/2	0.0273	0.0273	0.0273	—
Mean distance–random	25 labels	Accuracy (pp)	6/2	0.2031	0.2031	0.1562	0.6094
Mean distance–random	25 labels	Brier	4/4	0.3359	0.3359	0.4219	0.6719
Mean distance–random	25 labels	Flip rate	6/2	0.3438	0.3438	0.2734	0.6719
Mean distance–random	25 labels	nNLL	3/5	0.8086	0.8086	0.7266	—
Mean distance–random	5%	Accuracy (pp)	6/2	0.0391	0.0391	0.0391	0.1172
Mean distance–random	5%	Brier	6/2	0.1016	0.1016	0.1250	0.2031
Mean distance–random	5%	Flip rate	5/3	0.1016	0.1016	0.1250	0.2031
Mean distance–random	5%	nNLL	4/4	0.6367	0.6367	0.5781	—
Mean distance–Rep-RkNN	25 labels	Accuracy (pp)	6/2	0.4297	0.4297	0.2734	0.7812
Mean distance–Rep-RkNN	25 labels	Brier	7/1	0.1406	0.1406	0.0742	0.4219
Mean distance–Rep-RkNN	25 labels	Flip rate	6/2	0.3906	0.3906	0.2734	0.7812
Mean distance–Rep-RkNN	25 labels	nNLL	5/3	0.5977	0.5977	0.3711	—
Mean distance–random	15%	nNLL	6/2	0.4766	0.4766	0.3203	—
Mean distance–random	15%	Accuracy (pp)	6/2	0.0156	0.0156	0.0195	0.0469
Mean distance–random	15%	Brier	6/2	0.1016	0.1016	0.0977	0.1016
Mean distance–random	15%	Flip rate	6/2	0.0469	0.0469	0.0547	0.0938

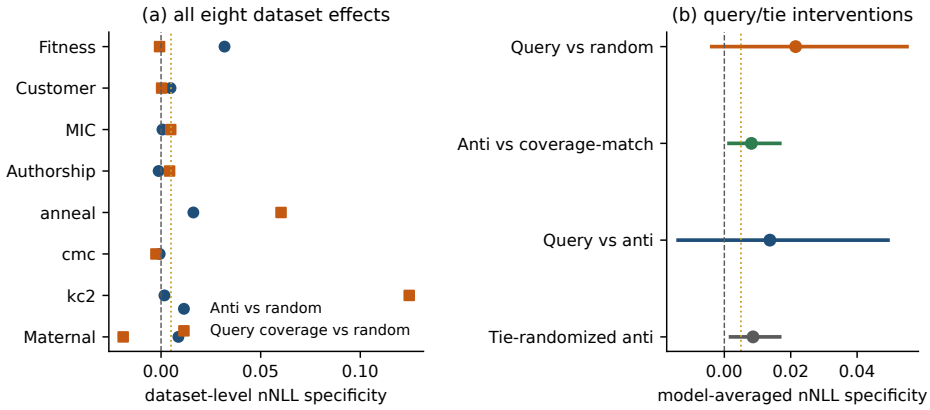

Figure 9: Query-aware attacks and repeated anti-hub selections at 5%. Left: the individual dataset effects on normalized loss. Right: model-averaged additional loss with 95% CIs. The gold line marks the reference effect size of 0.005.

F Row properties and influence correlations

Mean influence $I_{\text{mean}}(j)$ is defined in (4). We also use $I_{\max}(j) = \max_{c \neq y_j}(L_{j\to c} - L_0)$, the largest loss increase over replacement labels, and $I_{\text{map}}(j) = L_{j\to\pi(y_j)} - L_0$, the increase under the same class mapping used in the attack. The influence study uses binary nearest-neighbor membership for query coverage. Each query contributes a set of ten context neighbors, with the selected members of boundary ties retained from that neighbor search. The subsequent attack and matching experiments use the fractional convention in Section 3.2.

Table 15: All 40 comparisons of repeated row selections, query-coverage attacks, and coverage-matched controls. Accuracy is in percentage points. Exact p is the unadjusted sign-flip value; Holm adjusts accuracy, Brier, and flip jointly within each comparison and budget.

Comparison	Budget	Measure	Estimate	95% CI	Exact p	Holm
Repeated ties–random	25 labels	nNLL	+0.0038	[+0.0010,+0.0071]	0.0234	—
Repeated ties–random	25 labels	Accuracy (pp)	+0.58	[+0.11,+1.12]	0.0117	0.0117
Repeated ties–random	25 labels	Brier	+0.0060	[+0.0019,+0.0110]	0.0039	0.0117
Repeated ties–random	25 labels	Flip rate	+0.0103	[+0.0045,+0.0163]	0.0039	0.0117
Repeated ties–random	5%	nNLL	+0.0086	[+0.0019,+0.0167]	0.03125	—
Repeated ties–random	5%	Accuracy (pp)	+1.03	[+0.47,+1.67]	0.0039	0.0117
Repeated ties–random	5%	Brier	+0.0108	[+0.0046,+0.0181]	0.0078	0.0117
Repeated ties–random	5%	Flip rate	+0.0154	[+0.0071,+0.0239]	0.0039	0.0117
Row permutation–random	25 labels	nNLL	+0.0036	[+0.0006,+0.0071]	0.0273	—
Row permutation–random	25 labels	Accuracy (pp)	+0.63	[+0.15,+1.20]	0.0117	0.0156
Row permutation–random	25 labels	Brier	+0.0059	[+0.0019,+0.0110]	0.0039	0.0117
Row permutation–random	25 labels	Flip rate	+0.0100	[+0.0045,+0.0157]	0.0078	0.0156
Row permutation–random	5%	nNLL	+0.0086	[+0.0015,+0.0169]	0.0469	—
Row permutation–random	5%	Accuracy (pp)	+0.98	[+0.45,+1.57]	0.0039	0.0117
Row permutation–random	5%	Brier	+0.0104	[+0.0044,+0.0172]	0.0078	0.0117
Row permutation–random	5%	Flip rate	+0.0153	[+0.0071,+0.0238]	0.0039	0.0117
Query coverage–random	25 labels	nNLL	+0.0222	[-0.0041,+0.0620]	0.1602	—
Query coverage–random	25 labels	Accuracy (pp)	+1.02	[-0.06,+2.45]	0.1602	0.2891
Query coverage–random	25 labels	Brier	+0.0132	[-0.0063,+0.0381]	0.1445	0.2891
Query coverage–random	25 labels	Flip rate	+0.0186	[+0.0034,+0.0384]	0.0547	0.1641
Query coverage–random	5%	nNLL	+0.0215	[-0.0038,+0.0550]	0.1523	—
Query coverage–random	5%	Accuracy (pp)	+1.38	[+0.27,+2.79]	0.0234	0.0469
Query coverage–random	5%	Brier	+0.0159	[-0.0046,+0.0385]	0.1367	0.1367
Query coverage–random	5%	Flip rate	+0.0205	[+0.0057,+0.0381]	0.0156	0.0469
Anti-hub–coverage match	25 labels	nNLL	+0.0015	[-0.0067,+0.0080]	0.3164	—
Anti-hub–coverage match	25 labels	Accuracy (pp)	+0.63	[+0.14,+1.22]	0.0117	0.0234
Anti-hub–coverage match	25 labels	Brier	+0.0045	[+0.0004,+0.0103]	0.0352	0.0352
Anti-hub–coverage match	25 labels	Flip rate	+0.0096	[+0.0041,+0.0156]	0.0039	0.0117
Anti-hub–coverage match	5%	nNLL	+0.0082	[+0.0014,+0.0167]	0.0195	—
Anti-hub–coverage match	5%	Accuracy (pp)	+0.91	[+0.35,+1.61]	0.0039	0.0117
Anti-hub–coverage match	5%	Brier	+0.0104	[+0.0042,+0.0176]	0.0078	0.0117
Anti-hub–coverage match	5%	Flip rate	+0.0153	[+0.0080,+0.0230]	0.0039	0.0117
Query coverage–anti-hub	25 labels	nNLL	+0.0191	[-0.0052,+0.0577]	0.2031	—
Query coverage–anti-hub	25 labels	Accuracy (pp)	+0.38	[-0.34,+1.33]	0.3242	0.6484
Query coverage–anti-hub	25 labels	Brier	+0.0078	[-0.0096,+0.0310]	0.3750	0.6484
Query coverage–anti-hub	25 labels	Flip rate	+0.0086	[-0.0022,+0.0236]	0.1719	0.5156
Query coverage–anti-hub	5%	nNLL	+0.0137	[-0.0139,+0.0493]	0.2773	—
Query coverage–anti-hub	5%	Accuracy (pp)	+0.38	[-0.27,+1.22]	0.2500	0.6445
Query coverage–anti-hub	5%	Brier	+0.0057	[-0.0154,+0.0286]	0.3438	0.6445
Query coverage–anti-hub	5%	Flip rate	+0.0053	[-0.0059,+0.0171]	0.2148	0.6445

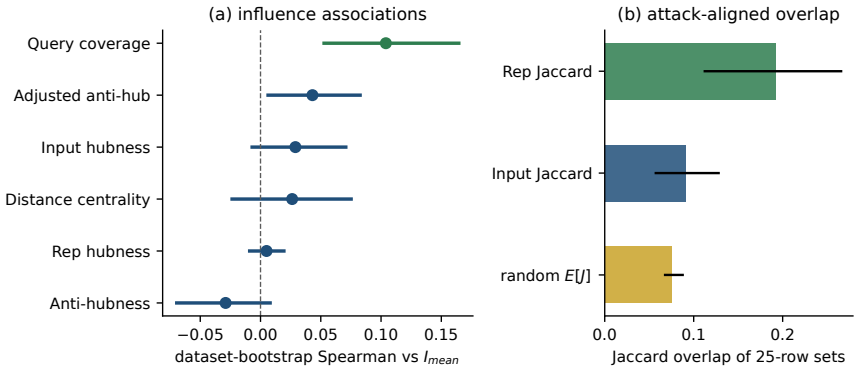

Figure 10: Exploratory row-influence correlations and selected-set overlap. The left panel reports rank correlations with mean influence. The right panel compares mean-distance and hub sets within sampled influence rows, alongside the random-overlap reference.

The row properties in Table 16 have the following definitions. Input and representation degree are the incoming weights h_j in the corresponding graphs. Anti-hub score is the negative input degree, and mean-distance centrality is the negative average representation distance to ten neighbors. Same-label support is the share of neighbor weight assigned to rows with the same class. Label disagreement is the share assigned to different classes, with each weight further multiplied by $\exp(-d/T)$, where d

is the neighbor distance and T is the median positive neighbor distance. Group size counts rows in the same feature-duplicate group, or the same mouse for MiceProtein; the analysis uses the logarithm of one plus that count. Class rarity is the negative logarithm of the original class's context frequency. Missingness is the fraction of numeric features without an observed value. The outlier score is the largest absolute transformed numeric feature value in a row.

Table 16: All 16 exploratory influence correlations. The outcome is I_{mean} unless another influence measure is named. Side 1/2 denotes a one-/two-sided test; the dataset column gives the number of estimable correlations. Holm and Benjamini–Hochberg (BH) adjust the 15 comparisons with at least four datasets. The query-coverage exact Holm p-value is 0.0586.

Row property and influence	Side	Datasets	Spearman	95% CI	Holm p	BH q
Adjusted anti-hub, I_{mean}	1	8	+0.043	[+0.006,+0.083]	0.1764	0.0945
Anti-hub score, I_{map}	1	8	-0.028	[-0.068,+0.008]	1.0000	0.9383
Anti-hub score, I_{mean}	1	8	-0.029	[-0.070,+0.008]	1.0000	0.9383
Anti-hub score, I_{max}	1	8	-0.029	[-0.069,+0.008]	1.0000	0.9383
Class rarity	2	7	-0.060	[-0.183,+0.069]	1.0000	0.5879
Input degree	1	8	+0.029	[-0.007,+0.071]	0.8436	0.3245
Representation degree	1	8	+0.005	[-0.009,+0.019]	1.0000	0.4764
Log group size	2	4	+0.051	[-0.075,+0.189]	1.0000	0.6359
Missingness	2	2	-0.026	[-0.046,-0.007]	—	—
Outlier score	2	8	-0.013	[-0.069,+0.044]	1.0000	0.9026
Distance centrality, I_{map}	1	8	+0.022	[-0.026,+0.066]	1.0000	0.4764
Distance centrality, I_{mean}	1	8	+0.026	[-0.024,+0.075]	1.0000	0.4764
Distance centrality, I_{max}	1	8	+0.021	[-0.026,+0.065]	1.0000	0.4764
Representation label disagreement	1	8	-0.041	[-0.109,+0.031]	1.0000	0.9383
Query coverage	1	8	+0.104	[+0.053,+0.165]	0.0015	0.0015
Representation same-label support	1	8	+0.016	[-0.045,+0.058]	1.0000	0.4764

Table 17 compares the rankings and selected sets of the centrality measures within the sampled candidate rows from the influence study. Spearman correlation measures rank agreement, and the Jaccard index measures selected-set overlap as defined in Section 5.3. The random reference is the expected overlap of random sets with the same size and candidate pool.

Table 17: Rank agreement and selected-set overlap among sampled candidate rows on the eight influence datasets. Within each sample, the 25 highest-degree rows are compared with the 25 rows of smallest mean distance. Intervals are dataset-bootstrap 95% CIs.

Comparison	Estimate	95% CI
Representation degree vs. distance centrality, Spearman	$+0.550$	$[0.458, 0.622]$
Input degree vs. distance centrality, Spearman	$+0.212$	$[0.143, 0.289]$
Representation hub vs. mean-distance sets, Jaccard	0.1925	$[0.112, 0.266]$
Input hub vs. mean-distance sets, Jaccard	0.0906	$[0.057, 0.128]$
Expected random-set Jaccard	0.0748	$[0.068, 0.088]$
Representation-set Jaccard minus random expectation	0.118	$[0.037, 0.189]$

The adjusted anti-hub association correlates residuals after linearly accounting for input and representation label disagreement, representation same-label support, log group size, class rarity, query coverage, and missingness. Residuals are the parts of the score and influence left after subtracting their fitted linear predictions from those properties. This is an exploratory association measure. The analysis contains sixteen comparisons. Fifteen have estimable correlations on at least four datasets and enter Holm and Benjamini–Hochberg adjustments; the latter controls the expected fraction of false discoveries among reported associations. Missingness is summarized on two datasets. Degree, anti-hub, centrality, neighbor-label, and coverage tests use a positive one-sided alternative. Outlier, group size, and rarity tests use two-sided alternatives.

Query coverage meets both bootstrap-based adjustments: Holm $p = 0.0015$ and Benjamini–Hochberg $q = 0.0015$. All eight dataset correlations are positive. The exact sign-flip p-value is 0.00390625, and applying Holm across the fifteen comparisons gives 0.05859375.