γ-CMI: NON-MYOPIC ACTIVE FEATURE ACQUISITION VIA PROBABILISTIC LOOKAHEAD

Anonymous authors
Paper under double-blind review

ABSTRACT

Active Feature Acquisition (AFA) is a sequential decision-making process that selectively acquires unobserved features for each data instance independently to improve predictive performance under a limited acquisition budget. A common approach is to greedily select the next feature by maximizing its conditional mutual information (CMI) with the prediction target given the currently observed features. However, this criterion considers only the immediate information gain of each candidate feature and may therefore overlook feature interactions, where a feature becomes more informative when considered jointly with other features that may be acquired subsequently. To address this limitation, we propose γ-CMI, a generalization of the CMI acquisition criterion that evaluates candidate features under probabilistically sampled future observation contexts. γ-CMI provides a continuous spectrum of lookahead, ranging from evaluating a candidate's immediate information gain through Greedy CMI ($\gamma = 0$) to evaluating its information gain under the fully sampled future context of all remaining features ($\gamma = 1$). Crucially, γ-CMI can be incorporated into existing CMI-based AFA methods in a plug-and-play manner, requiring only minor modifications to their acquisition criteria and no additional model retraining. We provide a theoretical analysis characterizing the acquisition behavior induced by γ-CMI, demonstrate its ability to capture feature interactions on diverse synthetic datasets, and show consistent improvements in early-stage acquisition performance across diverse real-world datasets and CMI-based backbones. Code is available here.

1 INTRODUCTION

Reliable decision-making in real-world applications often requires information that is not fully available at the outset and must instead be acquired sequentially through additional observations or tests (Kachuee et al., 2019a; Erion et al., 2022; He et al., 2022). For example, a clinician may decide whether to order blood tests or medical imaging based on a patient's symptoms and medical history (Kachuee et al., 2019b), while an equipment monitoring system may acquire additional diagnostic measurements based on sensor observations collected so far. This problem, known as *Active Feature Acquisition* (AFA), seeks to sequentially acquire the features needed for prediction based on the current observation state of each instance (Saar-Tsechansky et al., 2009; Early et al., 2016; He et al., 2016). Since each acquisition incurs time or monetary cost, AFA aims to maximize predictive performance under a limited acquisition budget. A common approach is to greedily select the next feature using *Conditional Mutual Information* (CMI) (Geman & Jedynak, 1996; Chen et al., 2015; Ma et al., 2019), which measures the information that an unobserved feature provides about the prediction target conditioned on the features observed so far.

However, existing Greedy CMI-based methods are inherently *myopic*. Since they evaluate candidates strictly under the current observation state, they cannot anticipate how a feature's predictive value might evolve as subsequent features are acquired. Consequently, they often overlook synergistic features – those that appear uninformative in isolation but become highly informative when combined with other unobserved features (Norcliffe et al., 2025). Hence, Greedy CMI-based methods routinely assign low priority to these features, leaving critical higher-order information unexploited.

Contribution. To address this limitation, we propose *γ-CMI*, a *non-myopic* generalization of the CMI acquisition criterion that evaluates features by anticipating hypothetical future observation

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.

states. Specifically, for each candidate feature, γ-CMI calculates the conditional information gain across sampled future contexts where subsets of the remaining unobserved features are additionally assumed to be known. The parameter $\gamma \in [0, 1]$ explicitly controls this effective lookahead horizon: higher values place greater weight on higher-order multi-feature interactions, while $\gamma = 0$ exactly recovers standard Greedy CMI. Since it modifies only the evaluation context at inference time, γ-CMI operates as a *plug-and-play* module without any architectural changes or additional retraining. We theoretically characterize how γ-CMI weights information contributions across different interaction orders and how generative model errors affect its acquisition behavior. We then validate these theoretical predictions on controlled synthetic datasets and evaluate the practical impact of γ-CMI across diverse real-world datasets and existing AFA methods. Our results demonstrate that accounting for potential future feature interactions can improve existing CMI-based acquisition methods.

2 Preliminaries

2.1 Problem Setup: Active Feature Acquisition

Let $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a random vector consisting of d features, and let $Y \in \mathcal{Y}$ denote a random variable representing the prediction target. We use $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathcal{Y}$ to denote their respective realizations. For C-class classification, we have $\mathcal{Y} = \{1, \ldots, C\}$, while for regression, $\mathcal{Y} = \mathbb{R}$. Active Feature Acquisition (AFA) considers the problem of sequentially acquiring unobserved features with an initial set of observed features to maximize the final predictive performance under a fixed acquisition budget (*e.g.,* a maximum number of acquisitions). More specifically, at each acquisition step, we use $\mathcal{O} \subseteq [d]$ to denote the set of feature indices observed so far, and $\mathcal{U} = [d] \setminus \mathcal{O}$ to denote the set of remaining unobserved feature indices. We denote the random variables corresponding to these subsets by $\mathbf{X}_\mathcal{O}$ and $\mathbf{X}_\mathcal{U}$, respectively. Given the current observations $\mathbf{x}_\mathcal{O}$, the model computes an acquisition score $R(i, \mathbf{x}_\mathcal{O})$ for each candidate feature X_i, $i \in \mathcal{U}$, and selects the next feature X_{i^*} to acquire according to

$$i^* = \arg\max_{i \in \mathcal{U}} R(i, \mathbf{x}_\mathcal{O}). \tag{1}$$

Once the value of the selected feature x_{i^*} is acquired, the index sets are updated as $\mathcal{O} \leftarrow \mathcal{O} \cup \{i^*\}$ and $\mathcal{U} \leftarrow \mathcal{U} \setminus \{i^*\}$, respectively. This procedure is repeated until the acquisition budget is exhausted.

2.2 Greedy CMI as Myopic Feature Acquisition

A canonical information-theoretic approach to AFA is to greedily maximize Conditional Mutual Information (CMI). Under this paradigm, the acquisition score of a candidate feature X_i is defined as the point-wise CMI between X_i and the target Y, given the current observations $\mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}$:

$$\begin{aligned} R_G(i, \mathbf{x}_\mathcal{O}) &= I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) \\ &= H(Y \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) - \mathbb{E}_{x_i \sim p(X_i \mid \mathbf{x}_\mathcal{O})} \left[H(Y \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, X_i = x_i) \right]. \end{aligned} \tag{2}$$

Here, $H(\cdot)$ denotes the Shannon entropy. Thus, the CMI acquisition score quantifies the expected reduction in predictive uncertainty from acquiring feature X_i. Since X_i is unobserved at the time of acquisition, evaluation requires marginalizing over the conditional distribution $p(X_i \mid \mathbf{x}_\mathcal{O})$. In practice, this distribution is typically approximated using conditional generative models or variational inference methods, which provide samples or estimates of plausible values for x_i.

Limitation of Greedy CMI. Many existing AFA methods rely on greedily maximizing CMI. However, this strategy is inherently myopic: it evaluates only the immediate information gain of a candidate feature X_i under the current observation $\mathbf{x}_\mathcal{O}$, without considering how acquiring X_i may affect the future observation trajectory. Let $\mathcal{S} \subseteq \mathcal{U} \setminus \{i\}$ denote the indices of a subset of additional unobserved features that may be acquired after X_i. To characterize how the predictive value of X_i changes when these additional features $\mathbf{X}_\mathcal{S}$ are observed, we define conditional interaction information, following prior work (McGill, 1954; Kim et al., 2024), as:

$$J(Y; X_i; \mathbf{X}_\mathcal{S} \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) = I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\mathcal{S}) - I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}). \tag{3}$$

The conditional interaction information quantifies how the information that X_i provides about Y changes when additional features $\mathbf{X}_\mathcal{S}$ are observed, given the current observations $\mathbf{x}_\mathcal{O}$. A positive

value ($J > 0$) indicates a *synergistic interaction*, meaning that X_i becomes more informative about Y when considered together with $\mathbf{X}_\mathcal{S}$. Conversely, a non-positive value ($J \leq 0$) indicates a *non-synergistic* (or redundant) interaction, where conditioning on $\mathbf{X}_\mathcal{S}$ does not increase and may reduce the information provided by X_i. From this perspective, Greedy CMI only evaluates the immediate information gain of a candidate feature X_i under the current observation context $\mathbf{x}_\mathcal{O}$, and thus does not account for interactions with potentially observable future features $\mathbf{X}_\mathcal{S}$. In particular, it may underestimate the value of synergistic candidate features whose informativeness increases substantially when observed together with other currently unobserved features. This motivates evaluating candidate features not only under the current context, but also under possible future feature contexts.

3 γ-CMI AS NON-MYOPIC FEATURE ACQUISITION

In this section, we extend the CMI-based acquisition criterion to account for potential future observation contexts by evaluating each candidate feature under randomly sampled future feature sets.

3.1 γ-CMI ACQUISITION SCORE

To address the myopic limitation of Greedy CMI, we propose γ-CMI (*Gamma-CMI*), which evaluates a candidate feature by accounting for its potential interactions with future observations. For a candidate feature index $i \in \mathcal{U}$, let $\Phi \subseteq \mathcal{U} \setminus \{i\}$ be a random set of potentially observable future feature indices, sampled from a distribution $p(\Phi)$ over all possible future feature subsets. While this distribution can be chosen flexibly, we adopt independent Bernoulli sampling in this work. Specifically, each remaining unobserved feature index $j \in \mathcal{U} \setminus \{i\}$ is independently included in Φ with probability $\gamma \in [0, 1]$. This sampling procedure induces a distribution $p_\gamma(\Phi)$, where γ directly controls the expected number of included features. For each realization $\Phi = \phi$, we sample the corresponding feature values from the conditional distribution $p(\mathbf{X}_\phi \mid \mathbf{x}_\mathcal{O})$. The γ-CMI acquisition score then averages the information gain of X_i over possible future observation contexts:

$$\begin{aligned} R_\gamma(i, \mathbf{x}_\mathcal{O}) &= \mathbb{E}_{\phi \sim p_\gamma} \mathbb{E}_{\mathbf{x}_\phi \sim p(\mathbf{X}_\phi \mid \mathbf{x}_\mathcal{O})} \left[I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\phi = \mathbf{x}_\phi) \right] \\ &= R_G(i, \mathbf{x}_\mathcal{O}) + \mathbb{E}_{\phi \sim p_\gamma} \left[J(Y; X_i; \mathbf{X}_\phi \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) \right]. \end{aligned} \tag{4}$$

See Appendix C.1 for derivation. Eq. (4) shows that γ-CMI takes the standard Greedy CMI and explicitly adds extra terms that capture the expected interaction information with subsequently acquired features. The hyperparameter γ acts as a lookahead regularizer that controls the horizon of these hypothetical future observations. Its structural role is most evident at its two extremes:

- **Boundary Case I ($\gamma = 0$).** The future context is strictly empty (*i.e.,* $\Phi = \varnothing$). Here, γ-CMI evaluates the feature solely under current evidence, reducing exactly to the myopic Greedy CMI:

$$R_0(i, \mathbf{x}_\mathcal{O}) = R_G(i, \mathbf{x}_\mathcal{O}) = I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}). \tag{5}$$

- **Boundary Case II ($\gamma = 1$).** The context includes all remaining unobserved features except the candidate itself (*i.e.,* $\Phi = \mathcal{U} \setminus \{i\}$). Here, γ-CMI evaluates the feature under a hypothetical terminal state where all other features have already been observed:

$$R_1(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\mathbf{x}_{\mathcal{U}\setminus\{i\}} \sim p(\mathbf{X}_{\mathcal{U}\setminus\{i\}} \mid \mathbf{x}_\mathcal{O})} \left[I\left(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_{\mathcal{U}\setminus\{i\}} = \mathbf{x}_{\mathcal{U}\setminus\{i\}}\right) \right]. \tag{6}$$

For intermediate values $0 < \gamma < 1$, γ-CMI defines a probabilistic continuum between these bounds. It smoothly interpolates between prioritizing the feature that is most informative immediately ($\gamma = 0$) and prioritizing the feature whose absence would matter most after all other unobserved features had been acquired ($\gamma = 1$). Consequently, tuning γ explicitly modulates the extent to which downstream feature interactions influence the immediate acquisition decision.

3.2 ANALYSIS OF THE γ-CMI ACQUISITION SCORE

Interaction-Order Analysis via Möbius Decomposition. To better understand the theoretical properties of γ-CMI, we apply the Möbius decomposition (Rota, 1964; Grabisch et al., 2000) to characterize the information contributions arising from different feature combinations. For any subset of unobserved feature indices $\mathcal{S} \subseteq \mathcal{U}$, the CMI between $\mathbf{X}_\mathcal{S}$ and the target Y, given the

Algorithm 1 Plug-and-Play Implementation of the γ-CMI Acquisition Score

Sample $\mathbf{x}_{\mathcal{U}}^{(1)}, \ldots, \mathbf{x}_{\mathcal{U}}^{(K)} \sim p_\theta(\mathbf{X}_{\mathcal{U}} \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}})$
Construct $\tilde{\mathbf{x}}^{(k)} = \mathbf{m}_{\mathcal{O}} \odot \mathbf{x}_{\mathcal{O}} + (1 - \mathbf{m}_{\mathcal{O}}) \odot \mathbf{x}_{\mathcal{U}}^{(k)}$ for $k = 1, \ldots, K$
Sample $\mathbf{m}^{(r)}$ for $r = 1, \ldots, M$, where $m_j^{(r)} \sim \text{Bernoulli}(\gamma)$ if $j \in \mathcal{U}$, and 0 otherwise.
for each $i \in \mathcal{U}$ **do**
  $R_\gamma(i, \mathbf{x}_{\mathcal{O}}) \leftarrow 0$, $\mathbf{m}_i \leftarrow \texttt{OneHot}(i)$ ▷ Initialize for candidate i
  $\mathbf{m}_\Phi^{(r)} \leftarrow \mathbf{m}^{(r)}$, $m_{\Phi,i}^{(r)} \leftarrow 0$ for $r = 1, \ldots, M$ ▷ Exclude candidate i from Φ
  for $r = 1, \ldots, M$ **do**
    $H_{\text{without}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} H(f_\psi(\tilde{\mathbf{x}}^{(k)} \odot (\mathbf{m}_{\mathcal{O}} + \mathbf{m}_\Phi^{(r)}), \mathbf{m}_{\mathcal{O}} + \mathbf{m}_\Phi^{(r)}))$
    $H_{\text{with}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} H(f_\psi(\tilde{\mathbf{x}}^{(k)} \odot (\mathbf{m}_{\mathcal{O}} + \mathbf{m}_\Phi^{(r)} + \mathbf{m}_i), \mathbf{m}_{\mathcal{O}} + \mathbf{m}_\Phi^{(r)} + \mathbf{m}_i))$
    $R_\gamma(i, \mathbf{x}_{\mathcal{O}}) \leftarrow R_\gamma(i, \mathbf{x}_{\mathcal{O}}) + \frac{1}{M}(H_{\text{without}} - H_{\text{with}})$
  end for
end for
return $\arg\max_{i \in \mathcal{U}} R_\gamma(i, \mathbf{x}_{\mathcal{O}})$

Red highlights the components introduced by the proposed γ-CMI extension.

current context $\mathbf{x}_{\mathcal{O}}$, decomposes exactly into the sum of its constituent interaction contributions: $I(Y; \mathbf{X}_{\mathcal{S}} \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}}) = \sum_{\mathcal{T} \subseteq \mathcal{S}} a(\mathcal{T})$. Here, $a(\mathcal{T})$ denotes the Möbius representation of the subset $\mathcal{T} \subseteq \mathcal{S}$. This term isolates the unique, higher-order information contribution of $\mathcal{T}$ that cannot be captured by evaluating any of its proper subsets independently.

Now, to evaluate a specific candidate feature X_i, we group these interaction terms by their interaction order. We define the aggregate r-th order information contribution involving feature X_i as

$$A_r(i) = \sum_{\mathcal{T} \subseteq \mathcal{U} \setminus \{i\}, |\mathcal{T}| = r-1} a(\mathcal{T} \cup \{i\}). \tag{7}$$

The first-order term, $A_1(i)$, represents the individual information contribution of feature X_i in isolation. Note that this reduces exactly to the Greedy CMI score: $A_1(i) = a(\{i\}) = I(Y; X_i \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}}) = R_G(i, \mathbf{x}_{\mathcal{O}})$. For $r \geq 2$, $A_r(i)$ aggregates all r-th order interaction contributions involving X_i and $r - 1$ other currently unobserved features. Using this notation, the γ-CMI acquisition score can be decomposed according to interaction order as in the following proposition.

Proposition 3.1. *Under independent Bernoulli sampling with parameter γ, the γ-CMI acquisition score can be given as a γ-weighted sum of information contributions across all interaction orders:*

$$R_\gamma(i, \mathbf{x}_{\mathcal{O}}) = \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} A_r(i) = A_1(i) + \gamma A_2(i) + \gamma^2 A_3(i) + \cdots + \gamma^{|\mathcal{U}|-1} A_{|\mathcal{U}|}(i). \tag{8}$$

The proof is provided in Appendix C.2.

This proposition demonstrates that γ directly controls how deeply future feature interactions influence the current acquisition score. Specifically, any r-th order interaction involving the candidate feature is weighted by γ^{r-1}. At $\gamma = 0$, the first-order contribution $A_1(i)$ is isolated, exactly recovering Greedy CMI. As γ increases, the attenuation of higher-order interactions is reduced, enabling complex synergies with unobserved features to increasingly drive the acquisition process.

Beyond its original motivation as a non-myopic extension of Greedy CMI, the order-wise decomposition in (8) reveals connections between γ-CMI and other established methodologies. Its γ-weighted interaction expansion is analogous to discounted returns in reinforcement learning, while its subset-averaged marginal contributions admit an exact probabilistic-value interpretation in cooperative game theory. We provide more details in Appendix A.

3.3 Practical Considerations for γ-CMI

Plug-and-Play Implementation. γ-CMI modifies only the observation context under which the acquisition score is evaluated, while leaving the underlying generative and predictive models in Greedy CMI-based AFA models unchanged. It can therefore be implemented as a *plug-and-play* module with only minor modifications to the acquisition criterion, requiring neither architectural changes nor additional model training.

We represent the current and hypothetical future observation contexts using two gate vectors:

- $\mathbf{m}_{\mathcal{O}} = (m_{\mathcal{O},1}, \ldots, m_{\mathcal{O},d}) \in \{0, 1\}^d$ indicates the set of features observed so far, where $m_{\mathcal{O},j} = 1$ if $j \in \mathcal{O}$ and $m_{\mathcal{O},j} = 0$ otherwise.

- $\mathbf{m}_\Phi = (m_{\Phi,1}, \ldots, m_{\Phi,d}) \in \{0,1\}^d$ specifies the future observation context for candidate feature X_i, where $m_{\Phi,j} \sim \texttt{Bern}(\gamma)$ for $j \in \mathcal{U} \setminus \{i\}$ and $m_{\Phi,j} = 0$ otherwise.

Given the current partial observation $\mathbf{x}_\mathcal{O} = \mathbf{m}_\mathcal{O} \odot \mathbf{x} + (1 - \mathbf{m}_\mathcal{O}) \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ denotes a baseline unobserved or noise value, we first impute the unobserved features via the conditional generative model $\mathbf{x}_\mathcal{U} \sim p_\theta(\mathbf{X}_\mathcal{U} \mid \mathbf{x}_\mathcal{O})$. The sampled future context is then constructed by gating these generated values, *i.e.,* $\mathbf{x}_\Phi = \mathbf{m}_\Phi \odot \mathbf{x}_\mathcal{U}$. Then, the predictor f_ψ estimates the conditional entropy under two distinct scenarios: one excluding the candidate feature i, and one including its imputed value x_i. This overall procedure is summarized in Algorithm 1.

In practice, the expectation over Φ can be approximated using M independently sampled future-context masks. With $|\mathcal{U}|$ remaining candidate features, this results in $O(M|\mathcal{U}|)$ predictor evaluations per acquisition step, compared with $O(|\mathcal{U}|)$ for Greedy CMI. Since the evaluations associated with different future-context masks are independent, they can be executed in parallel to reduce wall-clock overhead. We empirically evaluate the resulting inference-time overhead in Appendix J, observing an average wall-clock overhead of 1.08×.

Impact of Generative Model Error. As described in Algorithm 1, γ-CMI relies on a conditional generative model p_θ to generate values for unobserved features. As γ increases, more generated feature values are incorporated into the hypothetical future observation context and provided to the predictor f_ψ, making the acquisition score potentially more sensitive to errors in the generated values. Our theoretical analysis in Appendix B shows that increasing γ places greater weight on higher-order error terms. Importantly, however, these errors do not necessarily change the acquisition decision, since AFA depends on the relative ranking of candidate acquisition scores rather than their exact values. The complete theoretical analysis and proofs are provided in Appendix B and we empirically investigate the resulting trade-off in Section 5.2.

4 Related Work

CMI-Based Approaches. CMI-based methods are among the most widely adopted approaches for AFA. These methods typically use conditional generative models to estimate the expected information gain of unobserved features. Ma et al. (2019) model the conditional distribution of unobserved features using a partial VAE, while Li et al. (2020) perform conditional density estimation with ACFlow. Several discriminative approaches instead learn Greedy CMI policies without explicitly modeling the feature distribution. Covert et al. (2023) train a policy network through amortized optimization to directly select the next feature, while Gadgil et al. (2024) train a value network to regress the CMI of each candidate feature. Although these approaches employ policy or value networks and thus resemble reinforcement learning, they remain within the CMI-based paradigm because their training objectives are designed to reproduce Greedy CMI decisions at each acquisition step. More recently, Kobayashi et al. (2026) leverage pretrained models to estimate uncertainty in-context for unseen tasks, enabling CMI-based acquisition without task-specific retraining.

MDP-Based Approaches. Because AFA is inherently sequential, it can be formulated as a Markov decision process (MDP) (Rückstieß et al., 2013). Kachuee et al. (2019a) propose a DQN-based approach that uses changes in predictive uncertainty as the reward, while Li & Oliva (2021) incorporate a generative model into reinforcement learning to provide CMI-based intermediate rewards and auxiliary information during policy training. This illustrates that CMI-based and MDP-based approaches are not mutually exclusive formulations of AFA. Although reinforcement learning can, in principle, learn non-myopic acquisition policies by optimizing long-term returns, it introduces additional optimization challenges (Henderson et al., 2018; van Hasselt et al., 2018), and existing studies have not shown consistent advantages over CMI-based methods (Aronsson et al., 2026).

Other Approaches. Beyond CMI- and MDP-based approaches, several methods have explored alternative acquisition criteria. Norcliffe et al. (2025) use the magnitude of the change in predictive gradients induced by acquiring a candidate feature as the acquisition score, while Guney et al. (2025) learn acquisition policies from explainability-driven feature importance rankings. More recently, foundation models have been applied to sequential information acquisition; Nori et al. (2025) propose a sequential diagnostic framework in which a language model selects the next question or test based on the clinical information acquired so far and the current diagnostic hypotheses.

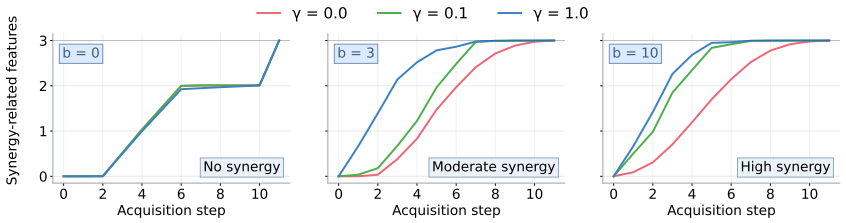


Figure 1: Acquisition of synergy-related features. The vertical axis shows the average number of times the two features in the active interaction pair and the switch feature are acquired. The covariance matrix is set to the identity matrix.

5 Experiments

In this section, we evaluate γ-CMI in two stages. We first investigate its underlying acquisition behavior using carefully designed synthetic datasets to answer four key questions:

- Can γ-CMI effectively capture *synergistic interactions* including unobserved features?
- How does *redundancy* including unobserved features affect γ-CMI acquisition behavior?
- How does γ affect the acquisition of features involved in *higher-order interactions*?
- How does *generative model error* affect γ-CMI across different values of γ?

We then evaluate γ-CMI on a diverse collection of real-world datasets spanning both classification and regression tasks. These experiments assess the practical benefit of incorporating γ-CMI into existing CMI-based AFA methods, demonstrating how the proposed criterion affects acquisition performance when used as a plug-and-play extension of pretrained CMI-based backbones.

5.1 Synthetic Experiments: Acquisition Behavior in a Gaussian Setting

Dataset Generation. Inspired by Yoon et al. (2019), we design the *Gaussian Dataset* for binary classification. Each data point is sampled from an 11-dimensional Gaussian distribution, *i.e.,* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\mathbf{x} \in \mathbb{R}^{11}$, where the features are constructed as follows: x_0 and x_1 provide independent, additive information; $x_2, \ldots, x_5$ form two potential synergistic pairs (x_2, x_3) and (x_4, x_5); $x_6, \ldots, x_9$ are purely noise; and x_{10} acts as a switch determining which synergistic pair is active. The binary label is sampled via $y \sim \texttt{Bern}(\texttt{sigmoid}(\ell))$ where $\ell = ax_0 + ax_1 + bx_2x_3$ if $x_{10} < 0$, and $\ell = ax_0 + ax_1 + bx_4x_5$ otherwise. Thus, while x_0 and x_1 are individually informative, the paired features contribute primarily through interactions whose relevance depends on the switch feature x_{10}. These interactions are thus difficult to identify from individual features alone.

Experimental Setup. Because both the feature distribution and the label-generation process are known in closed form, we use an oracle generative model and an exact predictor. This removes estimation error from feature generation and prediction, allowing us to attribute differences in acquisition behavior directly to the acquisition criteria. Throughout the following experiments, we fix the individual contribution parameter to $a = 0.5$ and vary the synergy parameter b and the correlation coefficients ρ_{ij} in Σ to independently control the strength of synergy and redundancy.

Behavior on Synergistic Interactions (Q1). We control the strength of the synergistic interactions compared with the individual contributions of x_0 and x_1. As shown in Figure 1, when no synergy is present ($b = 0$), γ-CMI exhibits acquisition behavior nearly identical to that of Greedy CMI. As the synergy strength increases, however, Greedy CMI tends to overlook the interactive features because their immediate information gains remain weak when considered in isolation. In contrast, for $\gamma > 0$, γ-CMI accounts for their synergistic interactions by evaluating their information gain under future contexts. Consequently, γ-CMI increasingly prioritizes these features as the synergy strength grows.

Observation 5.1 (γ-CMI captures synergistic features)**.** *The γ-CMI acquisition score accounts not only for the immediate information gain of a candidate feature, but also for the additional information that can emerge when other currently unobserved features are subsequently observed.*

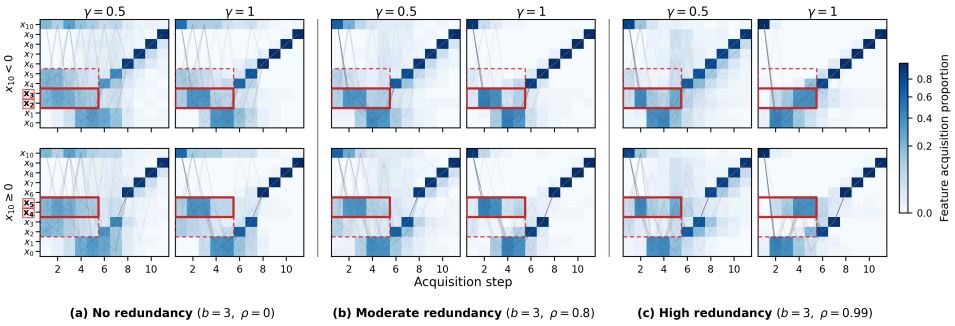


Figure 2: Feature acquisition trajectories under within-pair redundancy. Darker cells indicate higher feature acquisition proportions, and gray lines show representative trajectories. Solid and dashed red boxes mark the active and alternative interaction pairs, respectively, over the first five steps.

Behavior on Redundancy (Q2). We examine within-pair redundancy by setting $b = 3$ and $\rho_{23} = \rho_{32} = \rho_{45} = \rho_{54} = \rho$. As ρ increases, the features within each interaction pair become increasingly redundant. Under larger γ, the correlated counterpart of a candidate is more likely to appear in the sampled future context. Conditioning on this counterpart then reduces the candidate's incremental information about the target, resulting in a negative interaction contribution and a lower γ-CMI score. As shown in Figure 2, at $\rho = 0$, $\gamma = 1$ assigns high early acquisition priority to both the switch and interaction features. As ρ increases to 0.8 and 0.99 (moderate and high redundancy), the interaction features progressively lose priority, falling below the individually informative features x_0 and x_1 at $\rho = 0.99$. These results show that γ-CMI naturally downweights features whose information overlaps with sampled future contexts.

Observation 5.2 (γ-CMI avoids redundant features)**.** *By evaluating candidate features across potential future observation contexts, γ-CMI accounts for information overlap and assigns lower acquisition priority to features that are redundant with other unobserved features.*

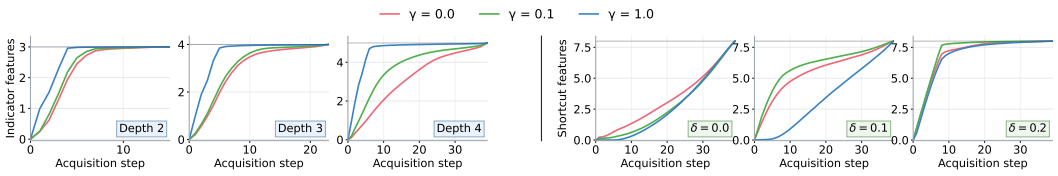
Additional analysis on the impact of redundancy can be found in Appendices D.2 and D.3.

5.2 Synthetic Experiments: Acquisition Behavior in an Indicator Tree Setting

Dataset Generation. To evaluate the scenario where a feature's relevance depends heavily on the sequential acquisition history, we construct the *Binary Indicator Tree Dataset*, which forms stacked layers of indicator features. The task is binary classification with binary features. The first layer contains feature x_0, telling us which feature in the next layer (out of x_1 and x_2) is the next indicator. Each layer doubles in size forming a tree, where each feature in a layer recursively determines which branch to take in the next layer. After n layers, the leaf node reached determines y.

Experimental Setup. Recovering the true label requires sequentially acquiring the feature path indicated by the active nodes. As the tree depth increases, the acquisition criterion must consider more features over longer horizons, capturing higher-order interactions. We vary the tree depth from 2 to 4. We use VAE-MLP, comprising a VAE generative model and an MLP predictor. To test the robustness of the acquisition criteria against myopic decisions, we introduce eight additional "shortcut" features that provide immediate but noisy information about the label. Each shortcut feature X_j is sampled from $P(X_j = y \mid Y = y) = \frac{1}{2} + \delta$ for $y \in \{0, 1\}$. Here, δ controls the strength of the immediate signal, with larger values making the shortcut features more informative and thus more attractive to myopic acquisition strategies than features along the true indicator path.

Behavior on Higher-Order Interactions (Q3). Figure 3 shows that, as the depth increases, smaller values of γ become less effective at prioritizing the relevant indicator features, while larger values of γ robustly maintain a high acquisition rate for the correct indicator path. This empirical trend is consistent with Proposition 3.1, which establishes that γ dictates the relative weighting of interaction orders, with larger values mitigating the attenuation of higher-order terms. These results show that γ-CMI can adapt its acquisition behavior to the structural depth of feature dependencies.



(a) vs. depth ($\delta = 0.0$) **(b) vs. noise (Depth 4)**

Figure 3: Acquisition of (a) indicator features across different tree depths and (b) shortcut features across different noise levels. The vertical axes follow the same definitions as in Figure 1.

Observation 5.3 (Larger γ enables deeper lookahead)**.** *By mitigating the attenuation of higher-order interaction terms, larger values of γ empower γ-CMI to successfully prioritize and acquire features that rely on deep, multi-step dependencies.*

Additional analysis of the shortcut signal strength δ and further results are provided in Appendix D.4.

Behavior on Generative Model Error (Q4). We evaluate γ-CMI using checkpoints of the conditional generative model p_θ from different stages of training. In addition to the final, low-error checkpoint, we consider two earlier checkpoints representing medium- and high-error regimes. As shown in Figure 4, γ-CMI improves performance for small to intermediate values of γ across all error regimes. However, as γ increases, the performance trajectories increasingly diverge: while the low-error model continues to benefit from deeper lookahead, larger generative errors limit these gains and can even cause performance degradation at high γ values. These results are consistent with our error analysis in Section 3.3. Generative errors do not necessarily change the acquisition decision when the resulting perturbations in acquisition scores are sufficiently small to preserve the candidate ranking. However, larger values of γ place greater weight on higher-order error terms, increasing the potential for generative errors to alter candidate rankings and thereby reducing the robustness of γ-CMI.

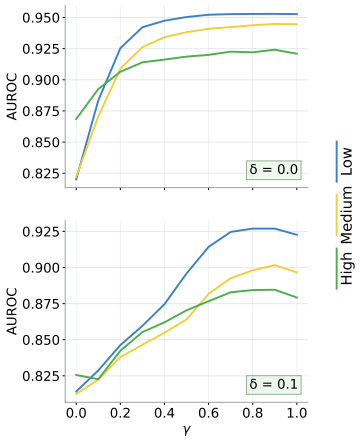


Figure 4: Performance under varying generative error regimes.

Observation 5.4 (Robustness to generative model error depends on γ)**.** *γ-CMI remains robust to generative model error when the resulting score perturbations preserve the candidate ranking, but its sensitivity to such errors increases with γ.*

5.3 Real-World Experiments

Datasets and Baselines. We evaluate γ-CMI on 12 real-world datasets, comprising six classification and six regression tasks (dataset details are provided in Appendix E.1). Rather than pursuing state-of-the-art performance, our primary goal is to examine whether γ-CMI can improve existing CMI-based AFA methods by replacing their acquisition criteria. We therefore focus on CMI-based baselines and apply γ-CMI in a plug-and-play manner to **ACFlow** (Li et al., 2020), **EDDI** (Ma et al., 2019), and **VAE-MLP** (hereafter VAE), which estimate CMI using generative models. We also conduct an exploratory evaluation with two amortized CMI-based acquisition methods, *i.e.,* **DIME** (Gadgil et al., 2024) and **GDFS** (Covert et al., 2023), to examine whether the future-context formulation can also be extended to methods that directly estimate acquisition scores. In this setting, we use a VAE to model the unobserved features when constructing hypothetical future contexts and fine-tune the pretrained amortized models, although fine-tuning is not required. For EDDI, we further include **EDDI+**, which adds an auxiliary prediction loss to the original training objective.

Evaluation Protocol. For a fair comparison, we maintain the original training and evaluation settings of each backbone and modify only the acquisition procedure by replacing Greedy CMI with

Table 1: Classification performance (AUROC for binary and Accuracy for multiclass tasks) on real-world datasets. Improvements are highlighted in **blue** and identical reported values are underlined.

			Bank Marketing	California Housing	MiniBooNE	MNIST	Fashion-MNIST	METABRIC
Explicit	ACFlow	Greedy	0.690 ± 0.036	0.504 ± 0.038	0.825 ± 0.053	0.510 ± 0.015	0.538 ± 0.007	0.440 ± 0.063
		γ-CMI	$\mathbf{0.811 \pm 0.029}$	$\mathbf{0.509 \pm 0.040}$	$\mathbf{0.855 \pm 0.033}$	0.510 ± 0.015	$\mathbf{0.540 \pm 0.007}$	$\mathbf{0.477 \pm 0.031}$
	EDDI	Greedy	0.656 ± 0.024	0.363 ± 0.047	0.854 ± 0.024	0.476 ± 0.005	0.530 ± 0.004	0.493 ± 0.019
		γ-CMI	$\mathbf{0.687 \pm 0.037}$	0.363 ± 0.047	$\mathbf{0.859 \pm 0.029}$	0.476 ± 0.005	$\mathbf{0.532 \pm 0.004}$	$\mathbf{0.537 \pm 0.026}$
	EDDI+	Greedy	0.647 ± 0.024	0.366 ± 0.062	0.890 ± 0.006	0.474 ± 0.014	0.538 ± 0.002	0.499 ± 0.017
		γ-CMI	$\mathbf{0.663 \pm 0.016}$	0.366 ± 0.062	0.890 ± 0.006	$\mathbf{0.480 \pm 0.012}$	0.538 ± 0.002	$\mathbf{0.531 \pm 0.033}$
	VAE	Greedy	0.849 ± 0.006	0.562 ± 0.006	0.908 ± 0.008	0.572 ± 0.003	0.594 ± 0.004	0.579 ± 0.014
		γ-CMI	$\mathbf{0.860 \pm 0.003}$	$\mathbf{0.565 \pm 0.007}$	$\mathbf{0.913 \pm 0.004}$	0.572 ± 0.002	0.594 ± 0.004	$\mathbf{0.584 \pm 0.010}$
Amortized	DIME	Greedy	0.888 ± 0.003	0.587 ± 0.010	0.935 ± 0.003	0.579 ± 0.003	0.603 ± 0.007	0.569 ± 0.018
		γ-CMI	$\mathbf{0.890 \pm 0.003}$	$\mathbf{0.589 \pm 0.007}$	$\mathbf{0.936 \pm 0.003}$	$\mathbf{0.585 \pm 0.002}$	$\mathbf{0.611 \pm 0.009}$	0.560 ± 0.030
	GDFS	Greedy	0.886 ± 0.003	0.573 ± 0.010	0.936 ± 0.001	0.591 ± 0.004	0.591 ± 0.014	0.571 ± 0.021
		γ-CMI	$\mathbf{0.888 \pm 0.001}$	$\mathbf{0.575 \pm 0.010}$	0.936 ± 0.001	$\mathbf{0.595 \pm 0.008}$	$\mathbf{0.596 \pm 0.008}$	$\mathbf{0.593 \pm 0.011}$

Table 2: Regression performance (RMSE) on real-world datasets. Improvements are highlighted in **blue** and identical reported values are underlined.

			Air Quality	California Housing	Energy Efficiency	Kin8nm	Parkinson	White Wine Quality
Explicit	ACFlow	Greedy	0.594 ± 0.090	0.920 ± 0.036	0.647 ± 0.116	0.950 ± 0.038	1.017 ± 0.015	0.959 ± 0.013
		γ-CMI	0.594 ± 0.090	$\mathbf{0.885 \pm 0.066}$	0.647 ± 0.116	$\mathbf{0.922 \pm 0.064}$	$\mathbf{1.006 \pm 0.015}$	0.959 ± 0.013
	EDDI	Greedy	0.410 ± 0.015	0.706 ± 0.013	0.602 ± 0.077	0.897 ± 0.034	0.978 ± 0.002	0.899 ± 0.004
		γ-CMI	0.410 ± 0.021	$\mathbf{0.699 \pm 0.008}$	0.602 ± 0.077	0.897 ± 0.034	$\mathbf{0.976 \pm 0.007}$	$\mathbf{0.894 \pm 0.003}$
	EDDI+	Greedy	0.442 ± 0.016	0.752 ± 0.033	0.541 ± 0.080	0.870 ± 0.034	0.976 ± 0.007	0.904 ± 0.004
		γ-CMI	$\mathbf{0.434 \pm 0.017}$	$\mathbf{0.735 \pm 0.017}$	$\mathbf{0.539 \pm 0.078}$	$\mathbf{0.866 \pm 0.027}$	$\mathbf{0.974 \pm 0.010}$	$\mathbf{0.900 \pm 0.004}$
	VAE	Greedy	0.314 ± 0.006	0.643 ± 0.004	0.354 ± 0.016	0.779 ± 0.005	0.944 ± 0.004	0.853 ± 0.007
		γ-CMI	$\mathbf{0.311 \pm 0.005}$	$\mathbf{0.642 \pm 0.004}$	$\mathbf{0.340 \pm 0.007}$	$\mathbf{0.769 \pm 0.003}$	$\mathbf{0.924 \pm 0.004}$	$\mathbf{0.843 \pm 0.005}$
Amortized	DIME	Greedy	0.411 ± 0.051	0.624 ± 0.005	0.487 ± 0.085	0.794 ± 0.007	0.887 ± 0.007	0.858 ± 0.018
		γ-CMI	$\mathbf{0.372 \pm 0.030}$	$\mathbf{0.623 \pm 0.005}$	$\mathbf{0.429 \pm 0.026}$	$\mathbf{0.792 \pm 0.003}$	$\mathbf{0.879 \pm 0.007}$	$\mathbf{0.839 \pm 0.011}$
	GDFS	Greedy	0.359 ± 0.021	0.644 ± 0.021	0.360 ± 0.009	0.784 ± 0.005	0.907 ± 0.013	0.845 ± 0.010
		γ-CMI	$\mathbf{0.357 \pm 0.018}$	0.646 ± 0.021	$\mathbf{0.348 \pm 0.018}$	$\mathbf{0.780 \pm 0.007}$	$\mathbf{0.894 \pm 0.009}$	$\mathbf{0.842 \pm 0.009}$

γ-CMI. We evaluate $\gamma \in \{0.0, 0.1, \ldots, 1.0\}$ on the validation set, select the value yielding the best validation performance, and report the corresponding performance on the test set. This sweep requires no additional model training, making hyperparameter tuning computationally lightweight. Since the effect of unobserved features is most relevant during the early stages of acquisition, we report the average performance over the first half of the total acquisition steps.

Results. Tables 1 and 2 show that γ-CMI improves acquisition performance across most datasets and backbones, while largely preserving performance otherwise. These results suggest that accounting for potential interactions with unobserved features can provide additional acquisition value beyond the immediate information provided by the current candidate feature. Notably, although the optimal value of γ varies across datasets and backbones, we observe a common overall trend: performance typically improves as γ increases from zero, reaches a peak at an intermediate value, and then declines as γ increases further. This pattern is consistent with the expected trade-off. Detailed performance curves across different γ values for each dataset are provided in Appendix I.2.

6 Conclusion

In this work, we introduce γ-CMI, a non-myopic generalization of Greedy CMI for active feature acquisition that accounts for potential future observation contexts. By evaluating candidate features under probabilistically sampled future observation contexts, γ-CMI captures information contributions arising from interactions with features that may be acquired subsequently, while recovering Greedy CMI as the special case $\gamma = 0$. Our theoretical analysis characterizes how γ controls the weighting of higher-order interactions and how generative model errors affect acquisition decisions. Controlled synthetic experiments validate these theoretical predictions, while experiments across diverse real-world datasets demonstrate that γ-CMI improves existing CMI-based AFA methods as a plug-and-play inference-time extension. Overall, our results suggest that explicitly accounting for potential future feature interactions provides a simple and effective way to move beyond myopic CMI-based acquisition.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing. We did not use generative AI tools for any other tasks requiring disclosure under the ICLR policy.

We reviewed all AI-assisted content and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses existing publicly available datasets and involves no new data collection or direct interaction with human subjects. Datasets containing human-derived clinical data are used solely for methodological benchmarking and in accordance with their respective terms of use.

REFERENCES

Linus Aronsson, Arman Rahbar, and Morteza Haghir Chehreghani. A Survey on Active Feature Acquisition Strategies. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 7740–7748, 2026.

Yuxin Chen, S. Hamed Hassani, Amin Karbasi, and Andreas Krause. Sequential Information Maximization: When is Greedy Near-optimal? In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 338–363, 2015.

Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009a. DOI: https://doi.org/10.24432/C56S3T.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4): 547–553, 2009b.

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 *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, 2012.

Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129:750–757, 2008.

Kirstin Early, Stephen E. Fienberg, and Jennifer Mankoff. Test time feature ordering with FOCUS: interactive predictions with minimal user burden. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 992–1003, 2016.

Gabriel Erion, Joseph D Janizek, Carly Hudelson, Richard B Utarnachitt, Andrew M McCoy, Michael R Sayre, Nathan J White, and Su-In Lee. A cost-aware framework for the development of AI models for healthcare applications. *Nature Biomedical Engineering*, 6(12):1384–1398, 2022.

Soham Gadgil, Ian Connick Covert, and Su-In Lee. Estimating Conditional Mutual Information for Dynamic Feature Selection. In *The Twelfth International Conference on Learning Representations*, 2024.

D. Geman and B. Jedynak. An active testing model for tracking roads in satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):1–14, 1996.

Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000.

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B Kolachalama. Active Feature Acquisition via Explainability-driven Ranking. In *Forty-second International Conference on Machine Learning*, 2025.

He He, Paul Mineiro, and Nikos Karampatziakis. Active Information Acquisition. *arXiv preprint arXiv:1602.02181*, 2016.

Weijie He, Xiaohao Mao, Chao Ma, Yu Huang, José Miguel Hernández-Lobato, and Ting Chen. BSODA: a bipartite scalable framework for online disease diagnosis. In *Proceedings of the ACM Web Conference 2022*, pp. 2511–2521, 2022.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning That Matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Mohammad Kachuee, Orpaz Goldstein, Kimmo Kärkkäinen, Sajad Darabi, and Majid Sarrafzadeh. Opportunistic Learning: Budgeted Cost-Sensitive Learning from Data Streams. In *International Conference on Learning Representations*, 2019a.

Mohammad Kachuee, Kimmo Kärkkäinen, Orpaz Goldstein, Davina Zamanzadeh, and Majid Sarrafzadeh. Cost-Sensitive Diagnosis and Learning Leveraging Public Health Data. *arXiv preprint arXiv:1902.07102*, 2019b.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

Chohee Kim, Mihaela Van Der Schaar, and Changhee Lee. Discovering Features with Synergistic Interactions in Multiple Views. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yuta Kobayashi, Zilin Jing, Jiayu Yao, Hongseok Namkoong, and Shalmali Joshi. Learning-To-Measure: In-Context Active Feature Acquisition. In *Forty-third International Conference on Machine Learning*, 2026.

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

Yang Li and Junier Oliva. Active Feature Acquisition with Generative Surrogate Models. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Yang Li, Shoaib Akbar, and Junier Oliva. ACFlow: Flow Models for Arbitrary Conditional Likelihoods. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

William J. McGill. Multivariate Information Transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Alexander Luke Ian Norcliffe, Changhee Lee, Fergus Imrie, Mihaela van der Schaar, and Pietro Lio. Stochastic Encodings for Active Feature Acquisition. In *Forty-second International Conference on Machine Learning*, 2025.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, 7(1):11479, 2016.

Byron Roe. MiniBooNE particle identification. UCI Machine Learning Repository, 2005. DOI: https://doi.org/10.24432/C5QC87.

Byron P. Roe, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543:577–584, 2005.

Gian-Carlo Rota. On the foundations of combinatorial theory I. Theory of Möbius Functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2:340–368, 1964.

Thomas Rückstieß, Christian Osendorfer, and Patrick van der Smagt. Minimizing Data Consumption with Sequential Online Feature Selection. *International Journal of Machine Learning and Cybernetics*, 4(3):235–243, 2013.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active Feature-Value Acquisition. *Management Science*, 55(4):664–684, 2009.

Athanasios Tsanas and Max Little. Parkinsons Telemonitoring. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C5ZS3N.

Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012a.

Athanasios Tsanas and Angeliki Xifara. Energy Efficiency. UCI Machine Learning Repository, 2012b. DOI: https://doi.org/10.24432/C51307.

Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, 2010.

Hado van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep Reinforcement Learning and the Deadly Triad. *arXiv preprint arXiv:1812.02648*, 2018.

Jan van Rijn. Kin8nm. OpenML, 2014. URL https://www.openml.org/d/189. Accessed: 2026-09-25.

Saverio Vito. Air Quality. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C59K5F.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature Selection using Stochastic Gates. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. INVASE: Instance-wise Variable Selection using Neural Networks. In *International Conference on Learning Representations*, 2019.

APPENDIX

A Additional Interpretations of γ-CMI **14**

A.1 Connection to Reinforcement Learning 14

A.2 Connection to Cooperative Game Theory 14

B Generative Model Error Analysis **17**

C Derivations and Proofs **18**

C.1 Derivation of Equation 4 18

C.2 Proof of Proposition 3.1 18

C.3 Proof of Corollary A.1 19

C.4 Proof of Proposition B.1 20

C.5 Proof of Corollary B.2 20

D Additional Synthetic Experiments **21**

D.1 Acquisition Trajectories on the Indicator and XOR Problems 21

D.2 Cross-Branch Correlation and Switch-Feature Redundancy 22

D.3 Additional Analysis of Redundancy 22

D.4 Additional Analysis of Higher-Order Synergy 23

E Dataset Details **24**

E.1 Real Datasets 24

E.2 Synthetic Datasets 25

F γ-CMI Extensions of Baselines **25**

F.1 γ-CMI Extension of ACFlow 26

F.2 γ-CMI Extension of EDDI 26

G Amortized CMI-Based Models **29**

H CMI Estimation for Regression Baselines **30**

I Predictive Performance Across γ **31**

I.1 Performance Across γ on Gaussian Synthetic Dataset 31

I.2 Performance Across γ on Real-World Datasets 31

J Inference-Time Overhead **33**

A Additional Interpretations of γ-CMI

The decomposition of the γ-CMI acquisition score developed in the main text admits two complementary interpretations. First, its γ-weighted interaction structure is closely related to discounted returns in reinforcement learning. Second, its averaging over randomly formed feature subsets admits a natural interpretation in cooperative game theory.

A.1 Connection to Reinforcement Learning

The γ-CMI acquisition score has an algebraic structure analogous to the discounted return $G_t = R_{t+1} + \gamma G_{t+1}$ in reinforcement learning when expressed recursively.

Corollary A.1. *Define the interaction return that accumulates all contributions from interaction depth h onward as*

$$G_h(i) = \sum_{k=0}^{|\mathcal{U}|-h-1} \gamma^k A_{h+k+1}(i), \qquad h = 0, \dots, |\mathcal{U}| - 1. \tag{9}$$

Then, the following recursive relation holds:

$$G_h(i) = A_{h+1}(i) + \gamma G_{h+1}(i). \tag{10}$$

The proof of Corollary A.1 is provided in Appendix C.3.

In particular, when $h = 0$, $G_0(i)$ corresponds to the γ-CMI acquisition score, so $R_\gamma(i, \mathbf{x}_\mathcal{O}) = G_0(i) = A_1(i) + \gamma G_1(i)$. Using $A_1(i) = R_G(i, \mathbf{x}_\mathcal{O})$, we obtain $R_\gamma(i, \mathbf{x}_\mathcal{O}) = R_G(i, \mathbf{x}_\mathcal{O}) + \gamma G_1(i)$.

In reinforcement learning, rewards obtained further into the future are weighted by higher powers of the discount factor. Analogously, in γ-CMI, interaction contributions that require a larger number of additional features to be observed are weighted by higher powers of γ. Under this interpretation, interaction depth in γ-CMI plays a role analogous to the time step in reinforcement learning, the interaction contribution corresponds to the reward, and γ plays the role of the discount factor.

A.2 Connection to Cooperative Game Theory

The same γ-CMI decomposition also admits a complementary interpretation from the perspective of cooperative game theory. The cooperative-game view interprets γ-CMI as averaging the marginal contribution of a candidate feature over randomly formed feature coalitions.

The set function $v(\mathcal{S})$, its Möbius coefficients $a(\mathcal{T})$, and the γ-CMI acquisition score introduced above can be naturally mapped to cooperative-game concepts. In particular, unobserved features can be viewed as players, subsets of features as coalitions, and $v(\mathcal{S})$ as the value assigned to a coalition. Under this correspondence, γ-CMI becomes a marginal contribution measure over feature coalitions, which allows us to connect it to probabilistic values, Harsanyi dividends, the Banzhaf value, and the Shapley value.

A.2.1 Feature Acquisition as a Cooperative Game

In cooperative game theory, a subset $\mathcal{S}$ of players is referred to as a coalition, and its value is represented by a characteristic function $v(\mathcal{S})$. In the active feature acquisition setting, this gives the following correspondence:

- unobserved feature: player
- feature index set $\mathcal{S}$: coalition
- $v(\mathcal{S})$ (information provided by $\mathcal{S}$): value of the coalition

More specifically, for a fixed observation $\mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}$, we define $v(\mathcal{S}) = I(Y; \mathbf{X}_\mathcal{S} \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O})$. Thus, $v(\mathcal{S})$ quantifies the conditional information about the prediction target Y provided by the features in coalition $\mathcal{S}$. Under this interpretation, active feature acquisition can be viewed as a cooperative game in which unobserved features act as players and the information carried by a feature coalition determines its value.

A.2.2 γ-CMI AS A PROBABILISTIC VALUE

For a candidate feature X_i, the increase in coalition value obtained by adding i to an existing coalition $\mathcal{S}$ is defined as its marginal contribution,

$$\Delta_i v(\mathcal{S}) := v(\mathcal{S} \cup \{i\}) - v(\mathcal{S}).$$

By the chain rule of conditional mutual information,

$$\Delta_i v(\mathcal{S}) = I(Y; X_i \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}}, \mathbf{X}_{\mathcal{S}}).$$

Therefore, the marginal contribution of candidate feature X_i to coalition $\mathcal{S}$ is exactly its CMI conditioned on the features already contained in $\mathcal{S}$. Greedy CMI evaluates only the marginal contribution at the empty coalition,

$$R_0(i, \mathbf{x}_{\mathcal{O}}) = \Delta_i v(\varnothing).$$

In contrast, γ-CMI averages this marginal contribution over randomly formed coalitions:

$$R_\gamma(i, \mathbf{x}_{\mathcal{O}}) = \mathbb{E}_{\phi \sim p_\gamma} \left[\Delta_i v(\phi) \right].$$

Here, p_γ denotes the distribution induced by independently including each unobserved feature except for the candidate i in the coalition with probability γ. Thus, rather than evaluating the contribution of X_i with respect to a single coalition, γ-CMI measures its expected marginal contribution over coalitions generated according to p_γ.

In cooperative game theory, values obtained by averaging a player's marginal contribution over a probability distribution on coalitions are commonly referred to as probabilistic values. From this perspective, γ-CMI can be interpreted as a probabilistic marginal contribution measure parameterized by the coalition inclusion probability γ.

A.2.3 CONNECTION TO HARSANYI DIVIDENDS

Recall that the Möbius decomposition expresses the coalition value as

$$v(\mathcal{S}) = \sum_{\mathcal{T} \subseteq \mathcal{S}} a(\mathcal{T}).$$

In cooperative game theory, the Möbius coefficient $a(\mathcal{T})$ corresponds to the Harsanyi dividend of coalition $\mathcal{T}$. It represents the contribution that is uniquely attributable to $\mathcal{T}$ after removing the contributions already explained by its proper subcoalitions.

From the decomposition derived in the main text,

$$R_\gamma(i, \mathbf{x}_{\mathcal{O}}) = \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \\ i \in \mathcal{T}}} \gamma^{|\mathcal{T}|-1} a(\mathcal{T}).$$

Therefore, γ-CMI can be interpreted as aggregating the Harsanyi dividends of all coalitions containing candidate feature i, while weighting each coalition $\mathcal{T}$ by $\gamma^{|\mathcal{T}|-1}$.

For the interaction contribution $a(\mathcal{T})$ to affect the acquisition score of i, all $|\mathcal{T}|-1$ features in $\mathcal{T} \setminus \{i\}$ must be included in the conditioning coalition. Since each feature is sampled independently with probability γ, this event occurs with probability $\gamma^{|\mathcal{T}|-1}$. Consequently, γ directly controls the extent to which higher-order coalition interactions contribute to the acquisition score.

A.2.4 Connections to the Banzhaf and Shapley Values

Banzhaf Value. Consider the special case $\gamma = \frac{1}{2}$. For every $\mathcal{S} \subseteq \mathcal{U} \setminus \{i\}$,

$$p_{1/2}(\mathcal{S}) = \left(\frac{1}{2}\right)^{|\mathcal{S}|} \left(\frac{1}{2}\right)^{|\mathcal{U}|-1-|\mathcal{S}|} = \frac{1}{2^{|\mathcal{U}|-1}}.$$

Hence, all possible coalitions are sampled with equal probability, and the γ-CMI score becomes

$$R_{1/2}(i, \mathbf{x}_{\mathcal{O}}) = \frac{1}{2^{|\mathcal{U}|-1}} \sum_{\mathcal{S} \subseteq \mathcal{U} \setminus \{i\}} \left[v(\mathcal{S} \cup \{i\}) - v(\mathcal{S}) \right].$$

This is exactly the uniform average of the marginal contribution of player i over all coalitions, which corresponds to the Banzhaf value. Therefore,

$$R_{1/2}(i, \mathbf{x}_{\mathcal{O}}) = \mathrm{Banzhaf}_i(v).$$

Thus, $\gamma = 1/2$ is a distinguished case of γ-CMI in which every possible conditioning coalition receives equal probability.

Shapley Value. We next consider averaging the γ-CMI score uniformly over all $\gamma \in [0, 1]$. Using the Harsanyi-dividend representation,

$$\begin{aligned}
\int_0^1 R_\gamma(i, \mathbf{x}_{\mathcal{O}})\, d\gamma &= \int_0^1 \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \\ i \in \mathcal{T}}} \gamma^{|\mathcal{T}|-1} a(\mathcal{T})\, d\gamma \\
&= \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \\ i \in \mathcal{T}}} a(\mathcal{T}) \int_0^1 \gamma^{|\mathcal{T}|-1}\, d\gamma \\
&= \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \\ i \in \mathcal{T}}} \frac{a(\mathcal{T})}{|\mathcal{T}|}.
\end{aligned}$$

The Shapley value admits the Harsanyi-dividend representation

$$\mathrm{Shapley}_i(v) = \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \\ i \in \mathcal{T}}} \frac{a(\mathcal{T})}{|\mathcal{T}|}.$$

Therefore,

$$\int_0^1 R_\gamma(i, \mathbf{x}_{\mathcal{O}})\, d\gamma = \mathrm{Shapley}_i(v).$$

This shows that, for each fixed value of γ, γ-CMI defines a probabilistic marginal contribution measure corresponding to a particular coalition inclusion probability, whereas averaging γ-CMI over the entire interval $[0, 1]$ recovers the Shapley value. These relationships can be summarized as

$$\begin{aligned}
\gamma = 0 &: \quad \text{Greedy CMI,} \\
0 < \gamma < 1 &: \quad \text{Probabilistic marginal contribution,} \\
\gamma = \frac{1}{2} &: \quad \text{Banzhaf value,} \\
\int_0^1 R_\gamma(i, \mathbf{x}_{\mathcal{O}})\, d\gamma &: \quad \text{Shapley value.}
\end{aligned}$$

Overall, γ-CMI can be viewed as a family of contribution measures that evaluates the expected marginal information of a candidate feature while controlling the probability with which other features participate in the conditioning coalition. From the cooperative-game perspective, γ therefore determines how strongly contributions arising from larger and higher-order feature coalitions are emphasized.

B Generative Model Error Analysis

Similar to the Möbius decomposition in Section 3.2, we analyze this error according to interaction order. Let $\eta_i(\mathcal{T})$ denote the generative model error associated with candidate feature X_i and an unobserved feature index subset $\mathcal{T} \subseteq \mathcal{U} \setminus \{i\}$. We define the aggregate r-th order generative model error as

$$E_r(i) = \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \setminus \{i\} \\ |\mathcal{T}| = r-1}} \eta_i(\mathcal{T}). \tag{11}$$

The resulting error in the γ-CMI acquisition score admits the following decomposition.

Proposition B.1. *Let $R_\gamma(i, \mathbf{x}_\mathcal{O})$ denote the oracle γ-CMI score and $\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O})$ denote the score computed using a learned generative model. Then,*

$$\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} E_r(i). \tag{12}$$

The proof is provided in Appendix C.4.

This decomposition shows that generative model errors are reflected in the acquisition score with the same γ-dependent weighting structure as the interaction contributions in Proposition 3.1. Therefore, as γ increases, γ-CMI can more strongly incorporate higher-order interactions between a candidate feature and other unobserved features, while errors arising from these higher-order terms are also reflected more strongly in the acquisition score.

From the perspective of feature acquisition, however, the primary objective is not necessarily to estimate CMI perfectly, but to correctly rank candidate features. Thus, even in the presence of estimation error, the acquisition decision may remain unchanged as long as the relative ordering of candidate scores is preserved. Suppose that the r-th order error satisfies $|E_r(i)| \leq \delta_r(i)$. Then,

$$\left| \widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) \right| \leq B_i(\gamma) := \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} \delta_r(i). \tag{13}$$

Let i^*_γ denote the optimal candidate feature index under the oracle γ-CMI score, and define the oracle score margin between the best and second-best candidates as $M_\gamma = R_\gamma(i^*_\gamma, \mathbf{x}_\mathcal{O}) - \max_{j \neq i^*_\gamma} R_\gamma(j, \mathbf{x}_\mathcal{O})$. The following result characterizes when the acquisition decision is preserved.

Corollary B.2. *Suppose that there exists an upper bound $B(\gamma)$ such that $B_i(\gamma) \leq B(\gamma)$ for every candidate $i \in \mathcal{U}$. If $M_\gamma > 2B(\gamma)$, then*

$$\arg\max_{i \in \mathcal{U}} \widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) = i^*_\gamma. \tag{14}$$

The proof is provided in Appendix C.5.

This result implies that, even in the presence of generative model error, the acquisition decision is preserved as long as the induced score perturbation is sufficiently small relative to the oracle margin between candidate features. Thus, γ-CMI can tolerate a certain level of generative model error without changing its acquisition decision. However, sufficiently large generative model errors may alter the candidate ranking.

C Derivations and Proofs

C.1 Derivation of Equation 4

Using the definition of conditional interaction information, we obtain

$$\begin{aligned} R_\gamma(i, \mathbf{x}_\mathcal{O}) &= \mathbb{E}_{\phi\sim p_\gamma}\mathbb{E}_{\mathbf{x}_\phi\sim p(\mathbf{X}_\phi|\mathbf{x}_\mathcal{O})}\left[I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\phi = \mathbf{x}_\phi)\right] \\ &= \mathbb{E}_{\phi\sim p_\gamma}\left[I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\phi)\right] \\ &= \mathbb{E}_{\phi\sim p_\gamma}\left[I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) + J(Y; X_i; \mathbf{X}_\phi \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O})\right] \\ &= I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) + \mathbb{E}_{\phi\sim p_\gamma}\left[J(Y; X_i; \mathbf{X}_\phi \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O})\right] \\ &= R_G(i, \mathbf{x}_\mathcal{O}) + \mathbb{E}_{\phi\sim p_\gamma}\left[J(Y; X_i; \mathbf{X}_\phi \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O})\right]. \end{aligned} \tag{15}$$

C.2 Proof of Proposition 3.1

Proof. Let $v(\mathcal{S}) := I(Y; \mathbf{X}_\mathcal{S} \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O})$ denote the information gain associated with a feature subset $\mathcal{S}$. By the Möbius decomposition, $v(\mathcal{S}) = \sum_{\mathcal{T}\subseteq\mathcal{S}} a(\mathcal{T})$.

For a candidate feature X_i, let $\mathcal{N}_i := \mathcal{U} \setminus \{i\}$, where $\mathcal{U}$ denotes the current set of unobserved feature indices. Under γ-CMI, each $j \in \mathcal{N}_i$ is independently included in the conditioning set Φ with probability γ. Hence, the probability that the conditioning set is exactly $\mathcal{S} \subseteq \mathcal{N}_i$ is $p_\gamma(\Phi = \mathcal{S}) = \gamma^{|\mathcal{S}|}(1-\gamma)^{|\mathcal{U}|-1-|\mathcal{S}|}$. Expanding the γ-CMI acquisition score over all possible conditioning sets gives

$$R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{\mathcal{S}\subseteq\mathcal{N}_i} p_\gamma(\Phi = \mathcal{S}) I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\mathcal{S}). \tag{16}$$

By the chain rule of conditional mutual information,

$$\begin{aligned} v(\mathcal{S}\cup\{i\}) &= I(Y; \mathbf{X}_{\mathcal{S}\cup\{i\}} \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) \\ &= I(Y; \mathbf{X}_\mathcal{S} \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) + I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\mathcal{S}) \\ &= v(\mathcal{S}) + I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\mathcal{S}). \end{aligned} \tag{17}$$

Therefore,

$$I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\mathcal{S}) = v(\mathcal{S}\cup\{i\}) - v(\mathcal{S}). \tag{18}$$

Thus, the CMI of candidate feature i can be interpreted as the increment in information obtained by adding i to the current conditioning set $\mathcal{S}$. Substituting (18) into (16) yields

$$R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{\mathcal{S}\subseteq\mathcal{N}_i} p_\gamma(\Phi = \mathcal{S}) \left[v(\mathcal{S}\cup\{i\}) - v(\mathcal{S})\right]. \tag{19}$$

By the Möbius decomposition,

$$v(\mathcal{S}\cup\{i\}) = \sum_{\mathcal{T}\subseteq\mathcal{S}\cup\{i\}} a(\mathcal{T}), \qquad v(\mathcal{S}) = \sum_{\mathcal{T}\subseteq\mathcal{S}} a(\mathcal{T}). \tag{20}$$

Taking their difference cancels all contributions that do not contain i, leaving

$$v(\mathcal{S}\cup\{i\}) - v(\mathcal{S}) = \sum_{\substack{\mathcal{T}\ni i \\ \mathcal{T}\setminus\{i\}\subseteq\mathcal{S}}} a(\mathcal{T}). \tag{21}$$

For an interaction term $a(\mathcal{T})$ to appear in (21), all elements of $\mathcal{T}\setminus\{i\}$ must be included in the conditioning set. Since each feature is independently included in Φ with probability γ,

$$\begin{aligned} \Pr(\mathcal{T}\setminus\{i\} \subseteq \Phi) &= \prod_{j\in\mathcal{T}\setminus\{i\}} \Pr(j \in \Phi) \\ &= \prod_{j\in\mathcal{T}\setminus\{i\}} \gamma \\ &= \gamma^{|\mathcal{T}|-1}. \end{aligned} \tag{22}$$

Hence, an interaction contribution $a(\mathcal{T})$ of order $|\mathcal{T}|$ enters the γ-CMI acquisition score with weight $\gamma^{|\mathcal{T}|-1}$. Summing over all interaction subsets containing i gives

$$R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \\ i \in \mathcal{T}}} \gamma^{|\mathcal{T}|-1} a(\mathcal{T}). \tag{23}$$

Recall that the total r-th order interaction contribution involving candidate feature i is defined as

$$A_r(i) := \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \\ i \in \mathcal{T},\ |\mathcal{T}|=r}} a(\mathcal{T}). \tag{24}$$

Grouping the terms in (23) by interaction order therefore gives

$$\begin{aligned} R_\gamma(i, \mathbf{x}_\mathcal{O}) &= \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} A_r(i) \\ &= A_1(i) + \gamma A_2(i) + \gamma^2 A_3(i) + \cdots + \gamma^{|\mathcal{U}|-1} A_{|\mathcal{U}|}(i). \end{aligned} \tag{25}$$

Finally, since $v(\emptyset) = 0$, the Möbius coefficient associated with the empty set is zero. Hence, the first-order contribution satisfies

$$\begin{aligned} A_1(i) &= a(\{i\}) \\ &= v(\{i\}) \\ &= I(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}). \end{aligned} \tag{26}$$

This is precisely the γ-CMI score at $\gamma = 0$, where no additional unobserved feature is included in the conditioning set, and therefore coincides with the Greedy CMI acquisition score. That is, $A_1(i) = R_0(i, \mathbf{x}_\mathcal{O}) = R_G(i, \mathbf{x}_\mathcal{O})$. Consequently,

$$R_\gamma(i, \mathbf{x}_\mathcal{O}) = R_G(i, \mathbf{x}_\mathcal{O}) + \gamma A_2(i) + \gamma^2 A_3(i) + \cdots + \gamma^{|\mathcal{U}|-1} A_{|\mathcal{U}|}(i). \tag{27}$$

Thus, γ-CMI can be interpreted as an acquisition criterion that retains the first-order information contribution of Greedy CMI while incorporating each r-th order interaction contribution with weight γ^{r-1}. □

C.3 Proof of Corollary A.1

Proof. Recall the interaction return that accumulates all contributions from interaction depth h onward:

$$G_h(i) = \sum_{k=0}^{|\mathcal{U}|-h-1} \gamma^k A_{h+k+1}(i), \qquad h = 0, \ldots, |\mathcal{U}| - 1. \tag{28}$$

Rearranging the terms gives

$$\begin{aligned} G_h(i) &= A_{h+1}(i) + \gamma A_{h+2}(i) + \gamma^2 A_{h+3}(i) + \cdots \\ &= A_{h+1}(i) + \gamma \left(A_{h+2}(i) + \gamma A_{h+3}(i) + \cdots \right) \\ &= A_{h+1}(i) + \gamma G_{h+1}(i). \end{aligned} \tag{29}$$

Therefore,

$$G_h(i) = A_{h+1}(i) + \gamma G_{h+1}(i). \tag{30}$$

□

C.4 Proof of Proposition B.1

Proof. For a subset of unobserved feature indices $\mathcal{S} \subseteq \mathcal{U} \setminus \{i\}$, let $g_i(\mathcal{S})$ denote the CMI computed using the oracle generative model, and let $\widehat{g}_i(\mathcal{S})$ denote the corresponding CMI computed using the learned generative model. The resulting γ-CMI acquisition scores are

$$R_\gamma(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi \sim p_\gamma}\left[g_i(\phi)\right], \qquad \widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi \sim p_\gamma}\left[\widehat{g}_i(\phi)\right]. \tag{31}$$

Define the generative-model-induced estimation error as $e_i(\mathcal{S}) := \widehat{g}_i(\mathcal{S}) - g_i(\mathcal{S})$. Then,

$$\begin{aligned}\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) &= \mathbb{E}_{\phi \sim p_\gamma}\left[\widehat{g}_i(\phi) - g_i(\phi)\right] \\ &= \mathbb{E}_{\phi \sim p_\gamma}\left[e_i(\phi)\right]. \end{aligned} \tag{32}$$

By the Möbius decomposition of the generative model error,

$$e_i(\Phi) = \sum_{\mathcal{T} \subseteq \Phi} \eta_i(\mathcal{T}). \tag{33}$$

Substituting (33) into (32) gives

$$\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\Phi \sim p_\gamma}\left[\sum_{\mathcal{T} \subseteq \Phi} \eta_i(\mathcal{T})\right]. \tag{34}$$

Rearranging the expectation as a sum over all $\mathcal{T} \subseteq \mathcal{U} \setminus \{i\}$ yields

$$\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{\mathcal{T} \subseteq \mathcal{U} \setminus \{i\}} \Pr(\mathcal{T} \subseteq \Phi)\eta_i(\mathcal{T}). \tag{35}$$

Under p_γ, each feature in $\mathcal{U} \setminus \{i\}$ is independently included in Φ with probability γ. Thus, for any subset $\mathcal{T}$, $\Pr(\mathcal{T} \subseteq \Phi) = \gamma^{|\mathcal{T}|}$. It follows that

$$\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{\mathcal{T} \subseteq \mathcal{U} \setminus \{i\}} \gamma^{|\mathcal{T}|}\eta_i(\mathcal{T}). \tag{36}$$

Grouping the terms according to their subset size gives

$$\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \setminus \{i\} \\ |\mathcal{T}| = r-1}} \eta_i(\mathcal{T}). \tag{37}$$

By the definition

$$E_r(i) := \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \setminus \{i\} \\ |\mathcal{T}| = r-1}} \eta_i(\mathcal{T}), \tag{38}$$

we finally obtain

$$\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} E_r(i). \tag{39}$$

□

C.5 Proof of Corollary B.2

Proof. By Proposition B.1,

$$\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} E_r(i). \tag{40}$$

Since $|E_r(i)| \le \delta_r(i)$, the triangle inequality gives

$$\begin{aligned}\left|\widehat{R}_\gamma(i, \mathbf{x}_\mathcal{O}) - R_\gamma(i, \mathbf{x}_\mathcal{O})\right| &\le \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} |E_r(i)| \\ &\le \sum_{r=1}^{|\mathcal{U}|} \gamma^{r-1} \delta_r(i) \\ &= B_i(\gamma). \end{aligned} \tag{41}$$

Let $B(\gamma) := \max_{i \in \mathcal{U}} B_i(\gamma)$. Recall that $i^\star_\gamma := \arg\max_{i \in \mathcal{U}} R_\gamma(i, \mathbf{x}_\mathcal{O})$ and define the oracle score margin as $M_\gamma := R_\gamma(i^\star_\gamma, \mathbf{x}_\mathcal{O}) - \max_{j \neq i^\star_\gamma} R_\gamma(j, \mathbf{x}_\mathcal{O})$. For any $j \neq i^\star_\gamma$, (41) implies

$$\begin{aligned}\widehat{R}_\gamma(i^\star_\gamma, \mathbf{x}_\mathcal{O}) - \widehat{R}_\gamma(j, \mathbf{x}_\mathcal{O}) &\ge R_\gamma(i^\star_\gamma, \mathbf{x}_\mathcal{O}) - B_{i^\star_\gamma}(\gamma) - (R_\gamma(j, \mathbf{x}_\mathcal{O}) + B_j(\gamma)) \\ &\ge R_\gamma(i^\star_\gamma, \mathbf{x}_\mathcal{O}) - R_\gamma(j, \mathbf{x}_\mathcal{O}) - 2B(\gamma) \\ &\ge M_\gamma - 2B(\gamma). \end{aligned} \tag{42}$$

Therefore, if $M_\gamma > 2B(\gamma)$, then $\widehat{R}_\gamma(i^\star_\gamma, \mathbf{x}_\mathcal{O}) > \widehat{R}_\gamma(j, \mathbf{x}_\mathcal{O})$ for every $j \neq i^\star_\gamma$. Hence, the learned generative model preserves the oracle γ-CMI acquisition decision. □

D Additional Synthetic Experiments

D.1 Acquisition Trajectories on the Indicator and XOR Problems

We compare the acquisition trajectories of Greedy CMI and γ-CMI on the Indicator and XOR problems using VAE-MLP. In the Indicator problem, one of the features other than the last feature contains information about the label, while the last feature indicates the index of that label-bearing feature. Therefore, the optimal trajectory is to first acquire the indicator feature and then acquire the label-bearing feature identified by the indicator. In the XOR problem, the label is determined by the combination of the values of the first two features. Thus, the optimal trajectory is to prioritize these two features in the first two acquisition steps.

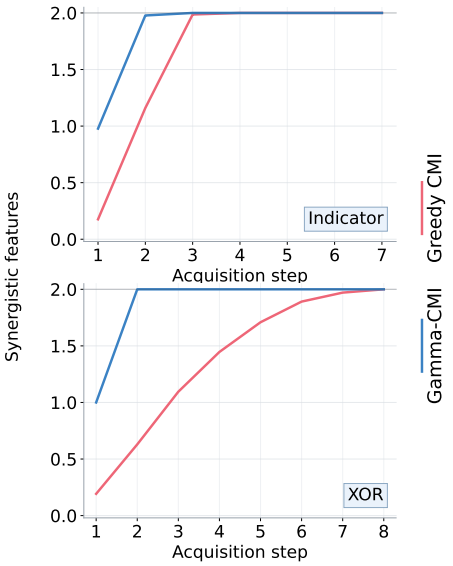


Figure D.1: Acquisition on the Indicator and XOR problems. The vertical axis shows the number of acquired features: the indicator and label-bearing features for Indicator and the first two features for XOR.

Both problems exhibit a synergistic structure in which important features are difficult to identify from their immediate individual information alone, and their predictive information becomes apparent only when multiple features are observed jointly. In the Indicator problem, observing only either the indicator feature or the label-bearing feature is insufficient to fully recover the label information. Similarly, in the XOR problem, observing only one of the two ground-truth features provides no information about the label. Consequently, Greedy CMI, which evaluates the individual CMI of each candidate feature under the current observation state, has difficulty prioritizing these ground-truth features during the early acquisition steps.

In contrast, γ-CMI evaluates the CMI of each candidate feature under hypothetical conditioning states in which other unobserved features are additionally observed. This allows the acquisition score to reflect the interaction contribution that emerges when the ground-truth features are jointly observed. As shown in Figure D.1, this difference enables γ-CMI to select the ground-truth features more frequently during the early acquisition steps than Greedy CMI.

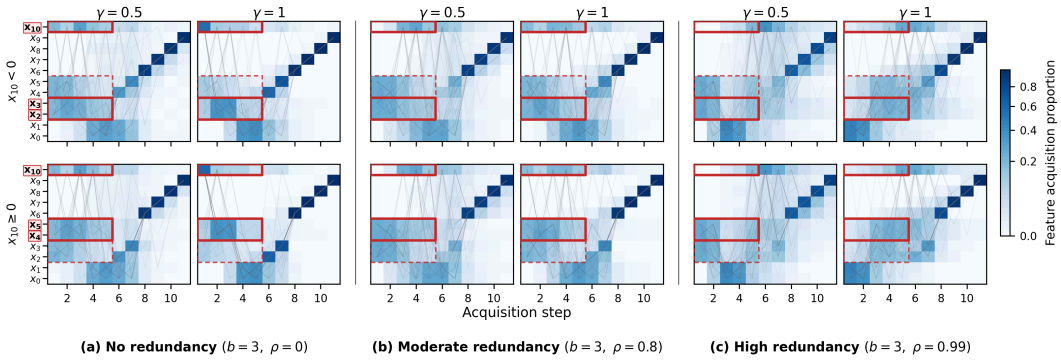


Figure D.2: Feature acquisition trajectories under cross-branch redundancy. Darker cells indicate higher feature acquisition proportions, and gray lines show representative trajectories. Over the first five acquisition steps, solid red boxes mark the active interaction pair and switch feature, whereas dashed red boxes mark the alternative interaction pair.

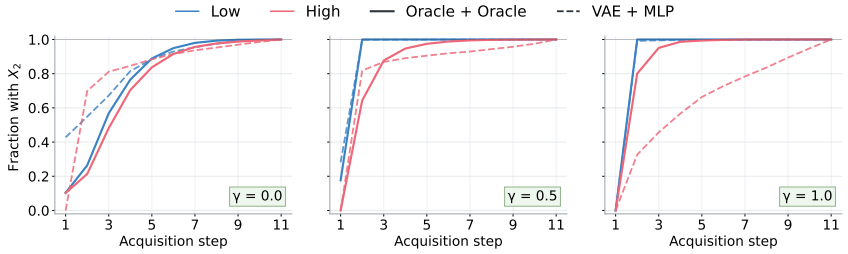


Figure D.3: Acquisition of a redundant feature on the Gaussian Synthetic Dataset across different redundancy levels and values of γ. The vertical axis shows the fraction of instances in which x_2 has been acquired.

D.2 Cross-Branch Correlation and Switch-Feature Redundancy

Using the Gaussian synthetic setting described in Section 5.1, we examine cross-branch redundancy by setting $b = 3$ and $\rho_{24} = \rho_{42} = \rho_{35} = \rho_{53} = \rho$. Unlike the within-pair correlation setting used for Observation 5.2, this correlates corresponding features from the two alternative interaction paths. As ρ increases, the information contributed by the two paths increasingly overlaps. Under larger γ, features from the alternative path are more likely to appear in the sampled future context, reducing the incremental information and γ-CMI scores of the active interaction features. Although the switch feature x_{10} is not directly correlated with these features, its predictive role is to distinguish which interaction path is active. Making the two paths redundant therefore also reduces the additional information provided by x_{10}, inducing structural redundancy in the switch feature.

Figure D.2 shows how this effect changes the acquisition order. At $\rho = 0$, $\gamma = 1$ assigns high early acquisition priority to both the switch feature x_{10} and the active interaction features. At $\rho = 0.8$, the interaction features remain concentrated in the early acquisition steps, whereas x_{10} loses its initial priority. At $\rho = 0.99$, x_{10} is delayed further. The interaction features remain prioritized early for $\gamma = 0.5$, whereas for $\gamma = 1$ they fall behind the individually informative features x_0 and x_1. Thus, strong cross-branch correlation, especially under larger γ, lowers the acquisition priority of both the interaction features and the switch feature.

D.3 Additional Analysis of Redundancy

To isolate the effect of redundancy, we set $a = 0$ and fix the synergy strength at $b = 10$. We then control the redundancy between x_2 and $x_3, \ldots, x_6$ by setting $\rho_{2j} = \rho$ for $j = 3, \ldots, 6$ and $\rho_{jk} = \rho^2$ for distinct $j, k \in \{3, \ldots, 6\}$, with $\rho \in \{0.2, 0.99\}$ corresponding to the low- and high-redundancy settings, respectively. When redundancy exists between a candidate feature and other

features, conditioning on additional unobserved features can reduce the incremental information provided by the candidate, thereby lowering its γ-CMI acquisition score.

As shown in Figure D.3, Greedy CMI, shown in the leftmost column, does not consider future context and therefore exhibits relatively little difference between the low- and high-redundancy settings. In contrast, under γ-CMI, features in the high-redundancy setting are consistently acquired later than their low-redundancy counterparts. This is consistent with nonzero γ allowing the acquisition score to account for information overlap with features included in the hypothetical future context.

D.4 Additional Analysis of Higher-Order Synergy

Using the γ-CMI score decomposition, we further analyze how the interaction order and the presence of individually informative competing features affect the acquisition behavior of γ-CMI on the Indicator Tree dataset. Consider a simplified setting in which a feature index set $\mathcal{T}^*$ forms a pure k-th order synergistic interaction, with $|\mathcal{T}^*| = k$ and $a(\mathcal{T}^*) = b > 0$. We assume that all other information contributions involving a candidate feature X_i with $i \in \mathcal{T}^*$ are zero, i.e., $a(\mathcal{T}) = 0$ for all $\mathcal{T} \ni i$ with $\mathcal{T} \neq \mathcal{T}^*$. Under this assumption, the γ-CMI acquisition score of feature X_i reduces to

$$R_\gamma(i, \mathbf{x}_\mathcal{O}) = \sum_{\substack{\mathcal{T} \subseteq \mathcal{U} \\ i \in \mathcal{T}}} \gamma^{|\mathcal{T}|-1} a(\mathcal{T}) = b\gamma^{k-1}. \tag{43}$$

Thus, the contribution of a k-th order synergistic interaction is weighted by γ^{k-1}. This weighting corresponds to the probability that all $k-1$ feature indices in $\mathcal{T}^* \setminus \{i\}$ are included in the randomly sampled future context, since each unobserved feature except the candidate is independently included with probability γ. For $\gamma < 1$, the contribution is increasingly attenuated as the interaction order k increases. Conversely, increasing γ reduces this attenuation, allowing higher-order interactions to exert a stronger influence on the acquisition score.

We next examine how this property affects the acquisition priority between synergistic features and competing features that provide immediate information about the target. In the Indicator Tree Dataset, eight additional features provide individual information about the label. To characterize the relative acquisition priority, consider an individually informative competing feature X_j with an individual information contribution $a(\{j\}) = c > 0$. We assume that all higher-order information contributions involving j are zero, i.e., $a(\mathcal{T}) = 0$ for all $\mathcal{T} \ni j$ with $|\mathcal{T}| \geq 2$. Under this assumption, the acquisition score of feature X_j is independent of γ, yielding $R_\gamma(j, \mathbf{x}_O) = c$. In contrast, the acquisition score of the synergistic feature X_i is $R_\gamma(i, \mathbf{x}_\mathcal{O}) = b\gamma^{k-1}$. Therefore, feature X_i receives a higher acquisition score than feature X_j if and only if $b\gamma^{k-1} > c$, or equivalently, $\gamma > (c/b)^{1/(k-1)}$. For $0 < c < b$ and $k \geq 2$, the critical value at which the two acquisition scores are equal is given by

$$\gamma^* = \left(\frac{c}{b}\right)^{\frac{1}{k-1}}. \tag{44}$$

This expression reveals two properties. First, for fixed b and k, the critical value γ^* increases with c. Hence, stronger individual information from a competing feature requires a larger γ for the synergistic feature to receive a higher acquisition score. Second, for fixed b and c, γ^* increases with the interaction order k. Thus, even when the total synergistic contribution remains unchanged, interactions involving more features require a larger γ to outweigh an individually informative competing feature. If $c \geq b$, the synergistic feature cannot receive a strictly higher score than the competing feature for any $\gamma \in [0, 1]$ under these assumptions.

These results provide an interpretation of the acquisition behavior observed in Figure 3. When γ is small, higher-order interaction contributions are strongly attenuated, which can favor features that provide immediate information. As γ increases, synergistic features receive greater weight from their interaction contributions and can gain priority over individually informative features. The derived threshold further shows that this change in relative priority depends jointly on the interaction order and the strength of the competing individual information.

E Dataset Details

This section provides additional details on the datasets used in our experiments that were omitted from the main text. We describe the prediction task, feature composition, preprocessing procedure, and the sizes of the training, validation, and test sets.

E.1 Real Datasets

Bank Marketing. Bank Marketing (Moro et al., 2014) is a binary classification dataset where the goal is to predict whether a client who received a marketing call from a bank subscribes to a term deposit. The label is set to 1 if the client subscribes and 0 otherwise. The dataset contains 15 features in total, consisting of 7 continuous and 8 categorical features. We split the full dataset of 45,211 samples into training, validation, and test sets with an 8:1:1 ratio, resulting in 36,168, 4,521, and 4,522 samples, respectively.

California Housing. California Housing (Kelley Pace & Barry, 1997) is a dataset for predicting house prices in California districts. We use this dataset for both classification and regression tasks. For classification, the continuous house-price target is discretized into four equally sized bins, resulting in a 4-class classification problem. The classification setting uses 8 continuous features, whereas the regression setting additionally includes one categorical feature, resulting in 9 features in total. We split the dataset into training, validation, and test sets with an 8:1:1 ratio, yielding 16,512, 2,064, and 2,064 samples, respectively.

MiniBooNE. MiniBooNE (Roe et al., 2005; Roe, 2005) is a binary classification dataset derived from the Fermilab MiniBooNE experiment, where the task is to distinguish signal events from background events. Label 1 denotes a signal event corresponding to a muon neutrino oscillating into an electron neutrino, while label 0 denotes a background event. The original dataset exhibits substantial class imbalance, with approximately three times more negative than positive samples. To alleviate this imbalance, we retain all 36,499 positive samples and randomly subsample 40,000 negative samples. In addition, among the original 50 continuous detector features, we select 20 features using XGBoost-based feature importance. The resulting training set contains 56,499 samples, while the validation and test sets each contain 10,000 samples.

MNIST and Fashion-MNIST. MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) are 10-class image classification datasets consisting of handwritten digits and clothing items, respectively. To reduce computational cost, we apply the feature selection method STG (Yamada et al., 2020) to the original 784 pixels and select 20 pixels as input features. For both datasets, we use 50,000 training samples, 10,000 validation samples, and 10,000 test samples.

METABRIC. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset (Curtis et al., 2012; Pereira et al., 2016) consists of gene-expression measurements from 1,898 breast cancer patients and is used to predict tumor molecular subtypes in a 6-class classification task. We use 12 continuous gene-expression features. The training, validation, and test sets contain 1,518, 189, and 191 samples, respectively.

Air Quality. Air Quality (De Vito et al., 2008; Vito, 2008) is a regression dataset where the goal is to predict the true carbon monoxide concentration (CO) from air-quality sensor measurements and meteorological variables. We use 12 continuous features. In the original dataset, missing values are encoded as -200. Samples with a missing CO target are removed, while missing feature values are replaced with 0. After preprocessing, the training, validation, and test sets contain 6,139, 767, and 768 samples, respectively.

Energy Efficiency. Energy Efficiency (Tsanas & Xifara, 2012a;b) is a regression dataset for predicting the heating and cooling loads of buildings from their physical characteristics. The dataset consists of 8 continuous features and provides two target variables: heating load and cooling load. We predict the two targets with separate output heads and report the mean of their RMSE values. The training, validation, and test sets contain 614, 76, and 78 samples, respectively.

Kin8nm. Kin8nm (van Rijn, 2014) is a synthetic regression dataset where the goal is to predict the position of an 8-link robotic arm from its joint states. The input consists of 8 continuous features representing the joint configuration of the robot arm. The training, validation, and test sets contain 6,553, 819, and 820 samples, respectively.

Parkinsons Telemonitoring. Parkinsons Telemonitoring (Tsanas et al., 2010; Tsanas & Little, 2009) is a regression dataset where the goal is to predict clinical symptom scores, measured by the Unified Parkinson's Disease Rating Scale (UPDRS), from voice recordings of patients with Parkinson's disease. We exclude age, sex, and recording-time variables and use the remaining 16 continuous features. The original dataset provides two target variables, motor UPDRS and total UPDRS. We predict the two targets with separate output heads and report the mean of their RMSE values. The training, validation, and test sets contain 4,700, 587, and 588 samples, respectively.

White Wine Quality. Wine Quality (Cortez et al., 2009b;a) contains physicochemical measurements of red and white wines together with quality scores assigned by human experts. In our experiments, we use only the white wine subset. Each sample contains 11 continuous physicochemical features. Although the original quality score is an integer-valued ordinal variable, we treat it as a continuous target and formulate the task as regression to preserve the ordering and similarity between adjacent quality scores. The training, validation, and test sets contain 3,918, 489, and 491 samples, respectively.

E.2 Synthetic Datasets

Indicator Problem. The Indicator Problem is a synthetic dataset containing a dedicated indicator feature that specifies which feature determines the label (Norcliffe et al., 2025). The dataset contains 7 features in total. The first six features, $x_0, \ldots, x_5$, are binary variables, while the final feature x_6 acts as the indicator. If the indicator takes the value j, the label is defined as

$$y = x_j, \qquad j = x_6.$$

Thus, before observing the indicator, it is unknown which individual feature is relevant to the label. Once x_6 is observed, the feature that determines the label becomes known. The dataset contains 8,000 training samples and 1,000 samples each for validation and testing. The optimal acquisition strategy is to acquire the indicator x_6 first and then acquire the feature x_j specified by the indicator.

XOR Problem. The XOR Problem is a binary classification dataset designed to evaluate feature interactions in which individual features are uninformative about the label but become fully informative when observed jointly. Each feature is randomly sampled from $\{0, 1\}$, and the label is determined by the interaction between the first two features, x_0 and x_1, as

$$y = \begin{cases} 0, & x_0 = x_1, \\ 1, & x_0 \neq x_1. \end{cases}$$

Hence, neither x_0 nor x_1 alone provides information about the label, whereas observing both features completely determines it. The remaining features are independent noise features. As in the Indicator Problem, the dataset contains 8,000 training samples and 1,000 samples each for validation and testing. The optimal acquisition strategy is therefore to prioritize x_0 and x_1 over the remaining features.

F γ-CMI Extensions of Baselines

In this section, we show how γ-CMI can be incorporated into two existing CMI-based acquisition methods, ACFlow and EDDI. ACFlow directly models arbitrary conditional distributions and computes conditional mutual information as an expected KL divergence, whereas EDDI uses a Partial VAE to approximate information gain in the latent space.

For a candidate feature index $i \in \mathcal{U}$, let Φ denote a random subset of future unobserved features whose realization $\phi \subseteq \mathcal{U} \setminus \{i\}$ is sampled according to the γ sampling distribution. For convenience, we define the augmented conditioning context as $\mathbf{c}_\phi = (\mathbf{x}_\mathcal{O}, \mathbf{x}_\phi)$, which contains both the currently observed features and the sampled future features.

We next derive the corresponding γ-CMI acquisition estimators for ACFlow and EDDI.

F.1 γ-CMI Extension of ACFlow

ACFlow uses the conditional mutual information between a candidate X_i and the prediction target Y as its acquisition criterion. By the definition of conditional mutual information,

$$\begin{aligned} I(X_i; Y \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}) &= \mathbb{E}_{(x_i,y)\sim p(X_i,Y\mid \mathbf{x}_\mathcal{O})}\left[\log \frac{p(X_i, Y \mid \mathbf{x}_\mathcal{O})}{p(X_i \mid \mathbf{x}_\mathcal{O})p(Y \mid \mathbf{x}_\mathcal{O})}\right] \\ &= \mathbb{E}_{(x_i,y)\sim p(X_i,Y\mid \mathbf{x}_\mathcal{O})}\left[\log \frac{p(Y \mid x_i, \mathbf{x}_\mathcal{O})}{p(Y \mid \mathbf{x}_\mathcal{O})}\right]. \end{aligned}$$

Since ACFlow can directly model arbitrary conditional distributions, extending it to γ-CMI only requires replacing the original conditioning context $\mathbf{x}_\mathcal{O}$ with the augmented context $\mathbf{c}_\phi = (\mathbf{x}_\mathcal{O}, \mathbf{x}_\phi)$. For fixed $\Phi = \phi$ and $\mathbf{x}_\phi$,

$$\begin{aligned} &I\left(X_i; Y \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\phi = \mathbf{x}_\phi\right) \\ &\quad = \mathbb{E}_{(x_i,y)\sim p(X_i,Y\mid \mathbf{c}_\phi)}\left[\log \frac{p(Y \mid x_i, \mathbf{c}_\phi)}{p(Y \mid \mathbf{c}_\phi)}\right]. \end{aligned}$$

Taking an additional expectation over both the sampled subset Φ and its realization x_ϕ yields the γ-CMI acquisition score for ACFlow:

$$\begin{aligned} R_\gamma^{\mathrm{ACFlow}}(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi\sim p_\gamma}\mathbb{E}_{\mathbf{x}_\phi\sim p(\mathbf{X}_\phi\mid \mathbf{x}_\mathcal{O})}\mathbb{E}_{(x_i,y)\sim p(X_i,Y\mid \mathbf{c}_\phi)} \\ \left[\log \frac{p(Y \mid x_i, \mathbf{c}_\phi)}{p(Y \mid \mathbf{c}_\phi)}\right]. \end{aligned}$$

Using $p(\mathbf{X}_\phi \mid \mathbf{x}_\mathcal{O})p(X_i, Y \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi) = p(\mathbf{X}_\phi, X_i, Y \mid \mathbf{x}_\mathcal{O})$, the same score can equivalently be written using joint sampling as

$$\begin{aligned} R_\gamma^{\mathrm{ACFlow}}(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi\sim p_\gamma}\mathbb{E}_{(\mathbf{x}_\phi,x_i,y)\sim p(\mathbf{X}_\phi,X_i,Y\mid \mathbf{x}_\mathcal{O})} \\ \left[\log \frac{p(Y \mid x_i, \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi)}{p(Y \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi)}\right]. \end{aligned}$$

For classification, the conditional mutual information can be expressed, as in the original ACFlow formulation, as an expected KL divergence:

$$\begin{aligned} R_\gamma^{\mathrm{ACFlow}}(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi\sim p_\gamma}\mathbb{E}_{\mathbf{x}_\phi\sim p(\mathbf{X}_\phi\mid \mathbf{x}_\mathcal{O})}\mathbb{E}_{x_i\sim p(X_i\mid \mathbf{c}_\phi)} \\ \left[D_{\mathrm{KL}}\left(p(Y \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y \mid \mathbf{c}_\phi)\right)\right]. \end{aligned}$$

Therefore, applying γ-CMI to ACFlow amounts to augmenting the conditioning context from $\mathbf{x}_\mathcal{O}$ to $(\mathbf{x}_\mathcal{O}, \mathbf{x}_\phi)$ and introducing an additional expectation over the sampled future context.

When $\gamma = 0$, we have $\Phi = \varnothing$ and hence $\mathbf{c}_\phi = \mathbf{x}_\mathcal{O}$. The acquisition score therefore reduces to

$$R_0^{\mathrm{ACFlow}}(i, \mathbf{x}_\mathcal{O}) = I(X_i; Y \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}),$$

recovering the original ACFlow acquisition criterion.

F.2 γ-CMI Extension of EDDI

Unlike ACFlow, EDDI uses a Partial VAE to model distributions under arbitrary partial observations and evaluates candidate features through expected information gain about the prediction target. Rather than directly computing the corresponding KL divergence in the target space, EDDI transforms it into a tractable expression involving KL divergences between latent-variable distributions.

To extend EDDI to γ-CMI, we first establish the relationship between the original EDDI reward and conditional mutual information, and then generalize the reward to the augmented conditioning context containing sampled future features.

F.2.1 Relationship Between the EDDI Reward and Conditional Mutual Information

Using Y to denote the prediction target, the information reward used by EDDI can be written as

$$R_{\mathrm{EDDI}}(i, \mathbf{x}_{\mathcal{O}}) = \mathbb{E}_{x_i \sim p(X_i \mid \mathbf{x}_{\mathcal{O}})} \left[D_{\mathrm{KL}} \left(p(Y \mid x_i, \mathbf{x}_{\mathcal{O}}) \,\|\, p(Y \mid \mathbf{x}_{\mathcal{O}}) \right) \right].$$

This quantity is equivalent to the Greedy CMI $I(Y; X_i \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}})$. By definition,

$$I(Y; X_i \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}}) = \int p(Y, X_i \mid \mathbf{x}_{\mathcal{O}}) \log \frac{p(Y, X_i \mid \mathbf{x}_{\mathcal{O}})}{p(Y \mid \mathbf{x}_{\mathcal{O}}) p(X_i \mid \mathbf{x}_{\mathcal{O}})} \, dy \, dx_i.$$

Using the probability chain rule $p(Y, X_i \mid \mathbf{x}_{\mathcal{O}}) = p(X_i \mid \mathbf{x}_{\mathcal{O}}) p(Y \mid x_i, \mathbf{x}_{\mathcal{O}})$, we obtain

$$\begin{aligned} I(Y; X_i \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}}) &= \int p(Y, X_i \mid \mathbf{x}_{\mathcal{O}}) \log \frac{p(Y \mid x_i, \mathbf{x}_{\mathcal{O}})}{p(Y \mid \mathbf{x}_{\mathcal{O}})} \, dy \, dx_i \\ &= \int p(X_i \mid \mathbf{x}_{\mathcal{O}}) \left[\int p(Y \mid x_i, \mathbf{x}_{\mathcal{O}}) \log \frac{p(Y \mid x_i, \mathbf{x}_{\mathcal{O}})}{p(Y \mid \mathbf{x}_{\mathcal{O}})} \, dy \right] dx_i \\ &= \mathbb{E}_{x_i \sim p(X_i \mid \mathbf{x}_{\mathcal{O}})} \left[D_{\mathrm{KL}} \left(p(Y \mid x_i, \mathbf{x}_{\mathcal{O}}) \,\|\, p(Y \mid \mathbf{x}_{\mathcal{O}}) \right) \right]. \end{aligned}$$

It follows immediately that

$$R_{\mathrm{EDDI}}(i, \mathbf{x}_{\mathcal{O}}) = I(Y; X_i \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}}).$$

Thus, the original EDDI information reward corresponds exactly to the $\gamma = 0$ case of γ-CMI.

F.2.2 γ-CMI Extension of the EDDI Reward

We now incorporate the sampled future feature subset Φ and its realization $\mathbf{x}_\phi$ into the conditioning context. As above, let $\mathbf{c}_\phi = (\mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi)$. For fixed $\Phi = \phi$ and $\mathbf{x}_\phi$, the EDDI reward becomes

$$R_{\mathrm{EDDI}}(i, \mathbf{c}_\phi) = \mathbb{E}_{x_i \sim p(X_i \mid \mathbf{c}_\phi)} \left[D_{\mathrm{KL}} \left(p(Y \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y \mid \mathbf{c}_\phi) \right) \right].$$

Applying the same derivation as above to the augmented conditioning context gives

$$\begin{aligned} &I\left(Y; X_i \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}}, \mathbf{X}_\phi = \mathbf{x}_\phi\right) \\ &\quad = \mathbb{E}_{x_i \sim p(X_i \mid \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi)} \left[D_{\mathrm{KL}} \left(p(Y \mid x_i, \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi) \,\|\, p(Y \mid \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi) \right) \right]. \end{aligned}$$

Taking expectations over $\phi \sim p_\gamma$ and $\mathbf{x}_\phi \sim p(\mathbf{X}_\phi \mid \mathbf{x}_{\mathcal{O}})$ therefore yields

$$\begin{aligned} R_\gamma^{\mathrm{EDDI}}(i, \mathbf{x}_{\mathcal{O}}) &= \mathbb{E}_{\phi \sim p_\gamma} \mathbb{E}_{\mathbf{x}_\phi \sim p(\mathbf{X}_\phi \mid \mathbf{x}_{\mathcal{O}})} \mathbb{E}_{x_i \sim p(X_i \mid \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi)} \\ &\qquad \left[D_{\mathrm{KL}} \left(p(Y \mid x_i, \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi) \,\|\, p(Y \mid \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi) \right) \right] \\ &= \mathbb{E}_{\phi \sim p_\gamma} \mathbb{E}_{\mathbf{x}_\phi \sim p(\mathbf{X}_\phi \mid \mathbf{x}_{\mathcal{O}})} \left[I\left(Y; X_i \mid \mathbf{X}_{\mathcal{O}} = \mathbf{x}_{\mathcal{O}}, \mathbf{X}_\phi = \mathbf{x}_\phi\right) \right]. \end{aligned}$$

Hence, $R_\gamma^{\mathrm{EDDI}}(i, \mathbf{x}_{\mathcal{O}}) = R_\gamma(i, \mathbf{x}_{\mathcal{O}})$. Furthermore, using

$$p(\mathbf{X}_\phi \mid \mathbf{x}_{\mathcal{O}}) p(X_i \mid \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi) = p(\mathbf{X}_\phi, X_i \mid \mathbf{x}_{\mathcal{O}}),$$

the reward can equivalently be written as

$$\begin{aligned} R_\gamma^{\mathrm{EDDI}}(i, \mathbf{x}_{\mathcal{O}}) = \mathbb{E}_{\phi \sim p_\gamma} \mathbb{E}_{\mathbf{x}_{\phi \cup \{i\}} \sim p(\mathbf{X}_{\phi \cup \{i\}} \mid \mathbf{x}_{\mathcal{O}})} \\ \left[D_{\mathrm{KL}} \left(p(Y \mid x_i, \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi) \,\|\, p(Y \mid \mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi) \right) \right]. \end{aligned}$$

F.2.3 Latent-Space Formulation of the γ-CMI Extension of EDDI Reward

EDDI avoids directly computing the KL divergence in the target space by expressing it in terms of the latent variable $\mathbf{Z}$ of the Partial VAE. The original EDDI derivation applies the chain rule of KL divergence twice. The same argument can be applied to the augmented conditioning context $\mathbf{c}_\phi = (\mathbf{x}_{\mathcal{O}}, \mathbf{x}_\phi)$.

Let $\mathbf{Z}$ denote the latent variable of the Partial VAE, and assume the conditional independence

$$p(Y \mid \mathbf{z}, x_i, \mathbf{c}_\phi) = p(Y \mid \mathbf{z}).$$

That is, conditioned on the latent variable $\mathbf{Z}$, the target Y is independent of both the candidate feature and the current conditioning context. This corresponds to the factorized likelihood used by the Partial VAE decoder, in which individual variables are generated conditionally on $\mathbf{Z}$.

Consider the target-space divergence $D_{\mathrm{KL}}(p(Y \mid x_i, \mathbf{c}_\phi) \| p(Y \mid \mathbf{c}_\phi))$. Applying the KL chain rule to the joint distribution of $(Y, \mathbf{Z})$ gives

$$\begin{aligned} &D_{\mathrm{KL}}\left(p(Y, \mathbf{Z} \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y, \mathbf{Z} \mid \mathbf{c}_\phi)\right) \\ &\quad = D_{\mathrm{KL}}\left(p(Y \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y \mid \mathbf{c}_\phi)\right) \\ &\qquad + \mathbb{E}_{y \sim p(Y \mid x_i, \mathbf{c}_\phi)}\left[D_{\mathrm{KL}}\left(p(\mathbf{Z} \mid y, x_i, \mathbf{c}_\phi) \,\|\, p(\mathbf{Z} \mid y, \mathbf{c}_\phi)\right)\right]. \end{aligned}$$

Rearranging yields

$$\begin{aligned} &D_{\mathrm{KL}}\left(p(Y \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y \mid \mathbf{c}_\phi)\right) \\ &\quad = D_{\mathrm{KL}}\left(p(Y, \mathbf{Z} \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y, \mathbf{Z} \mid \mathbf{c}_\phi)\right) \\ &\qquad - \mathbb{E}_{y \sim p(Y \mid x_i, \mathbf{c}_\phi)}\left[D_{\mathrm{KL}}\left(p(\mathbf{Z} \mid y, x_i, \mathbf{c}_\phi) \,\|\, p(\mathbf{Z} \mid y, \mathbf{c}_\phi)\right)\right]. \end{aligned}$$

We can alternatively apply the KL chain rule to the same joint divergence by decomposing with respect to $\mathbf{Z}$ first:

$$\begin{aligned} &D_{\mathrm{KL}}\left(p(Y, \mathbf{Z} \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y, \mathbf{Z} \mid \mathbf{c}_\phi)\right) \\ &\quad = D_{\mathrm{KL}}\left(p(\mathbf{Z} \mid x_i, \mathbf{c}_\phi) \,\|\, p(\mathbf{Z} \mid \mathbf{c}_\phi)\right) \\ &\qquad + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{Z} \mid x_i, \mathbf{c}_\phi)}\left[D_{\mathrm{KL}}\left(p(Y \mid \mathbf{z}, x_i, \mathbf{c}_\phi) \,\|\, p(Y \mid \mathbf{z}, \mathbf{c}_\phi)\right)\right]. \end{aligned}$$

Under the conditional independence assumption, $p(Y \mid \mathbf{z}, x_i, \mathbf{c}_\phi) = p(Y \mid \mathbf{z}, \mathbf{c}_\phi) = p(Y \mid \mathbf{z})$, and therefore the second KL divergence vanishes. Hence,

$$D_{\mathrm{KL}}\left(p(Y, \mathbf{Z} \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y, \mathbf{Z} \mid \mathbf{c}_\phi)\right) = D_{\mathrm{KL}}\left(p(\mathbf{Z} \mid x_i, \mathbf{c}_\phi) \,\|\, p(\mathbf{Z} \mid \mathbf{c}_\phi)\right).$$

Substituting this identity into the previous expression gives

$$\begin{aligned} &D_{\mathrm{KL}}\left(p(Y \mid x_i, \mathbf{c}_\phi) \,\|\, p(Y \mid \mathbf{c}_\phi)\right) \\ &\quad = D_{\mathrm{KL}}\left(p(\mathbf{Z} \mid x_i, \mathbf{c}_\phi) \,\|\, p(\mathbf{Z} \mid \mathbf{c}_\phi)\right) \\ &\qquad - \mathbb{E}_{y \sim p(Y \mid x_i, \mathbf{c}_\phi)}\left[D_{\mathrm{KL}}\left(p(\mathbf{Z} \mid y, x_i, \mathbf{c}_\phi) \,\|\, p(\mathbf{Z} \mid y, \mathbf{c}_\phi)\right)\right]. \end{aligned}$$

Consequently, the γ-CMI extension of the EDDI reward admits the following latent-space representation:

$$\begin{aligned} R_\gamma^{\mathrm{EDDI}}(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi \sim p_\gamma} \mathbb{E}_{\mathbf{x}_\phi \sim p(\mathbf{X}_\phi \mid \mathbf{x}_\mathcal{O})} \mathbb{E}_{x_i \sim p(X_i \mid \mathbf{c}_\phi)} \Bigg[& D_{\mathrm{KL}}\left(p(\mathbf{Z} \mid x_i, \mathbf{c}_\phi) \,\|\, p(\mathbf{Z} \mid \mathbf{c}_\phi)\right) \\ & - \mathbb{E}_{y \sim p(Y \mid x_i, \mathbf{c}_\phi)}\left[D_{\mathrm{KL}}\left(p(\mathbf{Z} \mid y, x_i, \mathbf{c}_\phi) \,\|\, p(\mathbf{Z} \mid y, \mathbf{c}_\phi)\right)\right] \Bigg]. \end{aligned}$$

Thus, the latent-space derivation used by EDDI remains valid when the original conditioning context $\mathbf{x}_\mathcal{O}$ is replaced by the augmented context $\mathbf{c}_\phi = (\mathbf{x}_\mathcal{O}, \mathbf{x}_\phi)$.

F.2.4 Recovery of the Original EDDI Reward at $\gamma = 0$

Finally, we verify that the γ-CMI extension contains the original EDDI acquisition rule as a special case. When $\gamma = 0$, no future feature is sampled, so $\Phi = \varnothing$ and $\mathbf{c}_\phi = \mathbf{x}_\mathcal{O}$. The resulting estimator becomes

$$\begin{aligned} \widehat{R}_0^{\mathrm{EDDI}}(i, \mathbf{x}_\mathcal{O}) = &\mathbb{E}_{\hat{x}_i \sim \hat{p}(X_i \mid \mathbf{x}_\mathcal{O})}\left[D_{\mathrm{KL}}\left(q(\mathbf{Z} \mid \hat{x}_i, \mathbf{x}_\mathcal{O}) \,\|\, q(\mathbf{Z} \mid \mathbf{x}_\mathcal{O})\right)\right] \\ &- \mathbb{E}_{(\hat{x}_i, \hat{y}) \sim \hat{p}(X_i, Y \mid \mathbf{x}_\mathcal{O})}\left[D_{\mathrm{KL}}\left(q(\mathbf{Z} \mid \hat{y}, \hat{x}_i, \mathbf{x}_\mathcal{O}) \,\|\, q(\mathbf{Z} \mid \hat{y}, \mathbf{x}_\mathcal{O})\right)\right]. \end{aligned}$$

This is exactly the latent-space information reward of the original EDDI formulation when the prediction target is denoted by Y. In particular, the second KL term removes the residual latent information introduced by the candidate feature that is not attributable to the prediction target. Therefore,

$$\widehat{R}_0^{\mathrm{EDDI}}(i, \mathbf{x}_{\mathcal{O}}) = \widehat{R}_{\mathrm{EDDI}}(i, \mathbf{x}_{\mathcal{O}}).$$

In summary, both ACFlow and EDDI exactly recover their original Greedy CMI-based acquisition rules when $\gamma = 0$. For $\gamma > 0$, both methods naturally extend to γ-CMI by incorporating sampled future features into the conditioning context, thereby allowing the acquisition score to account for interactions with features that may be acquired in subsequent steps.

G Amortized CMI-Based Models

In this section, we describe DIME and GDFS, two amortized CMI-based approaches that avoid explicitly modeling the distribution of unobserved features with a generative model. These methods learn to approximate Greedy CMI acquisition behavior through amortized optimization. Throughout this section, let $\mathcal{S} \subseteq [d]$ denote an arbitrary subset of observed feature indices.

DIME. DIME consists of a predictor network $f(\mathbf{x}_{\mathcal{S}}; \theta)$, which makes predictions from an arbitrary observed feature set $\mathbf{x}_{\mathcal{S}}$, and a value network $v(\mathbf{x}_{\mathcal{S}}; \phi) \in \mathbb{R}^d$, which amortizes the estimation of the CMI associated with each unobserved feature. Specifically, the i-th output of the value network approximates the CMI of candidate feature X_i as $v_i(\mathbf{x}_{\mathcal{S}}; \phi) \approx I(Y; X_i \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})$.

The two networks are trained jointly but optimize different objectives. The predictor network is trained to minimize the prediction loss over randomly sampled observed feature subsets:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{X},Y)} \mathbb{E}_{\mathcal{S}} \left[\ell \left(f(\mathbf{x}_{\mathcal{S}}; \theta), y \right) \right].$$

The value network is trained using the reduction in prediction loss obtained by additionally observing candidate feature x_i as its regression target:

$$\min_{\phi} \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{X},Y)} \mathbb{E}_{\mathcal{S}} \mathbb{E}_i \left[\left(v_i(\mathbf{x}_{\mathcal{S}}; \phi) - \Delta(\mathbf{x}_{\mathcal{S}}, x_i, y) \right)^2 \right],$$

where $\Delta(\mathbf{x}_{\mathcal{S}}, x_i, y) = \ell \left(f(\mathbf{x}_{\mathcal{S}}; \theta), y \right) - \ell \left(f(\mathbf{x}_{\mathcal{S}\cup\{i\}}; \theta), y \right)$ denotes the decrease in prediction loss induced by acquiring feature X_i. This training scheme relies on the property that, when the predictor is Bayes-optimal and cross-entropy loss is used, the expected reduction in prediction loss provides an unbiased estimator of CMI. In particular,

$$\mathbb{E}_{(x_i,y)\sim p(X_i,Y \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})} \left[\Delta(\mathbf{x}_{\mathcal{S}}, x_i, y) \right] = I(Y; X_i \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}).$$

Therefore, assuming sufficient model capacity and successful optimization of both objectives, we obtain $f(\mathbf{x}_{\mathcal{S}}; \theta^*) = p(Y \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})$ and $v_i(\mathbf{x}_{\mathcal{S}}; \phi^*) = I(Y; X_i \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})$ for $i \in [d] \setminus \mathcal{S}$. Thus, DIME amortizes the estimation of candidate-wise CMI using prediction-loss reduction as the learning signal, without requiring a separate generative model for the unobserved features, and performs Greedy CMI acquisition based on the learned value estimates.

GDFS. GDFS consists of a selector network $\pi(\mathbf{x}_{\mathcal{S}}; \phi) \in \Delta^{d-1}$ for feature acquisition and a predictor network $f(\mathbf{x}_{\mathcal{S}}; \theta)$ for prediction. Given the currently observed feature set $\mathbf{x}_{\mathcal{S}}$, the selector outputs a probability distribution over candidate features and ideally recovers the Greedy CMI policy

$$i^*(\mathbf{x}_{\mathcal{S}}) = \arg\max_{i \in [d]\setminus\mathcal{S}} I(Y; X_i \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}}).$$

GDFS jointly trains the selector and predictor. For a randomly sampled observed feature subset $\mathcal{S}$, the model minimizes the prediction loss after acquiring a candidate feature sampled from the selector:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{X},Y)} \mathbb{E}_{\mathcal{S}} \left[\mathbb{E}_{i\sim\pi(\mathbf{x}_{\mathcal{S}};\phi)} \left[\ell \left(f(\mathbf{x}_{\mathcal{S}\cup\{i\}}; \theta), y \right) \right] \right].$$

Covert et al. (2023) show that, when an optimal solution exists at each acquisition step, the predictor at the optimum corresponds to the Bayes-optimal classifier, $f(\mathbf{x}_{\mathcal{S}}; \theta^*) = p(Y \mid \mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}})$, while the selector recovers the Greedy CMI policy $i^*(\mathbf{x}_{\mathcal{S}})$.

Accordingly, GDFS does not explicitly estimate the CMI of each candidate feature. Instead, it amortizes the acquisition policy itself by directly minimizing the prediction loss after feature acquisition, such that the resulting selector coincides with the Greedy CMI policy under the optimal setting.

Since the selector ultimately operates over a discrete feature space, direct backpropagation through the sampled feature index is not straightforward. Although a score-function estimator such as REINFORCE could be used, its high variance can make optimization difficult. GDFS therefore adopts a continuous relaxation based on the Concrete distribution, enabling end-to-end gradient-based training of the selector together with the predictor.

H CMI Estimation for Regression Baselines

ACFlow, EDDI, and VAE require explicit computation of CMI or its corresponding information gain when estimating acquisition scores. In classification, the target distribution is discrete, allowing the corresponding information-theoretic quantities to be estimated relatively straightforwardly. In regression, however, the target Y is continuous, and entropy computation generally requires integration over the target density, which is typically intractable. In this section, we describe the regression-specific acquisition score estimators used for ACFlow, EDDI, and VAE.

ACFlow. For regression, ACFlow models the continuous target jointly with the input features. Specifically, instead of using a class-conditioned distribution as in classification, we concatenate the target with the feature vector and train a joint flow over $\mathbf{w} = [\mathbf{x}, y]$.

This formulation allows the model to evaluate the conditional density of the target given an arbitrary subset of observed features. For a candidate feature X_i and a future feature subset $\phi \subseteq \mathcal{U} \setminus \{i\}$, the acquisition score is estimated using the following conditional log-density difference:

$$R_\gamma^{\mathrm{ACFlow}}(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi\sim p_\gamma}\mathbb{E}_{(\mathbf{x}_\phi, x_i, y)\sim p_\theta(\cdot|\mathbf{x}_\mathcal{O})}\left[\log p_\theta\left(Y \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi, x_i\right) - \log p_\theta\left(Y \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi\right)\right].$$

Here, $\mathbf{x}_\phi$, x_i, and y are obtained from the same joint sample generated by the flow conditioned on $\mathbf{x}_\mathcal{O}$. Using the same target sample in both terms provides a Monte Carlo estimate of the expected conditional log-density ratio.

If the learned distribution p_θ matches the true conditional distribution, the above score recovers

$$\mathbb{E}_{\phi\sim p_\gamma}\mathbb{E}_{\mathbf{x}_\phi\sim p(\mathbf{X}_\phi|\mathbf{x}_\mathcal{O})}\left[I\left(Y; X_i \mid \mathbf{X}_\mathcal{O} = \mathbf{x}_\mathcal{O}, \mathbf{X}_\phi = \mathbf{x}_\phi\right)\right],$$

which corresponds to the γ-CMI acquisition score. In practice, both the expectation over future subsets and the expectation over generated values are approximated using Monte Carlo sampling.

EDDI. Rather than directly computing the predictive entropy of Y, EDDI approximates the target-relevant information provided by a candidate feature X_i using the difference between latent posterior KL divergences.

First, the total information that observing X_i provides about the latent variable $\mathbf{Z}$ is computed as

$$\mathrm{KL}_1(i) = D_{\mathrm{KL}}\left(q_\varphi\left(\mathbf{Z} \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi, x_i\right) \| \, q_\varphi\left(\mathbf{Z} \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi\right)\right).$$

After additionally conditioning on the target value y, the residual latent information provided by X_i is computed as

$$\mathrm{KL}_2(i) = D_{\mathrm{KL}}\left(q_\varphi\left(\mathbf{Z} \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi, x_i, y\right) \| \, q_\varphi\left(\mathbf{Z} \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi, y\right)\right).$$

The regression acquisition score of EDDI is then defined as the difference between these two KL divergences:

$$R_\gamma^{\mathrm{EDDI}}(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi, \mathbf{x}_\phi, x_i, y}\left[\mathrm{KL}_1(i) - \mathrm{KL}_2(i)\right].$$

For regression, the target decoder used to generate the target value y required for computing $\mathrm{KL}_2(i)$ is modeled as a Gaussian:

$$p_\theta(Y \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \sigma_\theta^2(\mathbf{z})\right).$$

Since the approximate latent posteriors are also diagonal Gaussians, both KL terms can be computed in closed form. EDDI+ uses the same acquisition score as EDDI, and differs only in its training objective.

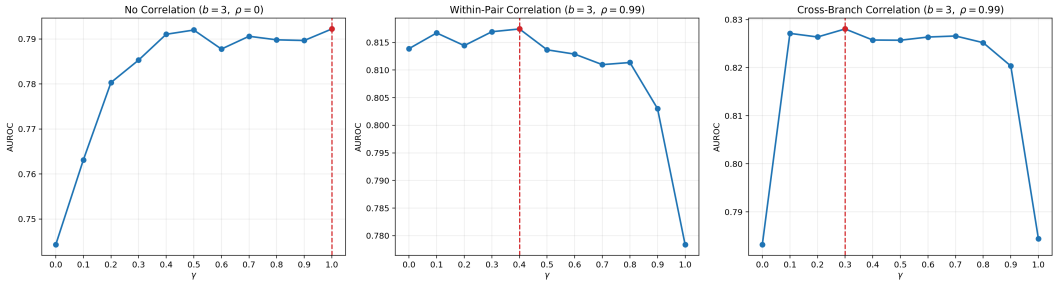


Figure I.1: Predictive performance across γ on Gaussian synthetic datasets. Each curve reports the AUROC averaged over the acquisition trajectory, from the first acquisition step to full acquisition. The red marker and dashed line indicate the best-performing γ in each setting.

VAE. For regression, the VAE predictor models the target distribution as a Gaussian:

$$p_\psi(Y \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi) = \mathcal{N}\left(\mu_\psi(\mathbf{x}_\mathcal{O}, \mathbf{x}_\phi), \sigma^2_\psi(\mathbf{x}_\mathcal{O}, \mathbf{x}_\phi)\right).$$

The differential entropy of a Gaussian distribution can be computed in closed form as

$$H_{\text{Gauss}}\left(Y \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi\right) = \frac{1}{2}\log\left(2\pi e\, \sigma^2_\psi(\mathbf{x}_\mathcal{O}, \mathbf{x}_\phi)\right).$$

The acquisition score of candidate feature X_i is defined as the expected reduction in predictive Gaussian entropy:

$$R^{\text{VAE}}_\gamma(i, \mathbf{x}_\mathcal{O}) = \mathbb{E}_{\phi\sim p_\gamma}\mathbb{E}_{(\mathbf{x}_\phi, x_i)\sim p_\theta(\cdot\mid\mathbf{x}_\mathcal{O})}\left[H_{\text{Gauss}}\left(Y \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi\right) - H_{\text{Gauss}}\left(Y \mid \mathbf{x}_\mathcal{O}, \mathbf{x}_\phi, x_i\right)\right].$$

In practice, each Gaussian entropy term is computed in closed form, while the expectations over future subsets and their corresponding generated values are approximated using Monte Carlo sampling.

I Predictive Performance Across γ

I.1 Performance Across γ on Gaussian Synthetic Dataset

Using the Gaussian synthetic setting and oracle models described in Section 5.1, we examine how the predictive performance of γ-CMI varies with γ. We consider the no-correlation and within-pair correlation configurations from that section, together with the cross-branch correlation setting described in Appendix D.2. Since these datasets are binary classification problems, we use AUROC as the evaluation metric and evaluate γ from 0 to 1 in increments of 0.1.

Figure I.1 summarizes the resulting performance curves across these three settings. When synergistic interactions dominate, larger values of γ can perform better because candidates are more likely to be evaluated under future contexts containing other features that contribute to the relevant synergistic interaction. In contrast, when redundancy dominates, smaller values of γ can perform better by limiting the influence of negative interaction contributions from redundant future contexts. Consistent with the preceding acquisition analyses, these results suggest that the best-performing γ within the evaluated sweep depends on the balance between synergy and redundancy in the underlying feature structure.

I.2 Performance Across γ on Real-World Datasets

In real-world datasets, feature synergy, redundancy, and model error may coexist in complex ways. As a result, the optimal value of γ can vary depending on the trade-off among these factors. Rather than exhibiting a monotonic relationship with γ, performance may improve and then degrade as γ varies between 0 and 1. This suggests that the effect of future-context conditioning is inherently dataset dependent.

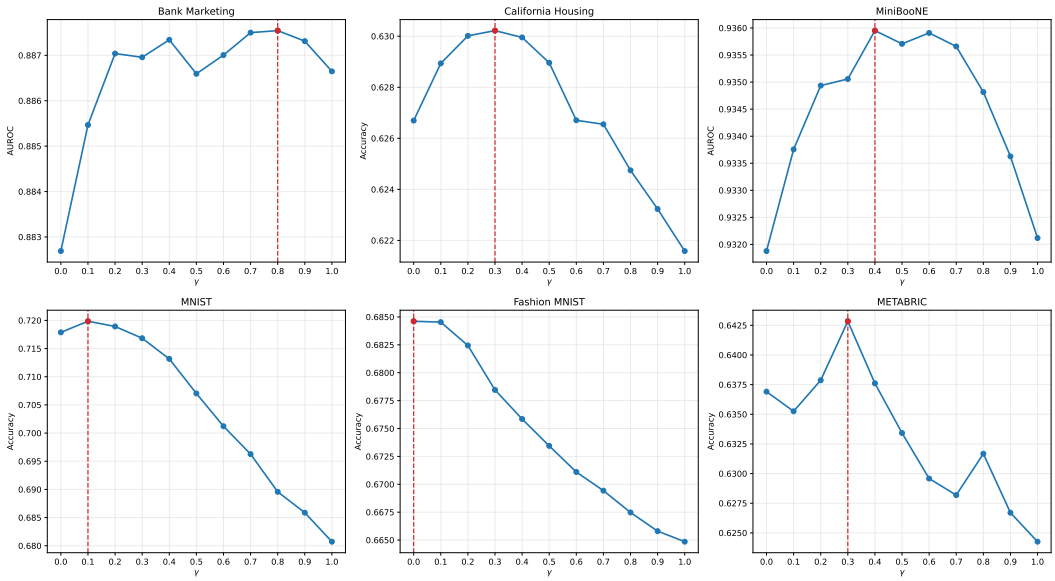


Figure I.2: Performance across different values of γ on VAE-based classification datasets. Each curve reports the AUROC averaged for binary and Accuracy averaged for multiclass tasks over the acquisition trajectory, from the first acquisition step to full acquisition. The red marker and dashed line indicate the best-performing γ in each setting.

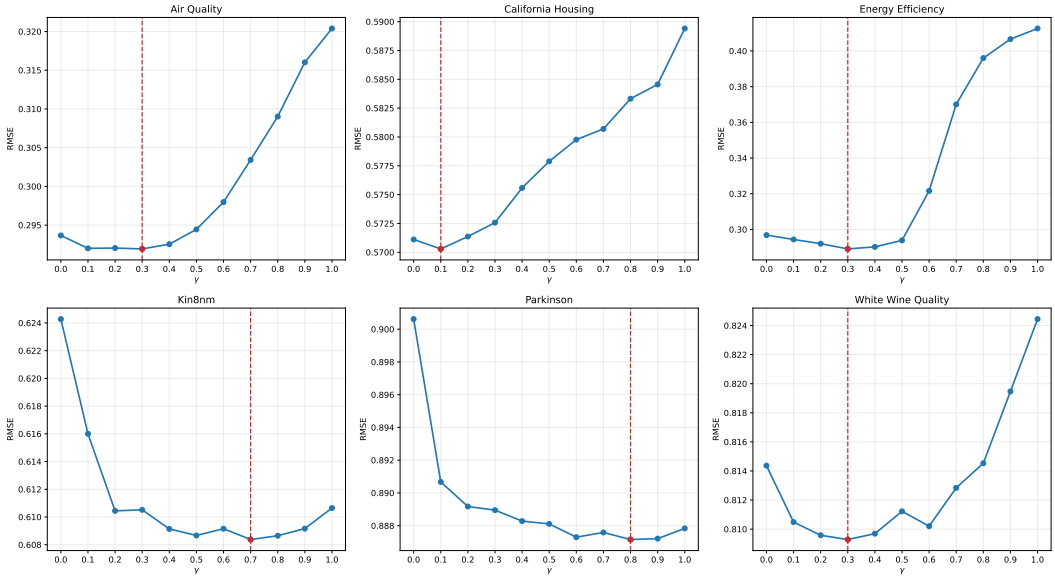


Figure I.3: Performance across different values of γ on VAE-based regression datasets. Each curve reports the RMSE averaged over the acquisition trajectory, from the first acquisition step to full acquisition. The red marker and dashed line indicate the best-performing γ in each setting.

As shown in Figures I.2 and I.3, the optimal value of γ varies across datasets. In general, the best performance is achieved at an intermediate value, $0 < \gamma < 1$, rather than at the extreme values $\gamma = 0$ or $\gamma = 1$. This empirical trend is consistent with our theoretical analysis, which characterizes the optimal γ as arising from the trade-off between feature interactions and model error.

Table J.1: VAE inference-time comparison over the complete test set and acquisition trajectory. Times are reported in seconds as mean ± standard deviation over five model repeats. Overhead is the ratio of the mean γ-CMI inference time to the mean Greedy CMI inference time.

Dataset	Best γ	Greedy CMI ($M = 1$)	γ-CMI ($M = 5$)	Overhead
Bank Marketing	1.0	9.3608 ± 1.8514	8.0274 ± 0.5273	$0.86\times$
California Housing	0.2	2.8660 ± 0.7244	2.3989 ± 0.1095	$0.84\times$
MiniBooNE	0.5	26.1004 ± 3.2050	26.8682 ± 1.3715	$1.03\times$
MNIST	0.1	21.0566 ± 0.1379	28.0682 ± 3.7941	$1.33\times$
Fashion-MNIST	0.0	20.8836 ± 0.3171	21.0273 ± 1.1525	$1.01\times$
METABRIC	0.5	0.9454 ± 0.0159	0.9447 ± 0.0189	$1.00\times$
Air Quality	0.1	1.8154 ± 0.2320	2.3959 ± 0.1340	$1.32\times$
California Housing (Reg.)	0.1	3.9067 ± 0.4454	3.9759 ± 0.1766	$1.02\times$
Energy Efficiency	0.4	0.7574 ± 0.0307	0.9236 ± 0.0684	$1.22\times$
Kin8nm	0.7	1.3147 ± 0.0170	1.3194 ± 0.0561	$1.00\times$
Parkinson	0.6	2.2042 ± 0.1631	3.0816 ± 0.1909	$1.40\times$
White Wine Quality	0.3	1.2888 ± 0.1271	1.2678 ± 0.0338	$0.98\times$
			Average overhead	$\mathbf{1.08}\times$

J Inference-Time Overhead

We measured the wall-clock inference time of VAE-based γ-CMI under the same experimental setting as in the main real-world evaluation, over the complete test set and the entire feature-acquisition trajectory. Greedy CMI uses no future context ($\Phi = \varnothing$) and therefore requires a single acquisition-score evaluation ($M = 1$), whereas γ-CMI evaluates five sampled future masks ($M = 5$). For γ-CMI, we use the dataset-specific value of γ selected in the main real-world evaluation. The five future-mask evaluations were stacked along the batch dimension and evaluated in parallel on a single NVIDIA RTX A6000 GPU. Model loading, metric computation, and result serialization were excluded from the timed region. Each reported value is the mean ± standard deviation over five independently trained model repeats. Overall, as shown in Table J.1, batched γ-CMI required approximately $1.08\times$ the inference time of Greedy CMI.