DISCOVERING FEATURE SUBSPACE THROUGH DATA LEARNABILITY: A KERNEL-LEARNING REALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern machine learning models can fit high-dimensional data despite irrelevant or redundant features. However, discovering a minimal feature subspace that preserves the underlying predictive structure remains challenging. A key difficulty is that feature utility is typically assessed through a particular model, and is therefore entangled with model-specific inductive biases. Here, we revisit the feature subspace discovery problem from a different perspective: the representation-induced learnability of the observed data. Our key idea is that an informative feature subspace should organize data into suitable structures, which benefit model performance across a range of complexities. To quantify this property, we construct kernel-based estimators and track empirical fitting performance as a function of model complexity, measured by the trace of the normalized Gram matrix. This defines a learning curve, and the Area Under the Learning Curve (AULC) provides a scalar measure of data learnability. AULC provides a principled criterion for comparing feature subspaces. Experiments across synthetic and real-world datasets demonstrate its effectiveness in feature subspace discovery and related scientific variable discovery tasks.

1 INTRODUCTION

Modern machine learning models can discover powerful hidden representations H that mediate the relationship between inputs X and outputs Y, typically by automatically filtering and transforming the original variables. In many scientific and engineering problems, however, the representation itself must remain explicit: the goal is to identify which observed variables organize the data into a form in which the input-output relationship is easier to learn. This gives rise to the problem of explicit feature subspace discovery, where one seeks a compact subset of candidate observables that preserves the informative structure relevant to a target quantity from samples $\mathscr{S}_n = \{(X_i, Y_i)\}_{i=1}^n$.

A standard approach is to evaluate a candidate feature subspace through the predictive performance of a learning model:

$$X_{\text{opt}} = \arg \min_{X^* \subset X} \min_{f \in \mathscr{F}} \left[\mathbb{E}_{(X,Y)\sim p}[\ell(Y, \hat{f}(X))] \right]. \tag{1}$$

Here, ℓ denotes a loss functional, for which we consider the mean squared error (MSE). Combined with subset-search strategies, this formulation underlies classical wrapper methods.

In finite-sample settings, however, the quality assigned to a feature subspace is inevitably entangled with the inductive bias of the model used for evaluation. Models with insufficient capacity may miss informative structures, whereas highly flexible models may compensate for poor feature representations through complex approximations. Predictive performance therefore reflects not only the quality of the selected features, but also the approximation ability of a particular model class.

This motivates a different perspective: rather than asking how well a specific model predicts from a feature subspace, we ask whether the subspace organizes the data into representations that are easy to learn. We refer to this property as *data learnability*. For example, a representation that exposes a linear input-output relationship should generally be easier to learn than one in which the same relationship is expressed through a more complicated geometry. A desirable feature subspace should therefore simplify the underlying data structure, rather than merely favor a particular predictor.

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.

To quantify this property, we develop a framework based on a Gaussian kernel that jointly tracks empirical fitting error and model complexity. Equivalently, the kernel bandwidth determines the neighborhood scale over which each sample is compared with surrounding observations, allowing us to examine how well the data preserve a coherent organization as this reference neighborhood is gradually expanded. Varying the effective complexity of the kernel estimator yields a learning curve that characterizes how rapidly fitting error decreases as model capacity increases. Its Area Under the Learning Curve (AULC) summarizes this trade-off: a smaller AULC indicates that lower fitting error can be achieved at comparable complexity, corresponding to greater data learnability. We use AULC as a criterion for comparing feature subspaces without selecting a single model complexity.

Combined with backward elimination, AULC provides an effective criterion for feature subspace discovery. Across controlled benchmarks and scientific problems, it captures nonlinear and higher-order dependencies while favoring compact feature sets, with particularly strong performance in systems without simple governing equations. Our main contributions are:

- We introduce *data learnability* as an alternative principle for feature subspace discovery, focusing on the structure induced by the selected feature subspace rather than the performance of a specific predictor.
- We develop AULC, a Gaussian-kernel realization that compares feature subspaces through data variation at matched empirical model complexity, with theoretical analysis under idealized asymptotic conditions.
- We combine AULC with backward elimination and demonstrate its effectiveness across symbolic, piecewise, latent-structure, and scientific discovery problems.

2 Related Work

While statistical dependence Spirtes et al. (2000) and information-theoretic measures Kraskov et al. (2004) provide general tools for quantifying relationships in data, identifying an effective feature subspace requires not only informativeness but also parsimony, so that redundant observables can be excluded Tibshirani (1996). Many feature-selection methods, including HSIC-based approaches Gretton et al. (2005) and mRMR Peng et al. (2005), typically provide feature rankings rather than directly determining the final feature subset. These approaches therefore often require the desired number of retained features to be specified in advance, or rely on additional regularization or model-selection criteria to determine the subset size, as in HSIC Lasso Yamada et al. (2014). Moreover, criteria restricted to linear associations can be insufficient for real-world systems exhibiting nonlinear and higher-order dependencies Casadiego et al. (2017); Bianconi (2021); Li et al. (2024).

Modern feature subspace selection therefore often relies on predictive models to capture complex structure. Advances in neural networks, symbolic regression, and related approaches have greatly improved the ability to model nonlinear relationships Rumelhart et al. (1986); Hornik et al. (1989); Udrescu & Tegmark (2020); Gao & Yan (2022). Predictive performance is consequently widely used both to learn representations implicitly Vaswani et al. (2017); Lemhadri et al. (2021) and to evaluate explicit feature subspaces Kohavi & John (1997); Guyon et al. (2002); Li et al. (2024). From this perspective, the quality of a feature subspace is naturally connected to the learnability of the input-output relationship it induces Bengio et al. (2013); Vapnik (2013); Champion et al. (2019).

Predictive performance, however, remains only an indirect proxy for learnability and representation quality Plachouras et al. (2025). A key confounding factor is model complexity: sufficiently flexible models may compensate for poor representations through increasingly complex approximations Zhang et al. (2021), particularly in data-scarce regimes where the underlying structure is underspecified D'Amour et al. (2022). Although model complexity is routinely controlled during training, less attention has been paid to the complexity actually realized by the fitted model, and to using it directly for feature subspace discovery. This distinction is important because nominal model complexity can differ substantially from empirical complexity, which is shaped by parameterization, optimization, and training dynamics Domingos (2012); Kingma & Ba (2017); Belkin et al. (2019).

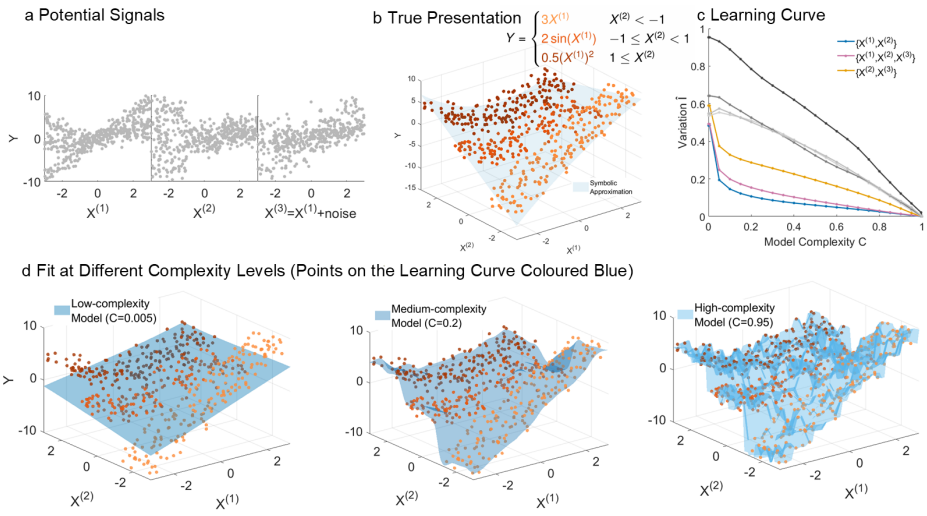

Figure 1: The AULC framework for feature subspace discovery.

3 AULC: A KERNEL-LEARNING FRAMEWORK

We construct AULC using a Gaussian-kernel estimator, where the kernel bandwidth controls empirical model complexity and the resulting fitting variation quantifies how easily the data can be described at the given complexity. For samples $\mathscr{S}_n = \{X_i, Y_i\}_{i=1}^n$, we define

$$K^X(X_i, X_j; \Sigma, h) = \frac{1}{(2\pi h)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{(X_i - X_j)^T \Sigma^{-1}(X_i - X_j)}{2h}\right\}, \tag{2}$$

where d is the dimension of the corresponding space, and h and Σ control the decay rate and shape of the kernel. Throughout the paper, superscripts indicate the associated space and subscripts denote sample indices. The corresponding density estimator is therefore

$$\hat{p}(x) = \frac{1}{n}\sum_i K^X(x, X_i; \Sigma, h) := \frac{1}{n}\sum_i K^X_{\cdot,i}. \tag{3}$$

3.1 KERNEL ESTIMATOR AND DATA VARIATION

Using a joint kernel K^{XY}, the conditional expectation is estimated as

$$\hat{f}(x) = \int y \frac{\hat{p}(x,y)}{\int \hat{p}(x,y)dy} dy = \frac{\sum_i \int y K^{XY}_{\cdot,i} dy}{\sum_i \int K^{XY} \cdot, i dy} = \frac{\sum_i \mu \cdot, i K^X_{\cdot,i}}{\sum_i K^X_{\cdot,i}} := \sum_i \mu_{\cdot,i} W^X_{\cdot,i}, \tag{4}$$

where $\mu^X_{\cdot,i} = Y_i + \Sigma_{XY}(x - X_i)$ and $W^X_{\cdot,i} = K^X_{\cdot,i} / \sum_i K^X_{\cdot,i}$. We define the joint covariance as

$$\Sigma = \begin{pmatrix} \Sigma_Y & \Sigma_{XY} \\ \Sigma_{XY}^T & \Sigma_X = I \end{pmatrix}, \tag{5}$$

where $\Sigma_X = I$ provides a compromise between flexibility and computational efficiency (see Appendix). Since Σ_Y does not enter Eq.4, positive definiteness can be ensured through an appropriate choice of Σ_Y without affecting the estimator.

We define the local variation at x as

$$\hat{R}^X(x; \mathscr{S}, h) = \sum_j W^X_{\cdot,j}\left[\mu^X_{\cdot,j} - \sum_i W^X_{\cdot,i}\mu^X_{\cdot,i}\right]^2 = \frac{\sum_{i,j}(\mu^X_{\cdot,i} - \mu^X_{\cdot,j})^2 W^X_{\cdot,i} W^X_{\cdot,j}}{2}, \tag{6}$$

where $T_i - T_j := \mu^X_{\cdot,i} - \mu^X_{\cdot,j} = (Y_i - Y_j) + \Sigma_{XY}(X_j - X_i)$ is independent of the query point x. The overall data variation is then estimated by

$$\hat{I}^X(\mathscr{S}; \Sigma_{XY}, h) = \frac{1}{n}\sum_{\mathscr{S}} \hat{R}^X(X_i) = \frac{1}{n}\sum_{k=1}^n \frac{\sum_{i,j}(T_i - T_j)^2 W^X_{ki} W^X_{kj}}{2}, \tag{7}$$

with Σ_{XY} optimized to minimize $\hat{I}$ at the given C. By default, both X and Y are z-standardized.

3.2 Model Complexity and AULC

We characterize empirical model complexity through the normalized Gram matrix $W^X = [W^X_{ij}]$:

$$C(\mathscr{S}; h, \Sigma = I) = \frac{\mathrm{tr}(W^X)}{n}. \tag{8}$$

Small h concentrates W^X_{ij} around $j = i$, allowing more localized and complex functions, whereas large h produces smoother estimators (see Eq.4). C decreases monotonically with h.

Varying h generates a family of variation-complexity pairs $(C, \hat{I})$, defining the *learning curve*. It describes how much model complexity is required to reduce the unexplained variation in a given representation. We summarize this trade-off through the area under the learning curve (AULC):

$$\mathrm{AULC}(\mathscr{S}; X, Y) = \int_0^1 \hat{I}(C) dC. \tag{9}$$

This is numerically achieved through the trapezoidal rule. A smaller AULC indicates that low variation can be achieved at lower complexity, and hence greater data learnability.

Geometrically, varying h changes the spatial scale over which samples interact. AULC therefore aggregates predictive conflict across multiple scales, motivating the feature subspace objective

$$X_{opt} = \arg \min_{X^* \subset X} \mathrm{AULC}(\mathscr{S}; X, Y). \tag{10}$$

As $C \to 1/n$, the weights become uniform and the estimator approaches a linear model (Eq.4). For consistent integration, we extend the domain to $C \in [0, 1]$ by setting $\hat{I}(0) = \hat{I}(C_{\min})$. The opposite limit, $C = 1$, corresponds to each sample referring only to itself, giving $\hat{I}(1) = 0$ for finite samples. This modification usually does not dramatically change the values of AULC.

3.3 Theoretical Properties

The following properties clarify the population-level mechanism underlying AULC subspace preference in an idealized asymptotic and noiseless setting:

(Lemma. 1) $\hat{I} \geq 0$, C decreases monotonically with h, and $\mathrm{AULC} \geq 0$. For fixed $C \in (0, 1)$, the required bandwidth satisfies $h = \Theta(n^{-2/d^X})$.

(Proposition. 2) Let $Y = m(X^*)$, where m is continuous almost everywhere and X^* is the minimal informative feature subspace. For every fixed non-degenerate complexity $C \in (0, 1)$, the variation associated with X^* converges to zero, whereas underselection leaves a non-vanishing variation floor.

(Proposition. 3) Under the stated regularity and redundancy assumptions, overselection yields asymptotically larger variation than X^* at matched complexity.

(Corollary. 1) These preferences extend to the discretized AULC evaluated on any fixed finite complexity grid.

3.4 Algorithms and Numerical Realization

The algorithms for AULC and feature subspace discovery are detailed in the Appendix.

4 Experiments

4.1 Experimental Settings

Most existing studies on feature subspace selection evaluate methods through downstream predictive performance. Since our goal is to explicitly recover the informative feature subspace, we instead evaluate feature-level recovery against the ground truth. We use the following metrics:

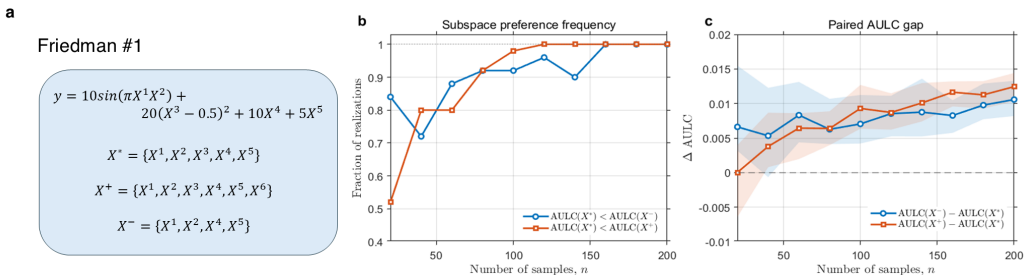

Figure 2: The AULC performance on the Friedman #1 benchmark, with 50 realizations. The shaded areas indicate the standard deviation. The discovery experiments are reported in the Appendix.

- EMR (Exact Match Rate): The proportion of experiments in which the selected feature subspace exactly matches the ground-truth subspace.
- PPV (Average Positive Predictive Value): The average proportion of selected features that are truly informative.
- Recall (Average Recall): The average proportion of ground-truth informative features that are successfully recovered.

For the baseline methods, we focus on approaches that do not require the target number of features in advance. Many feature selection methods only provide rankings or importance scores. An additional criterion, often based on predictive performance, is then needed to determine the final subspace size. We therefore consider the following baselines. The implementation is detailed in the Appendix.

- SR (Symbolic Regression): Informative features are identified from the variables retained in the discovered symbolic expression. This provides a natural baseline when the system admits a compact symbolic representation.
- RF-RFECV (Random Forest with Recursive Feature Elimination and Cross-Validation): A random forest is used to rank and recursively eliminate features, while cross-validation determines the final subset. We use 500 trees.
- ANN-BECV (Artificial Neural Network with Backward Elimination and Cross-Validation): A three-layer neural network is combined with backward elimination, and cross-validation determines the final subset.
- KRR-BECV (Kernel Ridge Regression with Backward Elimination and Cross-Validation): The same backward-elimination procedure is applied using kernel ridge regression as the predictive model.
- LassoNet: LassoNet performs nonlinear feature selection through hierarchical sparsity constraints, with cross-validation selecting the regularization level and feature subspace.

4.2 Empirical Validation of AULC Subspace Preference

We use the Friedman #1 benchmark to empirically examine the subspace preference induced by AULC (Fig. 2). The ground-truth subspace X^* attains a smaller AULC than both the under-selected subspace X^- and the over-selected subspace X^+. As the sample size n increases, these preferences become increasingly consistent across repeated realizations, in agreement with the asymptotic behavior established in Propositions 2 and 3. With sufficient samples, AULC consistently prefers X^* to both alternatives. All repeated realizations exhibit the correct preference once $n \geq 150$.

4.3 Recovering Structured Scientific Relations

Systems governed by symbolic equations provide an intuitive setting in which feature subspace discovery is of direct scientific interest. Many existing studies assume the existence of such governing relations and aim to recover them from observed data. To evaluate AULC-BE in this setting, we

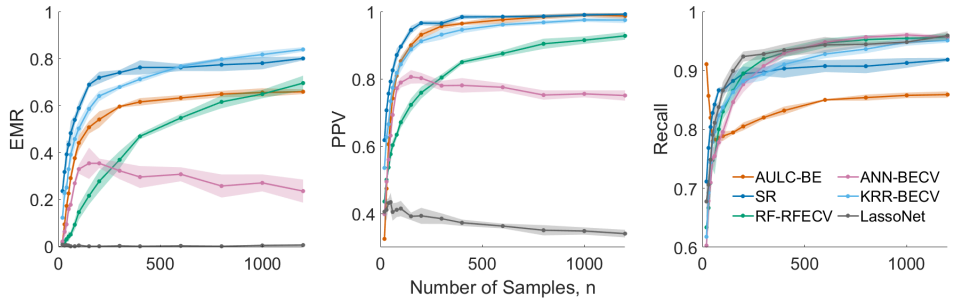

Figure 3: Results on Feynman datasets, calculated on 120 equations and 5 independent realizations.

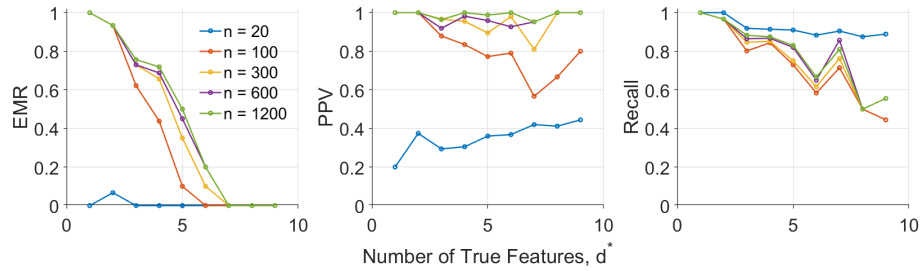

Figure 4: Feature recovery performance versus problem dimensions with the AULC-BE method.

apply it to the Feynman benchmark, which contains 120 physics equations. The data generation process is detailed in the Appendix.

The results are presented in Fig. 3. AULC-BE does not outperform the strongest structure-matched baselines, SR and KRR-BECV, on this benchmark, but exhibits a distinctive recovery pattern: as more data become available, false positives are strongly suppressed, while exact recovery becomes increasingly limited by the dimensionality of the ground-truth feature set.

A notable property of AULC-BE is its conservative selection behavior with sufficient data (Fig. 4). With very limited data, AULC-BE tends to over-select, as indicated by relatively high Recall but low PPV. As the sample size increases, PPV approaches 1, while exact recovery remains strongly dependent on the number of ground-truth variables. This suggests that AULC-BE may under-select when the available data are insufficient. Since AULC aggregates variation across multiple complexity levels, informative features need to provide sufficiently consistent improvements across the learning curve to be retained.

Overall, the results reveal a clear trade-off between sample size and dimension: recovering higher-dimensional informative subspaces roughly requires more data, while AULC-BE increasingly favors compact subsets composed almost exclusively of truly informative variables. This behavior is consistent with the dimensional dependence of kernel locality, for which maintaining a fixed complexity level requires increasingly dense sampling as the informative dimension grows.

4.4 Systems without Uniform Symbolic Forms

Beyond uniform symbolic forms, many systems are governed by piecewise or logical equations, meaning that the governing equations may not be continuous while the system may still possess an effective feature subspace. Here, we consider two tasks: a constructed piecewise equation and a system arising from a scientific discovery problem.

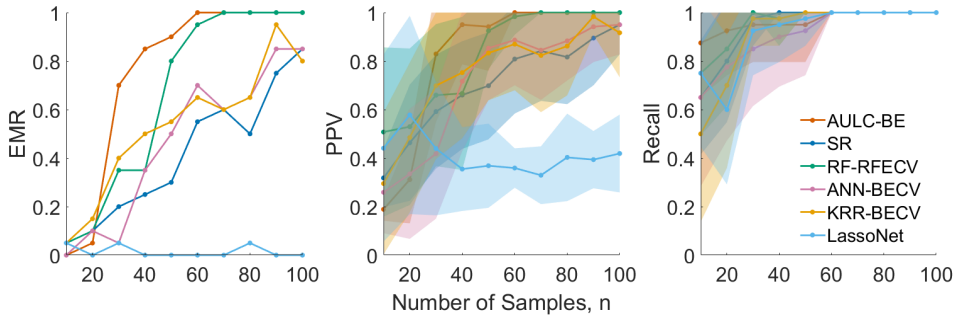

Figure 5: Results on the discontinuous system defined by Eq.11, with 20 realizations.

4.4.1 A PIECEWISE EQUATION

To evaluate AULC-BE beyond systems governed by a single global functional form, we construct a piecewise system in which only the first two features are informative:

$$y = \begin{cases} 3X^{(1)} & X^{(2)} < -1 \\ 3\sin(X^{(1)}) & -1 \le X^{(2)} < 1 \\ 0.5(X^{(1)})^2 & 1 \le X^{(2)} \end{cases} \tag{11}$$

We generate each dataset by independently sampling ten input variables from a uniform distribution over $[-3, 3]$ and computing y according to Eq.11. Thus, $X^{(1)}$ determines the response within each regime, while $X^{(2)}$ acts as a switching variable that determines which functional relation applies; the remaining eight variables are irrelevant.

The results are shown in Fig. 5. Compared with the Feynman systems, this benchmark cannot be represented compactly by a single smooth symbolic expression over the full input space. Consequently, SR tends to approximate the piecewise relation using a global expression rather than recovering its regime structure. For example, one discovered approximation is $Y = X^{(1)}\left[2.2918 + 0.1143(X^{(1)} - 6.1051)(X^{(2)} + 1.4265)\right]$. By contrast, methods that learn directly from the observed data remain applicable without requiring a single global symbolic form. RF-RFECV performs particularly well, which is consistent with the ability of decision trees to represent axis-aligned regime changes such as those induced by $X^{(2)}$.

AULC-BE performs well in this setting and achieves exact recovery with relatively few samples. This result is particularly relevant to the proposed learnability perspective: removing $X^{(2)}$ mixes samples governed by three distinct local relations, thereby increasing the variation that must be explained within the same representation.

4.4.2 IDENTIFYING CONTROL CONDITIONS FOR SIS DYNAMICS

Many dynamical systems are governed by threshold-like conditions that determine transitions between qualitatively different regimes. We examine this setting using the Susceptible–Infectious–Susceptible (SIS) process on networks.

Individuals are represented as nodes in a network G and transition between susceptible and infectious states according to the transmission rate β, recovery rate μ, and the states of neighboring nodes. For networks with relatively homogeneous degree distributions, such as Erdős–Rényi (ER) networks, epidemic persistence can be approximately characterized by a condition involving the mean degree, $f(d_{\text{mean}}, \mu, \beta) > 0$. For networks with heterogeneous degree distributions, a condition involving the largest eigenvalue of the adjacency matrix, $f(\Lambda_{\max}, \mu, \beta) > 0$, provides a more appropriate approximation to the epidemic threshold Wang et al. (2003).

Rather than deriving these conditions analytically, we ask whether the relevant observables can be recovered directly from data. The data-generation procedure is therefore important, since the representation that can be discovered is determined by the information contained in the observations. Details are provided in the Appendix.

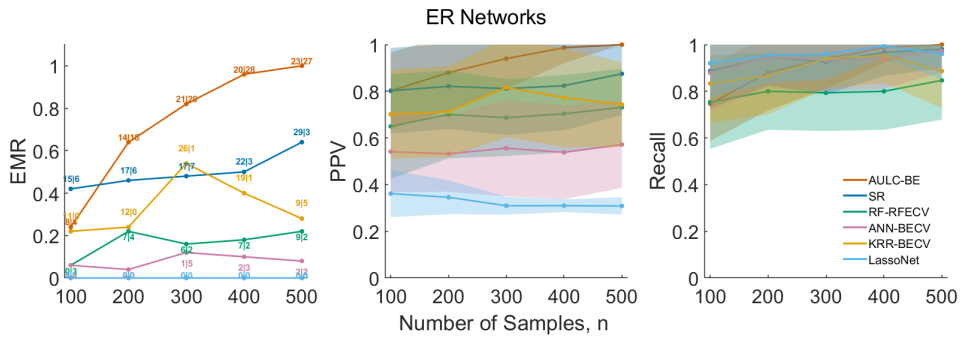

Figure 6: Identifying the controlling condition for SIS processes on ER networks, with 50 realizations. The numbers in the left panel indicate the numbers of $X^1|X^2$ in all realizations.

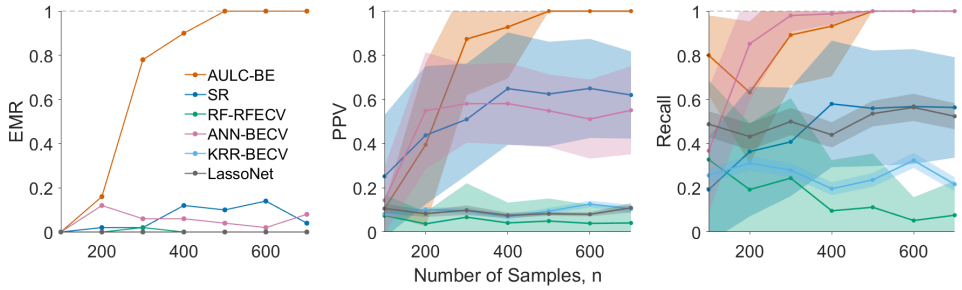

Figure 7: Results on the Madelon-like system, with 50 realizations.

Fig. 6 shows the results for ER networks. AULC-BE increasingly recovers the feature subspace corresponding to the expected control equations as the number of samples grows, reaching near-perfect PPV and Recall and exact recovery in all realizations at the largest sample sizes. The substantially lower EMR of most baselines indicates that identifying the minimal set of observables governing the transition is more demanding than retaining variables that are merely predictive.

More generally, the framework can recover different representations for distinct network ensembles when the underlying transition condition changes (Fig.13 in the Appendix). This illustrates an important property of data-driven scientific discovery: the representation identified by the method is determined jointly by the system dynamics and by the information made available through the data.

4.5 Madelon-like Systems

The Madelon benchmark was originally introduced for feature-selection evaluation Guyon et al. (2004). We construct a Madelon-like benchmark following the procedure detailed in the Appendix, with results shown in Fig. 7. Unlike the Feynman systems, the informative representation does not admit a simple explicit mapping to the output and must be recovered from nonlinear latent structure corrupted by noise and irrelevant variables. AULC-BE naturally extends to such classification tasks if we encode class labels as $Y \in \{0, 1\}$ and apply Eq.7 directly.

AULC-BE is the only evaluated method that consistently approaches exact support recovery as the sample size increases, with EMR, PPV, and Recall all converging to one. In contrast, ANN-BECV attains high Recall but substantially lower PPV, indicating persistent over-selection, while the remaining baselines fail to recover the ground-truth feature set reliably. These results highlight the advantage of evaluating representation learnability rather than predictive relevance alone.

4.6 Pattern Recognition in Medical Datasets

We further evaluate AULC on two real-world datasets. For the Breast Cancer Wisconsin Diagnostic (WDBC) dataset Street et al. (1993), AULC identifies a compact subspace of 6 features from 30

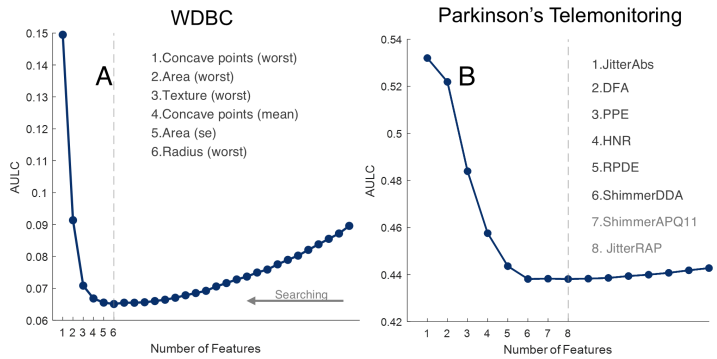

Figure 8: Experimental results on (A) WDBC and (B) Parkinson's Telemonitoring Datasets.

candidates (Fig. 8A). The largest reductions in AULC are associated with worst concave points, worst area, and worst texture, consistent with the original study, which emphasized the diagnostic value of "worst" measurements.

For the Parkinson's telemonitoring dataset, where dysphonia measures are used to predict the total UPDRS score, AULC selects 8 of 16 candidate features (Fig. 8B), with the first 6 providing most of the improvement. Four of these overlap with the parsimonious model reported by Tsanas et al. (2010). The remaining differences mainly involve closely related voice measures, suggesting that the selected subspaces capture similar underlying information.

5 Discussion

We show that focusing on the structure induced by observed data, rather than the predictive performance of a specific model, provides a conservative criterion for feature subspace discovery when no simple governing equation is available. This perspective, which we refer to as *data learnability*, offers an alternative formulation of feature discovery when the goal is to explicitly identify informative observables rather than construct an accurate predictor. While controlling model complexity is standard in statistical learning, AULC uses the empirically realized complexity as a common axis for comparing feature subspaces. Rather than penalizing complexity through a specified regularizer, it evaluates how rapidly the unexplained variation decreases as the realized complexity increases.

Like other data-driven approaches, AULC can only recover structure supported by the available samples and is therefore inherently sample-dependent. If the sampling process is unbalanced or under-represents important regimes, the discovered feature subspace may fail to reflect the more general system relation. Similarly, when different dynamical regimes are mixed within the same dataset, the method may preferentially identify features that dominate the aggregated observations rather than those specific to rare regimes. This dependence on the observed data is an important consideration when interpreting any representation obtained from the framework.

AULC lies between prediction-oriented modeling and causal discovery. It identifies relations supported by observational data without claiming to recover the underlying causal mechanisms. This makes it particularly useful as an exploratory tool when causal structure is difficult to establish or when the available observations are insufficient for causal identification.

A further limitation is computational cost, which increases with both the number of candidate observables and the number of samples. More efficient search and approximation strategies may improve scalability, although potentially at the cost of reduced sensitivity to subtle structure. While the present study focuses on continuous observables, the same principle can be extended to categorical inputs by treating categories as data partitions and evaluating whether they improve learnability.

AI use statement

Generative AI tools were used to assist with manuscript editing and language refinement, and to provide feedback on the presentation of the methodology, theoretical arguments, and experimental

design. AI tools were also used to assist with code editing and debugging. All theoretical claims, experimental procedures, implementations, and reported results were independently checked and validated by the authors. The authors take full responsibility for the content of the paper.

Reproducibility Statement

We provide the theoretical analysis, algorithmic details, numerical procedures, baseline implementations, and data-generation protocols in the Appendix. In particular, Appendix A contains the theoretical analysis, Appendix B details the computation of AULC and the backward-elimination procedure, Appendix D describes the baseline implementations and hyperparameter settings, and Appendix E specifies the data-generation procedures used in the experiments. Code and scripts for reproducing the experiments will be released upon publication.

References

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Ginestra Bianconi. *Higher-order networks*. Cambridge University Press, 2021.

Jose Casadiego, Mor Nitzan, Sarah Hallerberg, and Marc Timme. Model-free inference of direct network interactions from nonlinear collective dynamics. *Nature Communications*, 8(1):2192, 2017.

Kathleen Champion, Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116 (45):22445–22451, 2019.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

Oscar Fajardo-Fontiveros, Ignasi Reichardt, Harry R De Los Ríos, Jordi Duch, Marta Sales-Pardo, and Roger Guimerà. Fundamental limits to learning closed-form mathematical models from data. *Nature Communications*, 14(1):1043, 2023.

Ting-Ting Gao and Gang Yan. Autonomous inference of complex network dynamics from incomplete and noisy data. *Nature Computational Science*, 2(3):160–168, 2022.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(70):2075–2129, 2005. URL `http://jmlr.org/papers/v6/gretton05a.html`.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, 2002.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. *Advances in Neural Information Processing Systems*, 17, 2004.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial Intelligence*, 97 (1-2):273–324, 1997.

Alexander Kraskov, Harald Stogbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021.

Xin Li, Qunxi Zhu, Chengli Zhao, Xiaojun Duan, Bolin Zhao, Xue Zhang, Huanfei Ma, Jie Sun, and Wei Lin. Higher-order granger reservoir computing: simultaneously achieving scalable complex structures inference and accurate dynamics prediction. *Nature Communications*, 15(1):2506, 2024.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

Christos Plachouras, Julien Guinot, György Fazekas, Elio Quinton, Emmanouil Benetos, and Johan Pauwels. Position paper: Towards a unified representation evaluation framework beyond downstream tasks. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2025. doi: 10.1109/IJCNN64981.2025.11229333.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical Image Processing and Biomedical Visualization*, volume 1905, pp. 861–870. SPIE, 1993.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

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. doi: 10.1109/TBME.2009.2036000.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pp. 25–34. IEEE, 2003.

Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric P Xing, and Masashi Sugiyama. High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation*, 26(1):185–207, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A THEORETICAL ANALYSIS FOR AULC

Proving that AULC prefers the ideal feature subspace is difficult, especially in real-world scenarios where noise obscures the informative signals Fajardo-Fontiveros et al. (2023). We provide here a discussion of the properties of AULC under relatively restricted conditions.

We analyze the behavior of the proposed variation criterion at a fixed model complexity. Throughout this section, let

$$\mathscr{S}_n = \{(X_i^*, Y_i)\}_{i=1}^n, \quad d^* = \dim(X^*)$$

be an i.i.d. sample. For a candidate feature subspace $X \in \mathbb{R}^{d^x}$, define the normalized Gaussian-kernel weights:

$$W_{ki}^X(h) = \frac{\exp\{-||X_k - X_i||^2/(2h)\}}{\sum_{j=1}^n \exp\{-||X_k - X_j||^2/(2h)\}}. \tag{12}$$

The empirical model complexity associated with X and bandwidth h is:

$$C_n^X(h) = \frac{1}{n}\sum_{j=1}^n W_{kk}^X(h). \tag{13}$$

For a prescribed complexity level $c \in (0,1)$, let h_n^X denote a bandwidth satisfying

$$C_n^X(h_n^X) = c. \tag{14}$$

We exclude two degenerate endpoints $c = 0$ and $c = 1$ from the theoretical analysis.

For a linear correction $\Sigma \in \mathbb{R}^{1\times d^X}$, define

$$\hat{I}_n^X = \frac{1}{2n}\sum_{i,j,k=1} W_{ki}^X(h) W_{kj}^X(h) \left[(Y_i - \Sigma X_i) - (Y_j - \Sigma X_j)\right]^2. \tag{15}$$

Since Σ is learned by minimizing $\hat{I}_n^X$, the corresponding population-profiled criterion at complexity c is

$$\mathcal{I}_n^X(c) := \inf_{\Sigma} \mathbb{E}\left[\hat{I}_n^X(h_n^X(c), \Sigma)\right] \tag{16}$$

The following analysis characterizes the behavior of $\mathcal{I}_n^X(c)$ when informative variables are either removed or augmented by redundant variables.

A.1 VARIATION DECOMPOSITION

For a representation X, define:

$$m^X(x) := \mathbb{E}\left[Y|X = x\right], \quad v^X(x) := \mathrm{Var}(Y|X = x). \tag{17}$$

Also define, for an arbitrary function r,

$$\Omega_{r,n}^X(h) = \frac{1}{2n}\sum_{i,j,k} W_{ki}^X W_{kj}^X \left[r(X_i) - r(X_j)\right]^2. \tag{18}$$

Proposition 1. *(Variation Decomposition). When observed data $\mathscr{S}$ are organized by the feature subspace X,*

$$\mathbb{E}\left[\hat{I}_n^X(h, \Sigma) \mid X_{1:n}\right] = \Omega_{m^X - \Sigma X, n}^X(h) + \frac{1}{n}\sum_{i,k=1}^n W_{ki}^X(h)\left[1 - W_{ki}^X(h)\right] v^X(X_i). \tag{19}$$

Proof. For the regression model,

$$Y_i = m^X(X_i) + \xi_i, \quad \mathbb{E}\left[\xi_i \mid X_i\right] = 0, \quad \mathrm{Var}\left(\xi_i|X_i\right) = v^X(X_i)$$

Let:

$$r_\Sigma(x) = m^X(x) - \Sigma x.$$

Then:

$$(Y_i - \Sigma X_i) - (Y_j - \Sigma X_j) = r_\Sigma(X_i) - r_\Sigma(X_j) + (\xi_i - \xi_j).$$

The cross term vanishes when expanding the square and taking expectation conditional on $X_{1:n}$. Moreover, for $\forall i \neq j$,

$$\mathbb{E}\left[(\xi_i - \xi_j)^2 \mid X_{1:n}\right] = v^X(X_i) + v^X(X_j).$$

Therefore,

$$\mathbb{E}\left[\hat{I}_n^X(h, \Sigma) \mid X_{1:n}\right] = \Omega^X_{r_\Sigma, n}(h) + \frac{1}{2n} \sum_{k, i \neq j} W^X_{ki} W^X_{kj} \left[v^X(X_i) + v^X(X_j)\right].$$

With $\sum_j W^X_{kj} = 1$, the second term becomes

$$\frac{1}{n} \sum_{k,i} W^X_{ki}(1 - W^X_{ki}) v^X(X_i),$$

which finishes the proof. □

This decomposition separates the criterion into two conceptually different components. The first term measures local variation of the conditional mean that remains after the optimal linear correction, whereas the second term measures the unresolved information under the selected representation, potentially due to noise and/or ignorance of informative features.

A.2 Equal-complexity bandwidth scaling

The core conception underlying the AULC measure is that the comparison between representations organized by different feature subspaces must be performed at equal complexity rather than equal bandwidth. The following result characterizes the corresponding bandwidth scaling.

Assumption 1. *(Regular design) For every feature subspace X considered below, X has dimension d^X and admits a density p^X on a compact regular support $\mathcal{X}$, with*

$$0 < p_{min} \leq P^X(x) \leq p_{max} < \infty$$

on the interior of its support.

Lemma 1. *(Complexity-bandwidth scaling). Fix a non-generate complexity level $c \in (0, 1)$. Under Assumption. 1,*

$$h^X_n(c) = \Theta(n^{-2/d^X}). \tag{20}$$

Equivalently,

$$h^X_n(c) n^{2/d^X} = \Theta(1). \tag{21}$$

Proof. Since the Gaussian kernel is normalized row-wise,

$$W^X_{kk}(h) = \frac{1}{1 + \sum_{j \neq k} \exp\{-||X_k - X_j||^2/(2h)\}}.$$

Therefore, W^X_{kk}, and consequently C, monotonically decrease with h. Conditional on $X_k = x$,

$$\mathbb{E}\left[\sum_{j \neq k} e^{-||x - X_j||^2/(2h)} \mid X_k = x\right] = (n-1) \int e^{-||x-u||^2/(2h)} p_x(u) du$$

With $u = x + \sqrt{h}v$,

$$\int e^{-||z-u||^2/(2h)} p_X(u) du = h^{d^X/2} \int e^{-||v||^2/2} P_X(x + \sqrt{h}v) dv$$

The final integral is bounded above and below by positive constants under Assumption. 1. Hence the effective off-diagonal kernel mass is of order

$$n h^{d^X/2}.$$

Maintaining W^X_{kk} and consequently $C^X_n(h)$, at a fixed non-degenerate value therefore requires

$$n h^{d^X/2} = \Theta(1).$$

□

An immediate consequence is that, if $d_2 > d_1$,

$$\frac{h_{2,n(c)}}{h_{1,n(c)}} = \Theta(n^{2/d_1 - 2/d_2}) \xrightarrow{n\to\infty} \infty \tag{22}$$

Thus, at the same model complexity and the same data, a higher-dimensional feature subspace requires a substantially wider bandwidth h.

A.3 Local kernel regularity

To compare the signal variation across representations organized by different feature subspaces, we require that the local sample geometry is non-degenerate in every direction.

For each query point X_k, define the kernel-local pair covariance

$$Q_k^X(h) = \frac{1}{2}\sum_{i,j} W_{ki}^X(h) W_{kj}^X(h)(X_i - X_j)(X_i - X_j)^T. \tag{23}$$

Assumption 2. *(Kernel-local non-degeneracy). For every fixed c in a compact subset of* $(0, 1)$*, there exist constants*

$$0 < a_X(c) \le b_X(c) < \infty$$

such that, for $h = h_n^X(c)$*,*

$$a_X(c) hI \preceq \mathbb{E}\left[Q_k^X(h) \mid X_k\right] \preceq b_X(c) hI \tag{24}$$

as $n \to \infty$*, uniformly over* X_k *away from a vanishing boundary region.*

Assumption. 2 excludes locally degenerate designs in which all neighboring differences concentrate in a lower-dimensional direction. For regular densities and isotropic kernels, it is the natural local-geometric condition required to convert Euclidean locality into directional variation.

Under Assumption. 2, if $r \in C^1$, locally,

$$r(X_i) - r(X_j) = \nabla r(X_k)^T (X_i - X_j) + o(\sqrt{h}),$$

and therefore

$$\left[r(X_i) - r(X_j)\right]^2 = \left[\nabla r(X_k)^T (X_i - X_j)\right]^2 + o(h).$$

Consequently,

$$a_X h\mathbb{E}||\nabla r(Z)||^2 + o(h) \le \mathbb{E}\left[\Omega_{r,n}^X(h)\right] \le b_X h\mathbb{E}||\nabla r(Z)||^2 + o(h), \tag{25}$$

or equivalently,

$$\mathbb{E}\left[\Omega_{r,n}^X(h)\right] = \Theta\left(h\mathbb{E}||\nabla r(X)||^2\right) + o(h). \tag{26}$$

The C^1 regularity used in this subsection is required for deriving the first-order rate characterization in Eq.26, and will be used in the analysis of overselection below. It is not required for the basic consistency of the sufficient representation in Proposition .2, which only relies on boundedness and almost-everywhere continuity.

A.4 Rejection of Underselection

Let X^* denote the sufficient feature subspace that contains all variables needed to determine the response, and suppose for clarity that the underlying system is noiseless,

$$Y = m(X^*) \tag{27}$$

Consider a strict subspace

$$X^- \subsetneq X^*.$$

Define

$$V^-(x^-) = \mathrm{Var}\left(m(X^*) \mid X^- = x^-\right). \tag{28}$$

Assumption 3. *(Relevant omitted information). There exists* $\eta > 0$ *such that*

$$V^-(X^-) \ge \eta \quad \textit{almost surely} \tag{29}$$

This assumption states that the omitted variables contain response-relevant information that cannot be recovered from the retained representation.

Proposition 2. *(Underselection rejection). Suppose that the underlying system is noiseless,*

$$Y = m(X^*),$$

where m is bounded and continuous P_{X^*}-almost everywhere. Under Assumptions 1 and 3, for every fixed $c \in (0, 1)$,

$$\liminf_{n\to\infty} I_n^{X^-}(c) \geq \eta(1-c) > 0, \tag{30}$$

whereas

$$I_n^{X^*}(c) \to 0. \tag{31}$$

Consequently, for sufficiently large n,

$$I_n^{X^*}(c) < I_n^{X^-}(c). \tag{32}$$

If, in addition, m is Lipschitz continuous and Assumption 2 holds, then

$$I_n^{X^*}(c) = O\Big(h_n^{X^*}(c)\Big) = O\Big(n^{-2/d^*}\Big). \tag{33}$$

Proof. When the data are organized by X^-,

$$m^-(X^-) = \mathbb{E}\left[m(X^*) \mid X^-\right],$$

and

$$v_{X^-}(X^-) = V^-(X^-).$$

By Proposition. 1,

$$\mathbb{E}\left[\hat{I}_n^{X^-}(h,\Sigma)\right] = \mathbb{E}\left[\Omega^{X^-}_{m^- - \Sigma X^-,n}(h)\right] + \mathbb{E}\left[\frac{1}{n}\sum_{k,i} W_{ki}^{X^-}(1 - W_{ki}^{X^-})V^-(X_i^-)\right].$$

The first term is non-negative. Therefore, using Assumption. 3,

$$\mathcal{I}_n^{X^-}(c) \geq \frac{\eta}{n}\mathbb{E}\left[\sum_k \left(1 - \sum_i (W_{ki}^{X^-})^2\right)\right].$$

Since the self-weight is the largest weight in every row of Gram matrices induced by Gaussian kernel,

$$\forall i, \quad W_{ki}^{X^-} \leq W_{kk}^{X^-},$$

and hence

$$\sum_i \left(W_{ki}^{X^-}\right)^2 \leq W_{kk}^{X_-} \sum_i W_{ki}^{X^-} = W_{kk}^{X^-}.$$

It follows that

$$\mathcal{I}_n^{X^-}(c) \geq \frac{\eta}{n}\sum_k (1 - W_{kk}^{X^-}) = \eta(1-c) > 0.$$

However, for the full sufficient feature subspace X^*,

$$\mathrm{Var}(Y \mid X^*) = 0.$$

Therefore, by Proposition.1,

$$I_n^{X^*}(c) = \inf_{\Sigma} \mathbb{E}\left[\Omega^{X^*}_{m-\Sigma X^*,n}\left(h_n^{X^*}(c)\right)\right]. \tag{34}$$

Since the infimum is no larger than the value obtained at $\Sigma = 0$,

$$I_n^{X^*}(c) \leq \mathbb{E}\left[\Omega^{X^*}_{m,n}\left(h_n^{X^*}(c)\right)\right]. \tag{35}$$

For a query sample X_k, define the corresponding local variation

$$V_{k,n}(h) := \frac{1}{2}\sum_{i,j} W_{ki}^{X^*}(h) W_{kj}^{X^*}(h)\left[m(X_i) - m(X_j)\right]^2. \tag{36}$$

Then

$$\Omega_{m,n}^{X^*}(h) = \frac{1}{n}\sum_{k=1}^{n} V_{k,n}(h).$$

Using

$$(a-b)^2 \le 2(a-c)^2 + 2(b-c)^2,$$

together with $\sum_i W_{ki}^{X^*} = 1$, we obtain

$$V_{k,n}(h) \le 2\sum_i W_{ki}^{X^*}(h)\left[m(X_i) - m(X_k)\right]^2. \tag{37}$$

Consider a point x at which m is continuous. For any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\|u - x\| < \delta \quad \Longrightarrow \quad |m(u) - m(x)|^2 < \varepsilon.$$

Conditioning on $X_k = x$, we split the weighted sum into samples inside and outside the δ-neighborhood of x:

$$\begin{aligned}
&\sum_i W_{ki}^{X^*}(h)[m(X_i) - m(x)]^2 \\
&\quad = \sum_{\|X_i - x\| < \delta} W_{ki}^{X^*}(h)[m(X_i) - m(x)]^2 + \sum_{\|X_i - x\| \ge \delta} W_{ki}^{X^*}(h)[m(X_i) - m(x)]^2.
\end{aligned} \tag{38}$$

The first term is bounded by ε.

Since m is bounded, let $|m(x)| \le M$. For the second term,

$$[m(X_i) - m(x)]^2 \le 4M^2.$$

Moreover, because the denominator of $W_{ki}^{X^*}$ contains the self-kernel contribution equal to one,

$$\sum_{\|X_i - x\| \ge \delta} W_{ki}^{X^*}(h) \le n \exp\left(-\frac{\delta^2}{2h}\right). \tag{39}$$

By Lemma 1,

$$h_n^{X^*}(c) = \Theta\left(n^{-2/d^*}\right),$$

and hence

$$n \exp\left(-\frac{\delta^2}{2h_n^{X^*}(c)}\right) \to 0. \tag{40}$$

It follows that, at every continuity point of m,

$$V_{k,n}\left(h_n^{X^*}(c)\right) \to 0. \tag{41}$$

Since m is continuous P_{X^*}-almost everywhere, this convergence holds for P_{X^*}-almost every query point. Furthermore, boundedness of m provides a uniform bound on $V_{k,n}$. Therefore, by dominated convergence,

$$\mathbb{E}\left[\Omega_{m,n}^{X^*}\left(h_n^{X^*}(c)\right)\right] \to 0. \tag{42}$$

Consequently,

$$I_n^{X^*}(c) \to 0. \tag{43}$$

□

Under the stronger assumption that m is Lipschitz continuous with constant L, a quantitative rate follows from Assumption. 2. Indeed,

$$[m(X_i) - m(X_j)]^2 \le L^2\|X_i - X_j\|^2,$$

and therefore

$$V_{k,n}(h) \le \frac{L^2}{2} \sum_{i,j} W_{ki}^{X^*} W_{kj}^{X^*} \|X_i - X_j\|^2 \tag{44}$$

$$= L^2 \operatorname{tr}\left(Q_k^{X^*}(h)\right). \tag{45}$$

By Assumption. 2,

$$\mathbb{E}\left[V_{k,n}(h_n^{X^*}(c))\right] = O\left(h_n^{X^*}(c)\right).$$

Hence

$$I_n^{X^*}(c) = O\left(h_n^{X^*}(c)\right) = O\left(n^{-2/d^*}\right). \tag{46}$$

The mechanism behind this underselection rejection is therefore a limit separation: omitting useful information creates irreducible conditional variation of order $O(1)$, whereas the variation associated with the sufficient feature subspace vanishes as the sample size grows.

A.5 Rejection of Overselection

We next consider an overselected representation

$$X^+ = (X^\star, X^R),$$

where X^R contains redundant variables satisfying

$$X^R \perp X^*, \qquad Y = m(X^*).$$

Thus

$$Y \perp X^R \mid X^*.$$

Let

$$d^+ = d^* + d^R > d^*.$$

Unlike underselection, adding redundant variables introduces no irreducible conditional variance. The distinction instead arises because equal model complexity requires a wider bandwidth in the enlarged feature space.

Assumption 4. *(Non-affine signal). The regression function $m \in C^1$ satisfies*

$$J_m := \inf_{\Sigma} \mathbb{E}\left[||\nabla m(X^*) - \Sigma||^2\right] > 0$$

The condition excludes the degenerate cases in which m is exactly affine and therefore $\mathcal{I}_n^X(c) \to 0$ when data are organized by any sufficient feature subspace.

Proposition 3. *(Overselection rejection). Suppose Assumptions.1, 2, and 4 hold for both X^* and X_+. Then for every fixed $c \in (0,1)$,*

$$\frac{h_n^+(c)}{h_n^*(c)} = \Theta(n^{2/d^* - 2/d^+}) \to \infty.$$

Consequently,

$$\frac{\mathcal{I}_n^{X_+}(c)}{\mathcal{I}_n^{X^*}(c)} \to \infty$$

up to the non-degenerate constants specified in Assumption.2, and therefore for sufficiently large n,

$$\mathcal{I}_n^{X^*}(c) < \mathcal{I}_n^{X_+}(c).$$

Proof. For the sufficient feature subspace X^*, define

$$r_{\Sigma^*}(x^*) = m(x^*) - \Sigma^* x^*.$$

Its gradient is

$$\nabla r_{\Sigma^*}(X^*) = \nabla m(x^*) - \Sigma^*.$$

By Assumption. 2,

$$\mathbb{E}\left[\Omega^{X^*}_{r_{\Sigma^*},n}(h^*_n)\right] \leq b^*(c)h^*_n(c)||\nabla m(X^*) - \Sigma^*||^2 + o(h^*_n(c)).$$

Taking the infimum over Σ^*,

$$\mathcal{I}^{X^*}_n(c) \leq b^*(c)h^*_n(c)J_m + o(h^*_n(c))$$

Now consider the enlarged feature subspace $X^+ = \left(X^*, X^R\right)$. For

$$\Sigma^+ = \left(\Sigma^*, \Sigma^R\right),$$

the corresponding residual is

$$r_{\Sigma^+}(x^*, x^R) = m(x^*) - \Sigma^* x^* - \Sigma^R x^R.$$

Its gradient is

$$\nabla r_{\Sigma^+} = \left(\nabla m(x^*) - \Sigma^*, -\Sigma^R\right)^T.$$

Therefore,

$$\inf_{\Sigma^+} \mathbb{E}||\nabla r_{\Sigma^+}||^2 = \inf_{\Sigma^*.\Sigma^R} \left\{\mathbb{E}||\nabla m(X^*) - \Sigma^*||^2 + ||\Sigma^R||^2\right\} = J_m.$$

The redundant variable cannot reduce the irreducible nonlinear effects.

Using the lower bound in Assumption.2,

$$\mathcal{I}^{X^+}_n \geq a^+(c)h^+_n(c)J_m + o(h^+_n(c)).$$

Consequently,

$$\frac{\mathcal{I}^{X^+}_n}{\mathcal{I}^{X^*}_n} \geq \frac{a^+(c)}{b^*(c)} \frac{h^{X^+}_n}{h^{X^*}_n}(1 + o(1))$$

By Lemma. 1,

$$\frac{h^+_n(c)}{h^*_n(c)} = \Theta(n^{2/d^* - 2/d^+}) \to \infty$$

Hence, data organized by an overselected feature subspace have asymptotically large variation at equal model complexity. □

The mechanism behind overselection rejection is therefore a rate separation rather than a limit separation.

A.6 Consequence for the numerically evaluated AULC

In practice, AULC is evaluated on a finite set of non-degenerate complexity levels

$$0 < c_1 < \cdots < c_m < 1$$

using the trapezoidal rule, with the numerical endpoints defined as $I^X_n(0) = I^X_n(c_1)$ and $I^X_n(1) = 0$.

Let

$$\mathrm{AULC}^X_n = \sum_{j=0}^{m} \frac{I^X_n(c_j) + I^X_n(c_{j+1})}{2}(c_{j+1} - c_j),$$

where $c_0 = 0$ and $c_{m+1} = 1$.

Corollary 1. ***(Subspace preference of discretized AULC).*** *Under the assumptions of Proposition. 2, for every fixed finite complexity grid $\{c_j\}_{j=1}^m \subset (0,1)$ and every underselected representation X^- satisfying Assumption.3, there exists N^- such that, for all $n > N^-$,*

$$\mathrm{AULC}^{X^*}_n < \mathrm{AULC}^{X^-}_n.$$

Likewise, under the assumptions of Proposition.3, for every overselected representation X^+ satisfying Assumption 4, there exists N^+ such that, for all $n > N^+$,

$$\mathrm{AULC}^{X^*}_n < \mathrm{AULC}^{X^+}_n.$$

Algorithm 1 Estimating the learning curve and AULC

Input: Sample set with selected features $\mathscr{S}_{\text{selected}} = (X_{\text{selected}}, Y)$, and a sequence of target complexities $C_1 < \cdots < C_m$ in $(0, 1)$ ▷ C_1 should be close to 0

Output: Learning-curve points $\{(C_i, \hat{I}(C_i))\}$ and AULC

1: **for each** C_i **do**

2: Find a bandwidth h_i satisfying $C(\mathscr{S}; h_i, \Sigma = I) = C_i$ using Eq.8, if such an h_i exists ▷ Bisection

3: Compute

$$\hat{I}(C_i) \leftarrow \min_{\Sigma_{XY}} \hat{I}(\mathscr{S}; h = h_i, \Sigma = I, \Sigma_{XY})$$

using Eq.7 ▷ Newton's method

4: **end for**

5: Add the endpoint $(1, 0)$ to the learning curve

6: Add the endpoint $(0, \hat{I}(C_{\min}))$, where $C_{\min}$ is the smallest attainable complexity

7: Compute

$$\text{AULC} \leftarrow \sum_i \frac{\hat{I}(C_i) + \hat{I}(C_{i+1})}{2}(C_{i+1} - C_i)$$

▷ Trapezoidal rule

8: **return** $\{(C_i, \hat{I}(C_i))\}$ and AULC

B Algorithms and Numerical Methods

B.1 Calculating AULC

To compute AULC for a given system, we first specify a grid of target model complexities. Unless otherwise stated, we use

$$C = [0.005,\ 0.05,\ 0.10,\ 0.15, \ldots, 0.95,\ 0.995].$$

For each target complexity C, the corresponding kernel bandwidth h is obtained by solving Eq.8. Since the attainable complexity range depends on the finite sample, some predefined values may not admit a corresponding bandwidth and are therefore omitted. Our analysis suggests that, under asymptotic conditions, a single complexity level can in principle be sufficient. In practice, however, we use the above grid to evaluate AULC across a broader range of model complexities, since incorporating additional complexity levels increases the computational cost only linearly.

For each attainable complexity level, the variation $\hat{I}$ is minimized over Σ_{XY} according to Eq.7, yielding a point $(C, \hat{I}(C))$ on the learning curve. Before evaluating Eq.9 using the trapezoidal rule, we augment the curve with the endpoints

$$(1, 0) \quad \text{and} \quad (0, \hat{I}(C_{\min})),$$

where $C_{\min}$ denotes the smallest attainable complexity. The first endpoint represents the limiting case in which individual samples are effectively memorized, whereas the second extends the estimate at the lowest attainable complexity to $C = 0$. Alternatively, the latter endpoint may be estimated by linear regression. In practice, this choice has only a limited effect on the resulting AULC, while ensuring that AULC is evaluated over the common complexity range $[0, 1]$ across datasets.

The numerical integration error can be controlled by refining the complexity grid. An empirical analysis of the computational cost is provided in Fig.9.

B.2 Backward Elimination

Backward elimination is used to identify the feature representation that minimizes AULC. Ideally, finding the optimal feature subspace would require evaluating Eq.10 over all possible candidate subsets. However, exhaustive search becomes computationally prohibitive as the number of candidate

Algorithm 2 AULC-based backward elimination for feature subspace discovery

Input: Sample set $\mathscr{S} = (X, Y)$ and examined model complexities $C \in (0, 1)$
Output: Selected feature subset X^*, AULC sequence $\mathrm{S}_{\mathrm{AULC}}$, and feature-removal sequence $\mathrm{S}_{\mathrm{Feature}}$

$X_{\mathrm{current}} \leftarrow X$
$\mathrm{S}_{\mathrm{AULC}} \leftarrow [\mathrm{AULC}(X_{\mathrm{current}}, Y)]$
$\mathrm{S}_{\mathrm{Feature}} \leftarrow []$
while $|X_{\mathrm{current}}| > 1$ **do**
    for each feature $X_{\mathrm{test}} \in X_{\mathrm{current}}$ **do**
        $X_{\mathrm{selected}} \leftarrow X_{\mathrm{current}} \setminus \{X_{\mathrm{test}}\}$
        Compute $\mathrm{AULC}(X_{\mathrm{selected}}, Y)$ using Algorithm 1
    end for
    Let X_{remove} be the feature whose removal yields the smallest AULC
    $X_{\mathrm{current}} \leftarrow X_{\mathrm{current}} \setminus \{X_{\mathrm{remove}}\}$
    Append X_{remove} to $\mathrm{S}_{\mathrm{Feature}}$
    Append the corresponding minimum AULC to $\mathrm{S}_{\mathrm{AULC}}$
end while
Let k be the index of the minimum value in $\mathrm{S}_{\mathrm{AULC}}$
$X_{\mathrm{redun}} \leftarrow$ the first $k-1$ elements of $\mathrm{S}_{\mathrm{Feature}}$
$X^* \leftarrow X \setminus X_{\mathrm{redun}}$
Reverse $\mathrm{S}_{\mathrm{Feature}}$ ▷ More important features appear earlier
return X^*, $\mathrm{S}_{\mathrm{Feature}}$, $\mathrm{S}_{\mathrm{AULC}}$

features increases, motivating the use of a more scalable search strategy. We therefore adopt backward elimination, which starts from the full feature set and progressively removes variables, making it better suited than forward greedy search for preserving higher-order interactions among features.

Starting from the full set of observables, each remaining feature is tentatively removed in turn and the criterion is recomputed for the resulting subset. The feature whose removal produces the smallest criterion value is permanently eliminated, and the procedure is repeated until only one feature remains. The subset corresponding to the minimum value encountered along this elimination path is selected as the final representation.

The same search strategy is used for ANN-BECV and KRR-BECV, with cross-validated prediction error replacing AULC as the selection criterion. Because features that are more important to the representation tend to survive until later stages of backward elimination, the reverse removal order also provides an approximate ranking of feature importance.

B.3 Computational Complexity

The computation required for AULC-based relation discovery is highly parallelizable. For a fixed feature subset, different points on the learning curve can be evaluated independently. During backward elimination, the AULC values obtained by temporarily removing different candidate features can likewise be computed in parallel. Our implementation parallelizes the latter operation.

To characterize the computational requirements of AULC, we measure runtime and peak memory usage for systems with different sample sizes and numbers of candidate features. All results in Fig. 9 are averaged over five independent realizations. Experiments were conducted in MATLAB R2024b on a desktop equipped with a 13th Gen Intel(R) Core(TM) i7-1370P 1.90 GHz CPU and 32 GB RAM. Parallel computing was disabled for these experiments, such that the reported runtime corresponds to the cost of a single worker.

The computation of AULC consists of two main phases:

- **Phase I:** searching for the bandwidth h corresponding to the default complexity grid $0.05, 0.10, \ldots, 0.95 \cup 0.005, 0.995$;
- **Phase II:** optimizing the variation and constructing the corresponding learning curve.

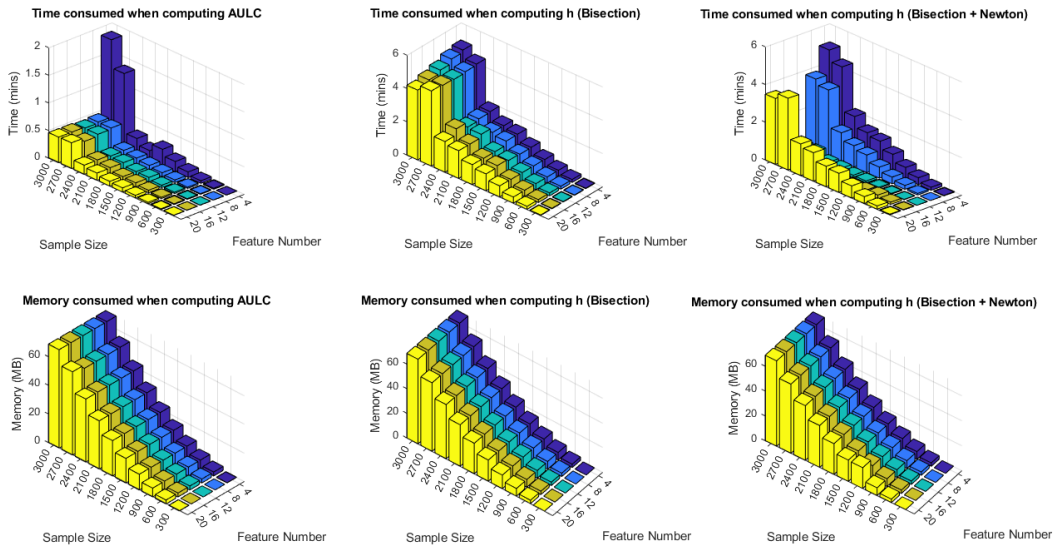

Figure 9: Computational cost of AULC.

Fig. 9 compares the computational costs associated with these operations. The top row reports runtime, while the bottom row reports peak memory usage. From left to right, the panels correspond to the complete AULC computation, bandwidth estimation using bisection, and bandwidth estimation using the hybrid bisection–Newton strategy.

Peak memory usage is governed primarily by sample size and remains relatively similar across the different computational stages. Runtime shows a stronger dependence on the particular operation, with the bandwidth search in Phase I generally constituting a substantial fraction of the total cost. Replacing pure bisection with the hybrid bisection–Newton strategy considerably reduces the bandwidth-search time. For Phase II, we additionally observe that, at a fixed sample size, runtime does not necessarily increase with the number of candidate features and may even decrease in some settings.

B.4 High-dimensional distance metrics and hyper-parameter selection

When extending AULC to higher-dimensional input spaces, the definition of neighborhood geometry becomes increasingly important. Different input variables may exhibit distinct scales and correlation structures, such that a standard Euclidean metric may assign disproportionate influence to particular directions. In the general kernel formulation, neighborhood relations are instead characterized through the squared Mahalanobis distance

$$d^2 = \Delta X^T \Sigma^{-1} \Delta X, \tag{47}$$

where the positive-definite matrix Σ determines how different directions in the input space are scaled and combined.

Fig. 10 illustrates this geometric interpretation in a two-dimensional input space. As in the one-dimensional case, samples can be viewed geometrically to examine whether the input representation induces a structured relationship with the output. The main distinction in higher dimensions is that differences across multiple input coordinates must be combined into a single notion of distance. The Mahalanobis metric provides such a construction and is naturally compatible with the Gaussian kernel used in AULC.

The role of Σ can be made explicit by decomposing its inverse as

$$\Sigma^{-1} = U^T \Lambda^2 U, \tag{48}$$

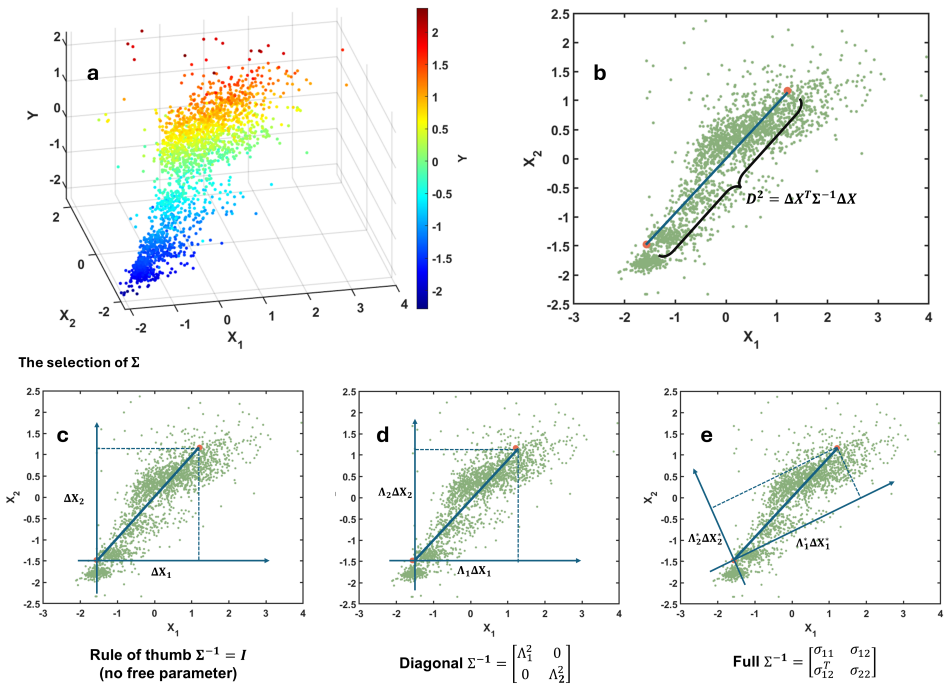

Figure 10: Geometric interpretation of the Mahalanobis metric.

which gives

$$d^2 = \Delta X^T U^T \Lambda^2 U \Delta X = \sum_i \left(\Lambda_i U_{i,\cdot} \Delta X\right)^2 . \tag{49}$$

Therefore, the Mahalanobis distance is equivalent to applying the linear transformation ΛU to the original coordinates and subsequently measuring standard Euclidean distance in the transformed space.

The examples in Fig. 10c–e illustrate progressively more flexible choices of this geometry. An identity metric corresponds to ordinary Euclidean distance. Allowing diagonal elements to vary permits different directions to be stretched independently. A fully parameterized metric further allows rotations and other linear coordinate transformations, enabling the neighborhood geometry to adapt to anisotropic structure in the data. The same interpretation extends directly to higher-dimensional systems.

Treating Σ as a tunable hyper-parameter increases the representational flexibility of the kernel, but also introduces additional optimization cost. Table 1 summarizes four possible configurations of the metric-related hyper-parameters. From Type I to Type IV, progressively more parameters are optimized, providing increasingly flexible neighborhood geometries at the expense of greater computational complexity. When Σ is fixed, the resulting optimization is either absent or can be efficiently solved using Newton's method with an analytically available Hessian. Allowing Σ itself to vary instead leads to constrained optimization problems with substantially more complicated Hessian structures, for which we employ interior-point methods with BFGS approximations.

In this study, we adopt the Type-II configuration, in which all input variables are first z-standardized, the input-space metric Σ is fixed to the identity matrix, and the input–output term Σ_{XY} is estimated from the data. This configuration introduces a local first-order input–output adjustment without requiring optimization of the full input-space metric, providing a practical compromise between structural flexibility, comparability across datasets, and computational tractability. More flexible parameterizations of Σ may be beneficial for systems with pronounced anisotropy, but their systematic evaluation is beyond the scope of this work.

Table 1: Optimization configurations for the metric parameters.

Index	Σ	Σ_{XY}	Variable number	Optimization Problem Type	Constraints	Solving Methods
I	Rules of thumb	Zero	0	Null	Null	Null
II	Rules of thumb	Non-zero	d	Unconstrained problem with known Hessian	Null	Newton's methods
III	Diagonal	Non-zero	2d	Inequality-constrained with complex Hessian	Positive diagonal terms in Σ	Interior-point methods with BFGS
IV	Full	Non-zero	d(d+3)/2	Inequality-constrained with complex Hessian	Positive-definite and symmetric Σ	Interior-point methods with BFGS

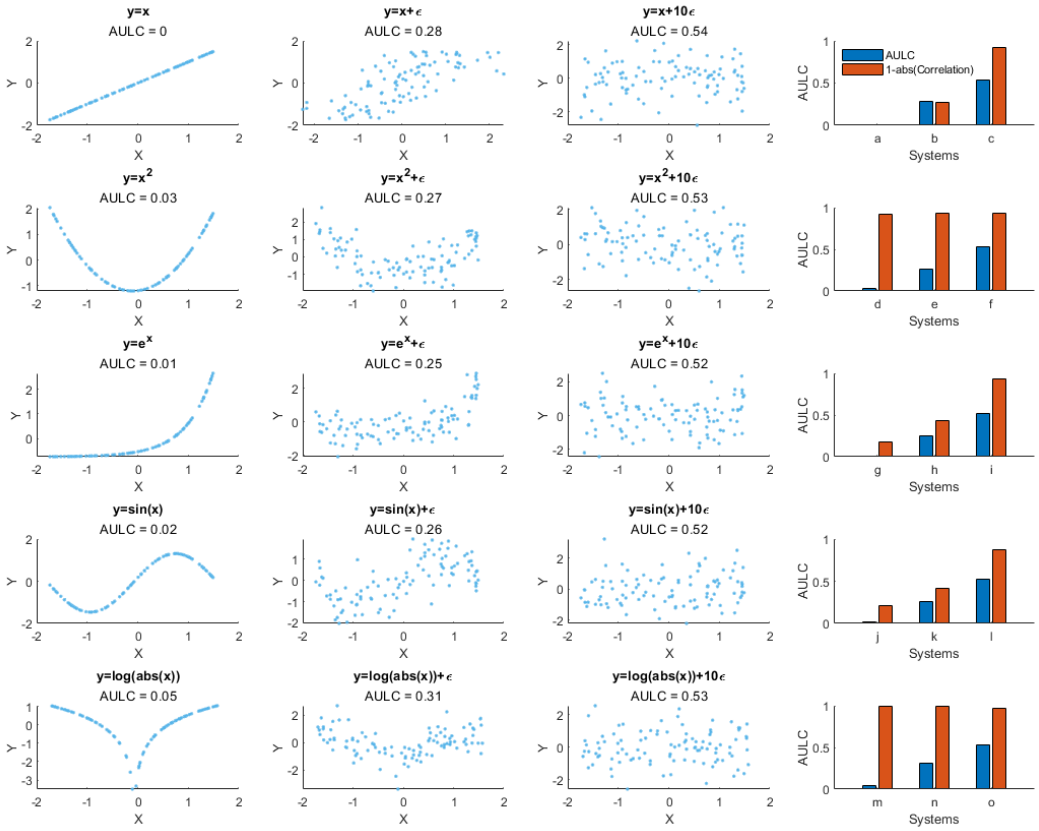

Figure 11: AULC values for different example systems.

C Other Examples

C.1 Systems with Noise

We further examine the behavior of AULC under different levels of observational noise. Fig. 11 shows several representative systems constructed from different base functions. Each row corresponds to the same underlying system, while the columns represent increasing levels of additive Gaussian noise, denoted by ϵ. For each system, the generated samples are first z-standardized, after which noise is added. This procedure ensures that the noise levels are comparable across systems and are not rescaled by the subsequent standardization. The resulting AULC values therefore illustrate how the estimated learnability of each system changes as the signal-to-noise ratio decreases.

C.2 The Friedman #1 Benchmark

We further evaluate AULC-BE feature discovery using the Friedman #1 benchmark, as summarized in Fig. 12. The system contains 10 candidate variables, of which only five determine the target

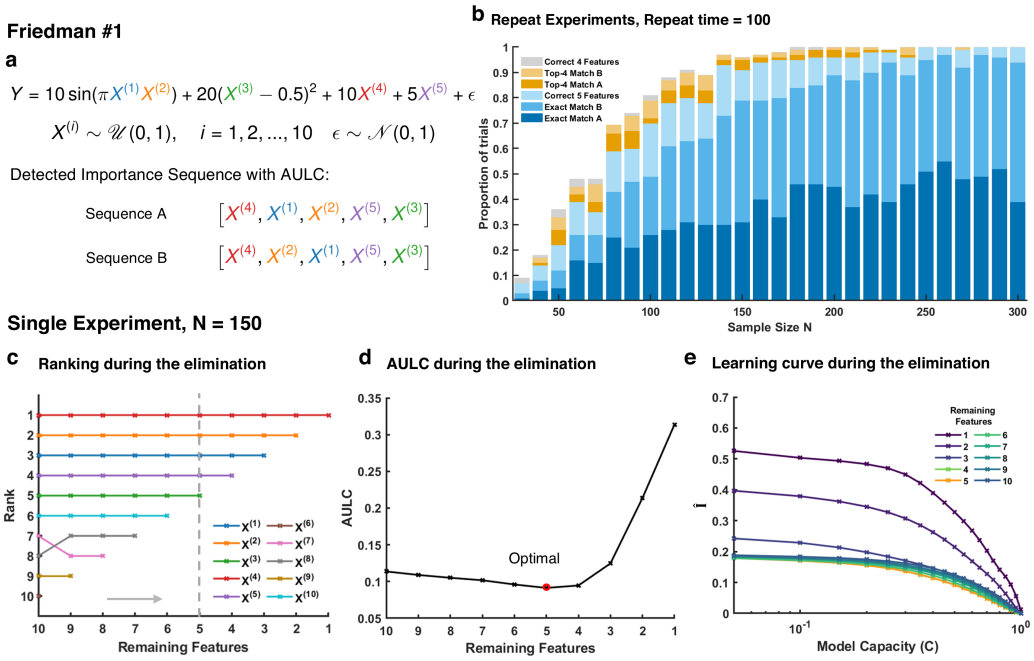

Figure 12: AULC on the Friedman #1 benchmark.

variable according to the governing equation. Across repeated experiments, the two most frequently observed feature-importance orderings produced by AULC are denoted as Sequence A and Sequence B in Fig. 12a.

Fig. 12b summarizes the results over 100 independent trials at different sample sizes. The stacked bars report the proportions of runs that follow Sequence A, Sequence B, or retain different subsets of the true variables. AULC is able to recover all five effective variables with relatively limited data, including at $N = 150$, while the frequency of the two dominant elimination sequences increases with sample size.

Fig. 12c–e presents a representative backward-elimination trajectory for $N = 150$. At each step, AULC is recomputed after temporarily removing each remaining feature, and the feature whose removal yields the smallest AULC is permanently eliminated. Fig. 12c shows the resulting evolution of feature ranks during this procedure. The dashed vertical line marks the point at which the minimum overall AULC is reached with five remaining variables, corresponding to the ground-truth feature subspace size.

The associated AULC trajectory is shown in Fig. 12d. In this example, removing the relatively weak true variable $X^{(3)}$ produces only a small increase in AULC, illustrating that different informative variables can contribute unequally to the overall learnability of the system. Fig. 12e further shows the corresponding learning curves throughout the elimination process, providing a direct view of how the underlying learnability changes as irrelevant and subsequently informative variables are removed.

D BASELINE IMPLEMENTATION DETAILS

Symbolic Regression (SR). Symbolic regression was implemented using PySR. For classification tasks, the symbolic expression was interpreted as a logit and optimized using binary cross-entropy. Binary problems were fitted using a single symbolic classifier. The final selected feature set was defined as the union of variables appearing in the selected symbolic expressions. We used 100 PySR iterations with the binary operators

$$+,\ -,\ \times,\ /,\ \wedge,$$

and the unary operators square, cube, square root, exponential, logarithm, sine, cosine, tangent, absolute value, and inverse sine. The final expression was selected using PySR's `best` model-selection rule.

Random Forest with Recursive Feature Elimination and Cross-Validation (RF-RFECV). RF-RFECV was implemented using a random-forest classifier together with recursive feature elimination and cross-validation. The random forest contained 500 trees with no explicit restriction on tree depth. Features were removed one at a time, and the elimination path was evaluated using stratified cross-validation with classification accuracy as the scoring criterion. Five folds were used whenever permitted by the class frequencies; otherwise, the number of folds was reduced to the number of samples in the smallest class. The feature subset achieving the best cross-validation performance along the RFECV path was returned as the selected representation.

Artificial Neural Network with Backward Elimination and Cross-Validation (ANN-BECV). ANN-BECV used a shallow feed-forward classification network with one hidden layer containing 16 neurons. The network was implemented using MATLAB patternnet, with scaled conjugate-gradient optimization, cross-entropy loss, and a softmax output layer. Training was limited to 200 epochs with a target training error of 10^{-6}. Model performance was evaluated using five-fold stratified cross-validation, with the mean held-out classification accuracy across folds used as the selection criterion.

Before backward elimination, candidate predictors were z-standardized. Starting from the full feature set, each remaining feature was tentatively removed in turn, and the corresponding cross-validation score was recomputed. The feature whose removal yielded the highest classification accuracy was permanently eliminated. This procedure was repeated until only one feature remained. The final representation was chosen as the subset with the highest cross-validation score encountered along the elimination path. At each elimination level, the same random seed was used for all candidate deletions to reduce variability caused by network initialization.

Kernel Ridge Regression with Backward Elimination and Cross-Validation (KRR-BECV). KRR-BECV followed the same greedy backward-elimination procedure, with predictive cross-validation performance used as the feature-subset criterion. An RBF kernel ridge model was employed. Candidate predictors were median-imputed when necessary and subsequently standardized. Hyper-parameters were re-tuned after an accepted elimination and then fixed when comparing candidate deletions within the same elimination step. Under the standard search configuration, the regularization parameter was selected from

$$\alpha \in \left\{10^{-3},\ 10^{-2},\ 10^{-1},\ 1,\ 10\right\},$$

while the RBF parameter was selected from

$$\gamma \in \left\{\frac{0.1}{p},\ \frac{1}{p},\ \frac{10}{p}\right\},$$

where p denotes the number of features in the candidate subset.

At each backward-elimination step, all possible one-feature deletions were evaluated, and the deletion yielding the highest cross-validation score was accepted. After constructing the complete elimination path, the subset with the highest cross-validation performance was selected as the final representation. For binary classification, class labels were fitted as $0/1$ targets and predictions were obtained using a threshold of 0.5. For multiclass classification, one-hot encoded targets were fitted and the class corresponding to the largest predicted score was selected.

LassoNet. LassoNet was implemented using its cross-validated regularization-path estimator. Candidate predictors were median-imputed and z-standardized before fitting. We used a neural network with a single hidden layer containing 16 units. The regularization path used a multiplicative path factor of 1.02 and $\gamma = 0$. Training used up to 1000 iterations for the initial dense model and 100 iterations for each regularized path point, with patience parameters of 100 and 10, respectively.

The regularization level, and therefore the retained feature subset, was selected according to cross-validation performance along the LassoNet path. To ensure consistency with the backward-elimination methods, which search only over non-empty feature subsets, the final LassoNet solution was constrained to retain at least one feature. If the unconstrained cross-validation optimum corresponded to the null model, the highest-scoring non-empty state along the regularization path was selected instead.

Regression Tasks. For regression tasks, the same feature-selection procedures and model configurations were used, with classification-specific objectives and stratified cross-validation replaced by their corresponding regression counterparts.

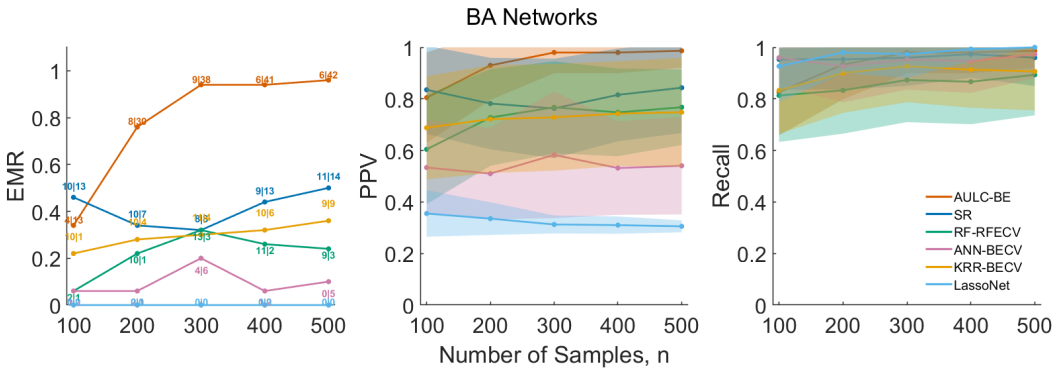

Figure 13: Identifying the controlling condition for SIS processes on BA networks, with 50 realizations.

E Data Generation

E.1 Feynman Physics Dataset

The dataset includes 120 equations. For each equation, we generate datasets at different sample sizes n and test the performance of different methods. The detailed data generation process is as follows:

- We slightly modify expressions when necessary to avoid singularities and undefined values during sampling. For example, $\sqrt{x}$ is replaced by $\sqrt{|x|}$.
- The input variables appearing in each equation are randomly sampled and treated as the ground-truth informative variables.
- The output is generated by evaluating the corresponding equation on the sampled inputs.
- Noisy copies of informative variables and independently generated noise variables are added as spurious features.
- Outliers are removed using an interquartile-range criterion, after which the first n retained samples are used to construct the final dataset.

E.2 Madelon-like Systems

The Madelon benchmark is generated by first sampling vertices $V_i \in \{-1, 1\}^5$ on a 5D hypercube. For each sample, the output is generated through a selected parent vertex V_i:

$$Y_i = \text{sign}(\prod_{k=1}^{5} V_{i,k}) \tag{50}$$

The five effective variables are then generated from V_i with additive Gaussian noise η:

$$X_i^{(\text{eff})} = V_i + \eta. \tag{51}$$

The 15 redundant variables in this case are generated via a linear transform of the five effective variables and another additive Gaussian noise ϵ:

$$X^{(\text{redun})} = W X^{(\text{eff})} + \epsilon. \tag{52}$$

An additional 30 noise variables are added to each dataset. The methods are then evaluated by their ability to recover the five effective variables from a total of 50 variables.

E.3 SIS PROCESS ON NETWORKS

The control condition for the SIS process on a network G has been discussed in many previous works under different approximation conditions. In general, the condition for endemic persistence has the form:

$$\frac{\beta}{\mu} > f(G) \tag{53}$$

where β and μ denote the transmission rate and the recovery rate respectively. It may not be that surprising that there can be different versions of $f(G)$ when the process operates on networks with different properties. For example, previous work has shown that while $f(G) = 1/\lambda_{max}(A)$ is a more general control condition for finite-sized networks, $f(G) = 1/\langle k \rangle$ is a simplified version when the process operates on homogeneous Erdős–Rényi networks Wang et al. (2003).

Randomly sampling the network and SIS parameters can produce highly imbalanced datasets, with most realizations concentrated in either the active or inactive regime. To ensure sufficient coverage of both sides of the transition, we therefore use the established SIS control condition only to guide sample generation. Importantly, the corresponding control variable is not provided to the subsequent feature-discovery procedure and is used solely for constructing a balanced sampling distribution.

Specifically, for each sample, we generate an ER (BA) network with $N \in [301, 800]$ nodes and an expected mean degree between 4 and 20 for ER networks, or between 2 and 15 for BA networks. Several topological descriptors are then extracted from each network, including the leading adjacency eigenvalue $\lambda_{\max}$. The SIS parameters are sampled through the auxiliary control variable $q = \lambda_{\max}\beta/\mu$, with approximately equal probability from the subcritical region $q \in [0.05, 0.90]$ and the supercritical region $q \in [1.10, 4.00]$. After sampling $\mu \in [0.05, 0.50]$, the transmission probability is set as $\beta = q\mu/\lambda_{\max}$, subject to $10^{-4} \leq \beta \leq 0.5$. The initial infected fraction is independently sampled from $[0.02, 0.10]$.

For each network–parameter configuration, we simulate a discrete-time stochastic SIS process for 300 updates. At each step, infected nodes recover independently with probability μ, while a susceptible node i with $n_i^I(t)$ infected neighbors becomes infected with probability $1 - (1 - \beta)^{n_i^I(t)}$. A realization is classified as active if the mean infection density over the final 150 time steps exceeds 0.05, and inactive otherwise.

The candidate features supplied to the discovery algorithms are:

- Active label (output): 1 if active and 0 otherwise
- $\langle k \rangle$: the average degree of the generated network
- Var(k): the variance of degree of the generated network
- Maximum degree
- Degree heterogeneity Var$(k)/\langle k \rangle^2 + 1$
- Mean clustering coefficient
- Largest eigenvalue of adjacency matrix
- Average shortest path
- Transmission rate β
- Recovery rate μ
- Initial infected fraction ρ_0