WHEN DOES SYNERGY HELP ACTIVE FEATURE ACQUISITION? A PID-BASED STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Active feature acquisition (AFA) sequentially selects informative features under budget constraints. However, existing policies rarely distinguish whether information contributed by interacting features is redundant, unique, or synergistic. We introduce SynAFA, a state-dependent AFA policy that combines pairwise joint information and conditional information, with Partial Information Decomposition (PID) characterizing its information structure. Across five tabular datasets and MNIST-loop, performance is heterogeneous. SynAFA shows its strongest gains at low budgets on PhysioNet, where synergistic and redundant feature pairs are supported by permutation tests, but its advantage diminishes as budgets increase and does not depend on pair proposals. SynAFA performs significantly worse than nearly all baselines on MiniBooNE, and than CAE on MNIST-loop. Controlled synthetic experiments further show that, within budget, SynAFA's advantage rises as joint information becomes more synergy-dominated, including when total joint information is held approximately constant. Further analyses show that the budget-dependent erosion is not resolved by non-greedy local search, which improves a diagnostic set-level objective but leaves predictive performance no better, often significantly worse, and that improvements in this objective are weakly aligned with the fixed classifier's predictive utility. These findings characterize when pairwise synergy can benefit AFA while exposing a persistent challenge in translating local information into effective acquisition objectives.

1 INTRODUCTION

Active feature acquisition (AFA) selects measurements sequentially for each test instance, balancing predictive performance against acquisition cost. Laboratory tests, sensor queries, and survey items are costly, and a measurement's value depends on what is already known and what remains affordable. Existing approaches estimate conditional information (Covert et al., 2023; Gadgil et al., 2024), learn generative surrogates (Ma et al., 2018; Zannone et al., 2019), optimize acquisition policies (Shim et al., 2017), or use acquisition-conditioned oracles (Valancius et al., 2023).

A central difficulty is that two unobserved features may be informative together even when neither is useful alone. Meanwhile, individually informative features may carry redundant information. Partial Information Decomposition (PID) (Williams & Beer, 2010) separates joint information into redundant, unique, and synergistic components. This distinction characterizes information structure, but does not determine whether an acquisition policy can exploit that structure under a budget.

We introduce SynAFA, which combines pairwise joint information and conditional information evaluated at the realized values of observed features, as a probe of when synergy helps rather than a claim of state-of-the-art performance. PID characterizes the information structure underlying its pair scores. The deployed policy uses total joint information, not separately weighted PID atoms. This study mainly addresses two questions: *when does pairwise information with rich synergy help acquisition, and what limits its benefit as budgets grow?*

Our contributions are:

1. **Method:** a state-dependent policy that, at each step, compares single-feature proposals (scored by marginal information or by information conditional on already-observed values) with proposals for a not-yet-observed feature pair (scored by its joint information;

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.

the *pair branch*) and acquires one feature from the best cost-adjusted proposal, with PID characterizing the information structure (§3).

2. **When synergy helps:** controlled synergy–budget sweeps (varying synergy strength and budget), including fixed-information controls and a pair-branch ablation, identify benefits when the corresponding joint acquisition is feasible. On most tested real datasets, the pair branch contributes little (§5).
3. **What limits the benefit as budgets grow:** non-greedy revision does not recover accuracy, and swap-level utility diagnostics and a duplicate-copy construction expose a limitation of aggregating local pairwise scores into a set objective (§6).

2 Related Work and Problem Formulation

Active feature acquisition. Greedy methods estimate the conditional value of the next feature (Covert et al., 2023; Gadgil et al., 2024). Generative methods infer unobserved measurements to evaluate candidate acquisitions (Ma et al., 2018; Zannone et al., 2019). Reinforcement learning (RL) methods optimize sequential acquisition and stopping (Shim et al., 2017; Mnih et al., 2015). Acquisition-conditioned oracles (Valancius et al., 2023) and information templates (Huang et al., 2025) explicitly address jointly informative acquisitions. Static methods such as concrete autoencoders (CAE) instead learn a population-wide subset against a downstream objective (Balın et al., 2019). SynAFA complements these approaches by relating acquisition outcomes to the redundant, unique, and synergistic composition of pairwise information. We compare policies within AFABench (Schütz et al., 2026), and see also Aronsson et al. (2025).

Problem and notation. For features $X_1, \ldots, X_D$, binary target Y, and costs c_j, the observed state is (S, x_S), where S is the acquired set. A policy chooses an unobserved feature or stops to predict, subject to a hard budget B, so $\sum_{j \in S} c_j \leq B$. Information tables are estimated from binarized training features. The test-time policy conditions on observed feature values. Under evaluation, a shared external classifier predicts using the acquired subset. Detailed preprocessing and the hard-budget protocol are in Appendices A.1 and A.9.

Pairwise information and PID. For two sources, PID decomposes their joint mutual information as

$$I(X_j, X_k; Y) = R + U_j + U_k + \mathrm{Syn}, \tag{1}$$

where R is shared redundant information, U_j, U_k are source-unique information, and Syn is synergistic information available only jointly. We use the Williams–Beer $I_{\min}$ definition (Williams & Beer, 2010) (see Appendix A.2). PID has also been used to characterize static feature selection (Wollstadt et al., 2023). Our focus is sequential acquisition with realized-value conditioning: $I(Y; X_j \mid X_k = x_k)$ can vary with the observed value x_k, unlike a population-level pair score.

3 SynAFA

3.1 Offline information quantities

Using the training data, we precompute

$$M_j = I(X_j; Y), \qquad V_{jk} = I(X_j, X_k; Y), \qquad C_{j|k,v} = I(Y; X_j \mid X_k = v). \tag{2}$$

These correspond to `single_value`, `pair_value`, and `cond_single_value`. PID decomposes V_{jk} for characterization and validation. Individual atoms are not directly used in the deployed acquisition score.

The conditional table stores $I(Y; X_j \mid X_k = v)$ separately for each observed value v of X_k, rather than only its average over v:

$$I(Y; X_j \mid X_k) = \sum_v p(X_k = v)\, C_{j|k,v}. \tag{3}$$

Different observed values can therefore produce different candidate scores. Conditioning uses one observed feature at a time; the maximum below does not estimate full-state information $I(Y; X_j \mid X_S = x_S)$.

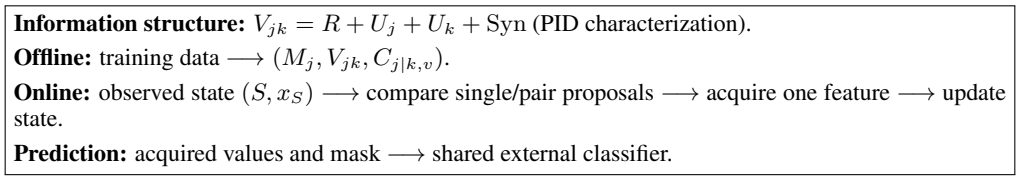

Figure 1: SynAFA in one schematic. PID characterizes joint information; test-time decisions use the three precomputed information tables. Pair proposals initiate one acquisition rather than committing to both features.

3.2 State-dependent acquisition

Given observed state (S, x_S), each unobserved feature receives

$$s_j = \max\left\{M_j, \max_{k \in S} C_{j|k,x_k}\right\}, \tag{4}$$

with $s_j = M_j$ when S is empty. The policy compares single-feature proposals $s_j - \lambda c_j$ with pair proposals $V_{jk} - \lambda(c_j + c_k)$, where $j, k \notin S$. If a pair wins, only its member with larger marginal information is acquired at that step. Its partner can subsequently become valuable through realized-value conditioning, while completion of the pair is not guaranteed. Acquisition stops when no cost-adjusted proposal is positive, subject to the hard-budget evaluation procedure in Appendix A.1.

The deployed pair branch does *not* additionally require the full pair cost to fit within the remaining budget. Although the acquired feature must consider the hard budget, this rule can start a pair that the budget cannot afford to complete. We retain this rule for the primary comparisons and test a feasibility mask as a post hoc mechanism diagnostic (§5.4). The policy is one-step-ahead and does not optimize a complete acquisition sequence.

Why compare single and pair proposals? For two independent fair bits with an XOR target, each bit alone carries no information about the target, but together they determine it (one bit). Single-feature scores would never select either bit, whereas a pair proposal can start the acquisition. Once one bit is observed, the conditional table makes the other informative. Costs and competing features can still prevent the pair from being completed.

3.3 Scaling to high dimensions: an amortized estimator (SynAFA v2)

Exact pairwise estimation needs $O(D^2)$ pair evaluations, each over $O(n)$ training samples, which is feasible for moderate D but expensive at image scale ($D = 784$, yielding $\binom{784}{2} = 306{,}936$ pairs). To scale SynAFA to this setting, SynAFA v2 trains an amortized estimator that maps a pair's empirical frequency statistics to its information values, including V_{jk} from one fixed sample of the whole training set, and $C_{j|k,v}$ from a fixed sample conditioned on $X_k = v$. This computes each pair's, or conditioned variable's, frequency counts at once, rather than estimating them exactly from the full training set. SynAFA v2 is therefore the variant of SynAFA designed for high-dimensional scaling, rather than a different policy. Estimator details and controls are in Appendix A.12.

4 Experimental Setup and Broad Benchmark

Evaluation protocol. We evaluate six datasets: five tabular datasets (CKD, ACTG175, BankMarketing, PhysioNet (Silva et al., 2012), and MiniBooNE), and one imaging dataset, MNIST-loop ($D = 784$), a binary relabeling of MNIST (LeCun et al., 1998). Dataset dimensions and test sizes are in Table A1. Hard budgets for the five tabular datasets follow AFABench's default configuration (Schütz et al., 2026), and MNIST-loop's budgets (5, 10, 20) and the synthetic sweep's budgets (3, 5, 8) were fixed for this paper and not tuned on validation or test performance. Comparisons within each dataset use the same hard budget and the same externally trained masked MLP. This controls the prediction model across policies, but does not guarantee equal generalization across their mask distributions. Real-data costs are uniform, while the controlled sweep uses heterogeneous costs. SynAFA uses $\lambda = 0.01$ in the tabular and synthetic hard-budget experiments, without

dataset-specific tuning. ForMNIST-loop (SynAFA v2), $\lambda = 0.001$ was fixed because pixel features are sparse, and a post hoc sensitivity analysis (Appendix A.14) finds it the best of three values tested.

For paired test instances, $\Delta\text{Acc} = n^{-1}\sum_i[\mathbf{1}(\hat{y}_i^{\text{SynAFA}} = y_i) - \mathbf{1}(\hat{y}_i^{\text{baseline}} = y_i)]$. Accuracy lies below the majority-class rate on PhysioNet and BankMarketing (Table A1), so only paired differences are interpreted. Significance uses a two-sided paired bootstrap with 10,000 resamples. Tests are per comparison and unadjusted for multiplicity. All results use the corrected hard-budget harness (Appendix A.7). With heterogeneous costs, a common budget is a cost ceiling, not a guarantee of equal feature counts.

The eight shared tabular baselines cover oracle-based acquisition (AACO), permutation importance, static selection (CAE), greedy methods (GDFS, DIME), and RL (JAFA and OL with/without an observation mask). EDDI and ODIN are additionally evaluated on the two smallest datasets. Appendix A.7 reports numerical comparisons for PhysioNet and MiniBooNE against all eight baselines, together with the available comparison summaries for the remaining datasets.

Heterogeneous performance. CKD and ACTG175 show mostly non-significant differences across budgets. BankMarketing is mixed and non-monotonic. PhysioNet gives the strongest positive tabular result at the lowest budget, but the advantage diminishes as the budget increases. MiniBooNE is a strong negative case, with significant losses in 22 of 24 budget–baseline comparisons. Thus the broad benchmark motivates a conditional explanation of success, rather than a claim of uniform superiority (Tables A2 and A3).

Table 1: Results of the tabular benchmark: accuracy differences at the lowest and highest tested budgets, against the adaptive AACO and static CAE. Positive values favor SynAFA. $^*p < 0.05$ without multiplicity correction. Endpoints summarize scope, not monotonicity, and all intermediate budgets are in Table A2. CKD ($n = 80$) stars are bootstrap-based and are not significant under an exact sign test (Table A13). Small CKD and ACTG175 test sets ($n = 80$ and 429) limit precision, non-significant differences thus do not establish equivalence.

	Budgets	vs. AACO		vs. CAE	
Dataset	low/high	Low	High	Low	High
CKD	2/7	-0.050^*	$+0.000$	$+0.038$	$+0.025$
ACTG175	3/10	-0.023	-0.009	$+0.016$	$+0.009$
BankMarketing	2/7	$+0.034^*$	-0.068^*	$+0.000$	-0.045^*
PhysioNet	5/15	$+0.011$	-0.038^*	$+0.037^*$	-0.005
MiniBooNE	5/15	-0.034^*	-0.033^*	-0.025^*	-0.044^*

Table 2: PhysioNet: mean accuracy difference against eight baselines and counts of significant wins/ties/losses at each budget ($n = 2{,}400$). Ties denote non-significant differences, not equivalence. Tests are unadjusted across comparisons. Full effects are in Table A3.

Budget	Mean ΔAcc	Wins	Ties	Losses
5	$+0.026$	6	2	0
10	$+0.005$	2	5	1
15	-0.011	1	5	2

5 When Does Synergy Help?

The benchmark raises a question about conditions for success. We first examine whether the strongest positive case contains synergistic pairwise structure supported by permutation tests. As real data doesn't allow us to change this structure, we use controlled sweeps to investigate whether the benefit comes from the information itself or from the acquisition feasibility.

5.1 Real-data validation: PhysioNet

Stratified permutation tests support synergistic and redundant feature structure in PhysioNet. In a scan of the 351 testable pairs (of 378 usable) with Benjamini–Hochberg FDR correction, GCS/PaO_2 was the strongest synergy pair ($\text{Syn} = 0.0127$ bits, raw permutation $p = 0.0005$, the smallest value

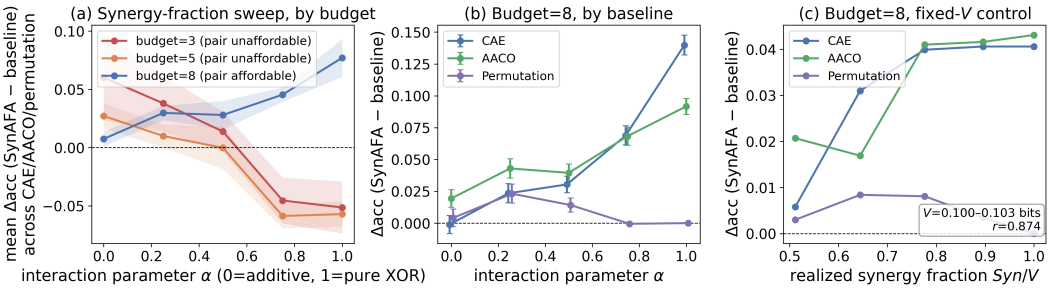

Figure 2: Controlled synergy sweeps. (a) Mean SynAFA advantage by budget (five dataset instances; band: ± 1 SD across instances); the designated pair costs 6, exceeding budgets 3 and 5 but fitting budget 8. (b) Individual baseline comparisons at budget 8, pooled over instances (95% paired-bootstrap intervals). (c) Fixed-V control: advantage versus realized synergy fraction with joint information approximately constant. Full results and calibration are in Appendix A.7.

obtained from 2,000 permutations). The pair survived a discovery/held-out-validation procedure. BUN/creatinine, tested separately as a clinically motivated candidate, supplies a confirmed redundant pair. The scan uses the full cohort, including the 2,400 evaluation patients, while SynAFA never uses its output. Procedures and a separate illustrative example (PaO_2/FiO_2) are in Appendix A.3.

In the held-out check, the scan in the discovery half (6,000 patients per half) selected 51 of 351 pairs after FDR correction. Only these pairs were tested on the validation half, with correction across that confirmatory set. 36 of the 51 (70.6%) replicated, including GCS/PaO_2. This separates discovery from confirmation within that check.

This structure coincides with SynAFA's advantage at low budgets(Table 2). However, removing the pair branch does not change that advantage (Appendix A.13), indicating that the advantage is not coming from pair proposals. The controlled experiments below vary the designated pair's information structure while holding the remaining construction fixed.

5.2 Controlled synergy–budget sweep

We construct a binary dataset with $D = 24$, including a designated pair S1 (cost 1) and S2 (cost 5), and 22 independent background features of decreasing informativeness. The designated pair's contribution to the outcome logit is

$$(1-\alpha)[w_1(2S_1-1)+w_2(2S_2-1)]+\alpha w_{\text{xor}}[2(S_1 \oplus S_2)-1], \tag{5}$$

where $w_1 = w_2 = 0.7$, $w_{\text{xor}} = 1.4$, and $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. The pair changes from additive to pure XOR structure. SynAFA and independently trained AACO, CAE, and permutation baselines use the same pipeline at budgets 3, 5, and 8, with 6,000 test instances per comparison and dataset instance (five independently generated dataset instances per α; Appendix A.8).

Figure 2 (a) and (b) show the budget-dependent trend for the deployed policy. The pair costs 6, so it is unaffordable at budgets 3 and 5 and affordable at budget 8. At budget 8, SynAFA's mean advantage (five dataset instances) rises from $+0.007$ at $\alpha = 0$ to $+0.077$ at $\alpha = 1$ (instance-bootstrap 95% interval $[0.063, 0.089]$), reaching $+0.140$ against CAE and $+0.092$ against AACO, with permutation roughly tied. The descriptive correlation across the five prescribed levels is $r = +0.946$. At budgets 3 and 5, the correlations are -0.974 and -0.945. Each correlation has the same sign in all five dataset instances. Full matrices are in Tables A4 and A8. Removing the pair branch reverses the trend at budget 8 (advantage falls from $+0.006$ to -0.043; Table A22), so the rise is due to pair proposals (Appendix A.13).

The positive trend at budget 8 is not strictly monotonic and is strongest against CAE and AACO, while permutation returns to no advantage at high α. Moreover, α changes information composition without holding total information fixed. Under this symmetric $I_{\min}$ construction, $U_a = U_b = 0$, and Syn/V rises from 0.52 to 1.00 even though neither Syn nor V is monotonic (Table A6). The experiment therefore concerns increasingly synergy-dominated information, not an isolated change to one PID atom.

5.3 FIXED-INFORMATION CONTROL

There is still another question: does the benefit at budget 8 reflect information composition or simply more joint information? To address this, a second sweep recalibrates the additive and XOR weights for target fractions $\mathrm{Syn}/V \in \{0.52, 0.65, 0.78, 0.90, 1.00\}$ while holding V approximately fixed. Realized V is 0.1003–0.1032 bits and realized Syn/V is 0.512–1.000 (Table A10). This construction cannot be reliably calibrated for synergy fraction below about 0.5, so the control covers only the upper half of the range, roughly 0.5 to 1, not the full 0 to 1.

Figure 2 (c) shows that mean advantage rises from $+0.010$ to $+0.028$ with synergy fraction ($r = 0.874$) at budget 8. The largest-fraction effects are $+0.043$ against AACO and $+0.041$ against CAE, while permutation remains nearly no advantage (Table A11). This trend cannot be explained simply by an increase in total pair information. Together, the controls suggest that the benefit comes from synergy-dominated pairwise information within the tested constructions when the joint acquisition is feasible. Ablating the pair branch shows that this benefit is concentrated at high synergy fractions and comes from pair proposals (Appendix A.13).

To check whether this trend replicates across datasets, we pooled five separately generated dataset instances, each with its own trained policies, into 30,000 paired test comparisons. Within each instance, the correlation between synergy fraction and mean advantage is positive, and each instance improves from the lowest fraction to the highest (Table A11). The significance markers in the pooled tables only reflect randomness in which patients are included in the test sample. They do not test whether the trend remains across these dataset instances.

5.4 MECHANISM: CAN THE POLICY COMPLETE THE PAIR?

Looking at what acquisition records explain the tight-budget reversal. At budget 3, increasing α causes SynAFA to abandon cheap S1 (in four of five dataset instances at $\alpha = 1$), because S1 carries no information by itself under pure XOR. At budget 5, it acquires expensive S2 alone (four of five instances at $\alpha = 1$). At budget 8, both features are acquired in 100% of test instances in all dataset instances at $\alpha \geq 0.75$, whereas completion at $\alpha = 0$ is unreliable (0–100%, mean 30%; Table A7). Affordability alone therefore does not guarantee that SynAFA completes the pair.

We repeat the sweep with one change that is masking pair proposals whose combined cost exceeds the remaining budget. The mask does not change the results at budget 8. At budgets 3 and 5 it removes the negative mean relative effects at high synergy fractions. Table 3 gives the $\alpha = 1$ case, and full sweeps are in Tables A5 and A9. This intervention supports a feasibility limitation of the deployed proposal rule as an explanation for the high-synergy losses, rather than saying that synergy is generally harmful under tight budgets. This mask is a post-hoc diagnostic used only here, and primary results retain the deployed rule.

Table 3: Feasibility diagnostic at pure XOR ($\alpha = 1$): mean accuracy advantage over AACO, CAE, and permutation, averaged over five dataset instances. Masking restores positive relative performance when the pair is unaffordable (budgets 3 and 5: the mean advantage is negative before and positive after masking in all five dataset instances). Budget 8 is unchanged.

Budget	Deployed	Feasibility-masked
3	-0.051	$+0.064$
5	-0.057	$+0.063$
8	$+0.077$	$+0.077$

6 WHAT EXPLAINS THE EROSION OF THE BENEFIT?

On real data SynAFA's relative performance worsens as budgets grow (PhysioNet, MNIST-loop), and under uniform costs feasibility alone doesn't explain this. One natural explanation is irreversible greedy selection, in which a feature that scores well at the current step may make later features redundant, and a feature that scores poorly at the current step may only become valuable once another feature is acquired first. Greedy scoring cannot see either possibility in advance, and the choice cannot be reversed once made. We test this explanation by revising acquired sets, then examine whether the diagnostic objective is aligned with prediction.

6.1 MNIST-LOOP STRESS TEST AND BUDGET EROSION

MNIST-loop relabels digits $\{0, 6, 8, 9\}$ against $\{1, 2, 3, 4, 5, 7\}$, providing an imaging task in which spatially separated pixels can carry joint information. SynAFA v2 significantly outperforms AACO at all tested budgets but loses to CAE, with the gap widening from budget 5 to 20 (Table 4).

Table 4: MNIST-loop: accuracy(SynAFA v2)−accuracy(baseline), evaluated with the shared classifier ($n = 12{,}000$). * denotes an unadjusted paired-bootstrap $p < 0.05$.

Baseline	Budget 5	Budget 10	Budget 20
AACO	$+0.127^*$	$+0.149^*$	$+0.070^*$
CAE	-0.023^*	-0.034^*	-0.070^*
Permutation	$+0.020^*$	$+0.001$	-0.018^*

Separate controls show that the hard-budget fallback is never invoked (all 12,000 test instances reach each budget by the policy's own choice), and that the gain over the static ablations is due to the realized value adaptation. Estimator controls indicate a contribution from amortization error that increasing the precompute sample shrinks the accuracy gap from 0.0128 to 0.0005 on a reduced $D = 90$ subset. Under exact scoring at $D = 90$, local search improves accuracy at budgets 5 and 20 but not 10. This comparison changes dimensionality and estimator together and does not show whether the aggregation objective is aligned with predictive accuracy (Appendices A.5 and A.12). Therefore, the erosion on PhysioNet and MNIST-loop motivates testing whether revising irreversible greedy selections recovers predictive utility.

6.2 NON-GREEDY REVISION OF A DIAGNOSTIC OBJECTIVE

For a fixed instance, define

$$T(S) = \sum_{j \in S} \max \left\{ M_j, \max_{k \in S \setminus \{j\}} C_{j|k,x_k} \right\}, \tag{6}$$

where a singleton contributes M_j. $T(S)$ is a diagnostic aggregation of local scores, not true set-level mutual information. SynAFA itself does not define a global set objective, so this search is not an exact non-greedy optimization of the deployed policy. Starting from each instance's acquired set, the corrected search evaluates replacements by fully recomputing $T(S')$, accepting the best swap only if $T(S') > T(S)$. Every accepted move therefore improves the diagnostic objective; all instances converged within 200 rounds. This is a retrospective diagnostic of revisability, not a deployable acquisition rule that refunds previous measurements. Full methodology is in Appendix A.11.

Across nine settings, which are three budgets each for MNIST-loop ($D = 784$, amortized), PhysioNet ($D = 41$, exact), and MiniBooNE ($D = 50$, exact), local search yields no significant accuracy improvement and seven significant decreases (Table 5; p-values in Table A20). Since increasing $T(S)$ does not recover accuracy, we next examine the predictive utility of accepted swaps.

6.3 DO OBJECTIVE IMPROVEMENTS PREDICT UTILITY IMPROVEMENTS?

For each accepted swap, we measure ΔT and ΔU, where U is the fixed external classifier's predicted probability of the true class. A random-swap control uses the same removal and a uniformly random replacement. Although $\Delta T > 0$ by construction, mean ΔU is small in every setting (from -0.018 to $+0.0014$) and negative in six of nine. Pearson correlations range from -0.068 to $+0.107$; 0.419–0.533 of accepted swaps increase utility, with setting-dependent differences from random replacements (Table A21; all results use the full test sets).

What the alignment diagnostic measures. These correlations are conditional on a swap being selected and accepted, so $\Delta T > 0$ throughout. They describe the relationship between improvement magnitudes along this search, not ranking accuracy over all candidate swaps. Moreover, swaps from the same instance are dependent, so the reported confidence intervals use an instance-level cluster bootstrap and the correlations are treated as descriptive.

On MNIST-loop, selected swaps improve utility more often than random replacements (e.g., 0.506 versus 0.400 at budget 5, a 10.6-percentage-point difference), yet their mean ΔU is only $+0.0014$

Table 5: Non-greedy revision under the diagnostic objective, full test sets. Accuracy changes compare revised and original sets; mean utility changes and correlations use accepted swaps. These point estimates summarize Tables A20 and A21; * denotes unadjusted $p < 0.05$ for accuracy.

Dataset	Budget	ΔAcc	Mean ΔU	$\rho(\Delta T, \Delta U)$
MNIST-loop	5	-0.0025	$+0.0014$	$+0.073$
MNIST-loop	10	-0.0213^*	$+0.0007$	$+0.107$
MNIST-loop	20	-0.0322^*	-0.0024	$+0.044$
PhysioNet	5	-0.0371^*	-0.0116	-0.010
PhysioNet	10	-0.0150^*	-0.0022	-0.068
PhysioNet	15	-0.0042	$+0.0004$	-0.048
MiniBooNE	5	-0.0404^*	-0.0180	$+0.095$
MiniBooNE	10	-0.0148^*	-0.0038	$+0.060$
MiniBooNE	15	-0.0065^*	-0.0039	$+0.006$

there and -0.0024 at budget 20 (0.485 versus 0.418). Improvement frequency therefore does not imply a positive average gain, nor do these results establish universal equivalence to random selection. In MiniBooNE at budgets 10 and 15, selected swaps improve utility slightly less often than random ones.

The pattern occurs with both exact and amortized scores. Both ΔU and accuracy depend on the external classifier, and these diagnostics do not establish misalignment independent of the classifier. Missing higher-order interactions remain possible, but the observations directly implicate the alignment of the tested objective rather than higher-order information as the primary cause.

The following construction shows that overcounting can occur without estimation error or higher-order synergy. Let $M \geq 2$ and $X_1 = \cdots = X_M = Z$ be identical copies with $I(Z;Y) = I_0 > 0$, and $S = \{1, \ldots, M\}$. Then

$$T(S) = MI_0, \qquad I(X_S;Y) = I_0, \qquad T(S)/I(X_S;Y) = M. \tag{7}$$

Given any copy $X_k = x_k$ with positive probability, every other copy X_j is forced to equal x_k. X_j therefore becomes deterministic and carries no remaining information about Y, yielding $C_{j|k,x_k} = 0$. Each term in Eq. 6 therefore reduces to $\max(I_0, 0) = I_0$, whereas observing all copies together reveals only the information in Z, namely I_0.

The marginal information floor fails to discount acquired redundancy. By the chain rule, true joint information has zero increments after the first identical copy, while $T(S)$ counts M_j for every copy, which causes a structural failure. It doesn't mean that redundancy explains every bad swap, or that all sets are overestimated by size.

7 Discussion and Limitations

Across the six real datasets and the controlled synthetic family, this paper studies one question through three stages: *when does synergy-rich pairwise information help acquisition, and what limits that benefit as budgets grow?* The broad benchmark in §4 shows that SynAFA's advantage is heterogeneous, motivating a search for the conditions under which it succeeds. §5 located one such condition on PhysioNet, where permutation tests confirm synergistic and redundant pairwise structure. Controlled sweeps show that the advantage rises with increasing synergy fraction when the pair can be acquired within budget, though PhysioNet's own advantage does not depend on the pair branch. §6 then asks why this benefit erodes as budgets grow. Non-greedy revision of the acquired set did not recover accuracy, and accepted revisions were only weakly correlated with true predictive utility, indicating a mismatch between the diagnostic objective and the utility it approximates rather than greedy irreversibility alone. The picture that emerges is conditional, not uniform. Pairwise synergy can help, but only when the acquisition is both structurally present and budget-feasible. Improving a set-level score is not equivalent to improving downstream predictive performance.

We observe three distinct failure modes: unfinishable pairs with different costs under tight budgets (synthetic sweeps), state-independent acquisition when population-level scores dominate (MiniBooNE), and budget erosion with aggregated scores weakly aligned with predictive utility (PhysioNet, MNIST-loop). PhysioNet supports the existence of relevant pairwise structure in real data

via permutation tests, but removing the pair branch does not change its acquisition gains essentially. The synthetic controls more directly connect information composition and feasibility to relative performance. Together, these findings suggest why synergistic/redundant pairs alone are insufficient. Deploying them also requires enough budget to complete the acquisition, an acquisition rule that responds to realized values rather than collapsing to population-level scores, and an aggregation objective that aligns with downstream utility as more features are acquired.

MiniBooNE shows a strong negative case. Its state-dependent ranking collapses to an effectively static acquisition pattern dominated by population-level scores rather than realized conditional values. SynAFA acquires the identical feature subset for all test instances at every tested budget, unlike the other four tabular datasets. Holding the acquired features fixed, retraining the classifier on the real masks substantially narrows the gap to AACO, CAE, and permutation, though most comparisons remain significant losses (Table A16). However, these losses do not imply that loss of adaptivity causes the residual gap. Our acquisition comparisons and objective alignment results therefore concern utility under the evaluation classifier, which can favor some mask distributions over others. Appendix A.10 gives the trajectory, binarization, and sample-size diagnostics.

The evidence concerns the tested policy, aggregation, datasets, and budgets. The fixed-V control covers the upper range of the synergy fraction. The post-hoc feasibility mask requires prospective validation. Our implementation binarizes features and assumes a binary outcome. Pairwise PID cannot represent interactions requiring three or more sources, while higher-order extensions introduce computational challenges. Amortized bias and variance are not fully characterized, and the reduced-dimensional exact control changes more than the amortized type. The real-data benchmarks use uniform costs, while heterogeneous costs are tested only in the controlled sweep. Bootstrap tests are not adjusted for multiple comparisons, and the small test sets of CKD and ACTG175 limit precision. Benchmark policies use a single training seed, so the bootstrap reflects test sampling, not training variability. The shared classifier supports comparisons under a fixed prediction model, rather than implying that the predictive utility would hold with a different classifier.

The results motivate evaluating whether candidate scores align with incremental utility given the acquired set before investing in more extensive search. Prospective comparisons could combine feasibility within the remaining budget and objectives that discount already acquired information, and test robustness across prediction models and policy-specific mask distributions. These are design directions, not validated improvements. Higher-order PID remains complementary. The duplicate-copy construction shows that this aggregation can fail even without higher-order synergy. Atom-specific reweighting also requires an active pair branch, so its effect depends on the weight of acquisition costs. Also, it does not generally rescue the negative benchmark pattern (Appendix A.15).

8 Conclusion

We studied when pairwise information structure can support active feature acquisition through SynAFA, a policy combining joint information and realized-value conditional information. Across the evaluated benchmark datasets, its benefits are heterogeneous. Feature structure supported by permutation tests accompanies its strongest gains on PhysioNet at low budgets, which do not depend on the pair branch. However, these gains do not extend consistently across datasets or budgets. Controlled sweeps provide more targeted evidence. Within the tested synthetic family, SynAFA's advantage increases as pairwise information becomes more synergy-dominated, provided the relevant acquisition is feasible in terms of budget. Additionally, pair branch ablation shows that this benefit comes from pair proposals, which contribute little on most tested real datasets. A feasibility-mask diagnostic links the high-synergy losses under tight budgets to pairs that cannot be completed.

A distinct limitation emerges when local scores are aggregated into a set-level objective. Non-greedy single-swap local search improves the diagnostic objective, yet produces no accuracy gain in any of the tested settings. Accepted objective improvements are weakly correlated with changes in the predictive utility of the shared classifier. The duplicate-copy construction identifies one structural reason why this aggregation can overcount information. These findings are limited to the tested objective, single-swap local search, and the masked evaluation classifier. They do not imply that pairwise information or PID is ineffective more generally. They motivate acquisition objectives that jointly account for feasibility within the remaining budget, information already available in the acquired subset, and downstream predictive utility.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing throughout the manuscript, and for retrieval and discovery when checking related-work citations. We also used generative AI tools, to a limited extent, for feedback on research methodology (for example, flagging an inconsistency that motivated an ablation removing the pair-proposal branch) and for assistance in clarifying the wording of one existing proof. We have not used generative AI tools to generate synthetic datasets, develop the paper's theoretical framework, formulate its mathematical claims, implement its methods, or clean or reformat any dataset, and these tasks are not applicable to this work. We have reviewed all AI-assisted work: suggested text and interpretations were checked against the underlying code, data, and experimental results before being incorporated, and any AI-suggested methodological checks were run and verified by the authors before being reported in the paper. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available benchmark datasets (CKD, ACTG175, BankMarketing, PhysioNet, MiniBooNE, and MNIST), accessed and preprocessed through the AFABench benchmark, together with a purpose-built synthetic dataset. No new data involving human subjects was collected for this study. The clinical datasets (CKD, ACTG175, PhysioNet) are de-identified, publicly released benchmark datasets, used under their existing terms of release and consistent with their original intended research use. This work studies a general methodology for active feature acquisition and does not deploy a clinical or other decision-making system. We are not aware of discriminatory applications, dual-use concerns, or other potential harms specific to this work.

REPRODUCIBILITY STATEMENT

Preprocessing, the evaluation protocol (including the hard-budget harness and its correction), and baseline configurations are detailed in Appendix A.1. Formal definitions of the Partial Information Decomposition quantities used throughout, and the proof of the redundancy-overcounting proposition, are given in Appendix A.2; the PhysioNet permutation-test procedure is described in Appendix A.3. The synthetic dataset's generative process, including all weights and calibration, is specified in Appendices A.4 and A.8. The amortized estimator's architecture, training procedure, and hyperparameters are given in Appendix A.12, and the external classifier's architecture and training are described in Appendix A.9. Dataset sizes, splits, and sources are listed in Appendix A.6 and Table A1. The pair branch ablation, cost-weight sensitivity analyses, and local-search methodology are fully specified in Appendices A.13, A.14, A.15, and A.11, respectively.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Linus Aronsson, Arman Rahbar, and Morteza Haghir Chehreghani. A survey on active feature acquisition strategies. *arXiv preprint arXiv:2502.11067*, 2025.

Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, pp. 444–453. PMLR, 2019.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pp. 6424–6447. PMLR, 2023.

Soham Gadgil, Ian Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *International Conference on Learning Representations*, volume 2024, pp. 34979–35010, 2024.

Hung-Tien Huang, Dzung Dinh, and Junier B Oliva. Information templates: A new paradigm for intelligent active feature acquisition. *arXiv preprint arXiv:2508.18380*, 2025.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Valter Schütz, Han Wu, Reza Rezvan, Linus Aronsson, and Morteza Haghir Chehreghani. Afabench: A generic framework for benchmarking active feature acquisition. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 9765–9776, 2026.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Why pay more when you can pay less: A joint learning framework for active feature acquisition and classification. *arXiv preprint arXiv:1709.05964*, 2017.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 computing in cardiology*, pp. 245–248. IEEE, 2012.

Michael Valancius, Max Lennon, and Junier Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition. *arXiv preprint arXiv:2302.13960*, 2023.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *Journal of Machine Learning Research*, 24(131):1–44, 2023.

Sara Zannone, José Miguel Hernández-Lobato, Cheng Zhang, and Konstantina Palla. Odin: Optimal discovery of high-value information using model-based deep reinforcement learning. In *ICML Real-world Sequential Decision Making Workshop*, 2019.