CLASP: CORRELATION-LEVERAGED ACROSS-SAMPLE SEPARABILITY PRIORITIZATION FOR UNSUPERVISED BIOMARKER DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomedical data science operates in a regime of high data dimensionality, low sample sizes, and structured noise. As a result, many applications suffer from a reproducibility crisis, in which biomarkers identified in one study fail to replicate in others. To address this challenge, we introduce Correlation-Leveraged Across-sample Separability Prioritization (CLASP), an unsupervised feature selection framework that identifies reproducible feature subsets by maximizing cross-subsample cluster separability. CLASP frames feature selection as a stochastic optimization problem over inclusion probabilities; it uses a Cross-Entropy Method paired with a correlation-augmented Gumbel sampling scheme to find a joint feature set that maximizes cluster distinctiveness. Evaluated both on simulation environments and seven diverse real-world biomedical datasets, CLASP shows state-of-the-art feature selection stability, cluster quality, and out-of-sample cluster stability, thus preserving the reproducibility of selected biomarkers.

1 INTRODUCTION

Biomedical datasets are notoriously high-dimensional and suffer from the $p \gg n$ problem of having far more input features than data samples (Yang et al. (2021); Zitnik et al. (2019)). These datasets also contain high levels of noise relative to the condition under study. This environment is a key contributor to the ongoing "reproducibility crisis" in biology and medicine, in which biomarkers identified in one study fail to replicate across independent experiments and analyses (Rahnenführer et al. (2023); Perng & Aslibekyan (2020)). Examples include the inconsistent biomarkers linked to ADHD using the same publicly available dataset (Wang et al. (2017)) and the inability to identify stable genes that influence cancer outcome prediction from microarray data (Michiels et al. (2005))

One strategy to mitigate the $p \gg n$ problem is to select a small set of "informative features" for the analysis while discarding the remaining input features as potentially dominated by noise. While popular, many feature selection methods optimize signal variance, association with a predictive target, or preservation of estimated data structure; these criteria can lead to unstable results on biomedical datasets. High sample variances and correlated noise patterns, both of which are common in biomedical datasets, can confound variance- and structure-based algorithms by drowning out the meaningful signal (Nadler (2009); Xiu et al. (2025)). This scenario can produce fragile representations that fail to replicate across independent cohorts. Likewise, it is common for prediction-based algorithms to overfit to nuisance features due to finite sampling effects. This overfitting leads to "phantom biomarkers" that achieve good quantitative performance on a specific dataset but lack a biological foundation and fail to generalize to other cohorts (Wang et al. (2024)).

Alongside feature selection, another key goal of biomedical data analysis is to group samples into cohesive subtypes. For example, skin cells in a histopathology image might be grouped into distinct morphological populations, with each responding to a different intervention. Sparse clustering methods bridge the domains of feature selection and data stratification by simultaneously partitioning the dataset and identifying the subset of features that define the clusters. Sparse clustering methods commonly rely on penalty-based optimization frameworks (Witten & Tibshirani (2010)) or model-based formulations with sparsity-inducing priors (Bouveyron & Brunet-Saumard (2014)). However, structured noise and highly-correlated features can confound these methods (Cobey et al. (2024),

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.

Le et al. (2024)). As a result, minor perturbations in the dataset can lead to vastly different results, especially in high-dimensional and low sample size biomedical datasets.

This paper introduces Correlation-Leveraged Across-sample Separability Prioritization (CLASP), a framework for robust unsupervised feature selection in high-dimensional, low-sample regimes. Our unique strategy is to learn a sparse feature set that produces a *highly separable partitioning of the dataset across random subsamples*. Mathematically, CLASP casts this task as a stochastic optimization problem that maximizes the average silhouette score of the cluster partitions obtained from random subsamples of the data. The sampling procedure uses a correlation-augmented variant of Gumbel sampling (Jang et al. (2017); Huijben et al. (2022)) that exploits the network structure of the feature space and reduces the effective search dimension. We employ a Cross-Entropy Method (CEM) (De Boer et al. (2005)) to accommodate the non-differentiable objective and optimize the feature selection probabilities. The main contributions of CLASP are as follows:

1. Feature selection is optimized using cluster separability objective that favors highly separable groupings across random subsamplings of the dataset. We demonstrate that this strategy improves both the feature selection and clustering stability in noisy datasets.
2. An efficient stochastic optimization routine that optimizes our non-differentiable objective. Feature selection masks are generated using a correlation-augmented Gumbel sampling scheme, while CEM identifies the joint feature set that optimizes cross-subset consensus.
3. Extensive simulation study under various noise regimes demonstrating that CLASP can recover the ground-truth structure better than baseline methods. CLASP is on-par with baselines in the case of uncorrelated and low noise levels, while achieving up to 95% better feature recovery when noisy correlated structures overwhelm the true sparse signal.
4. Real-world evaluation on 7 biomedical datasets across distinct modalities that contain structured and high-variance noise. CLASP shows a 34% improvement in Nogueira feature selection stability across dataset perturbations, 13% in clustering stability, and 11% in Davies-Bouldoin Index (DBI) over the next highest-performing baseline.

2 Related Work

Unsupervised Feature Selection: Unsupervised feature selection (UFS) reduces the complexity of high-dimensional data by selecting a small number of discriminative features for downstream analysis (Dwivedi et al. (2024)). Broadly, UFS methods can be grouped into three categories based on the association that they optimize (Li et al. (2024)). Graph-based UFS methods are the most represented in the current literature; they use an assumed graph structures to select a sparse set of features that are most representative and informative of underlying data topology (Tang et al. (2023); He et al. (2005); Huang & Yang (2022)). Statistical UFS methods use the distribution of each feature within a dataset to discover the subset which maximizes said statistical objective (Shaham et al. (2022); He et al. (2011); Yuan et al. (2021)). Lastly, task-based UFS methods assess the significance of features for a downstream task and select the feature subsets that drive the performance (Zhu et al. (2019); Covões & Hruschka (2011); Hassan et al. (2022)). More recently, deep learning approaches have been proposed for UFS. The Concrete Autoencoder (Balın et al. (2019)) introduces a continuous relaxation of discrete feature selection masks to enable end-to-end backpropagation, while Restricted Autoencoders (Covert et al. (2020)) iteratively filter features based on reconstruction performance. While powerful, these overparameterized networks are prone to capturing spurious correlations and overfitting to noise in the $p \gg n$ regime of biomedical datasets.

To mitigate sensitivity to noise and sample variance, consensus-based UFS stabilizes selection across data subsamples, a strategy first introduced by Meinshausen and Buhlman (Meinshausen & Bühlmann (2010)). By penalizing features that only appear under specific data splits, the consensus loss reduces variance to select only features with robust signals across distributions. Following this work, consensus and ensemble-based approaches have become increasingly common for UFS (Hotta et al. (2026); Zhang & de Amorim (2026); Cao et al. (2023); Malarkkan et al. (2026)). While these sparse representations improve generalization to unseen data, existing methods target large sample sizes and do not extend to $p \gg n$ single-view settings. CLASP proposes the novel strategy of using *clustering stability across random subsamples* as an objective for the low-sample size regime, thereby overcoming the limitations of prior work.

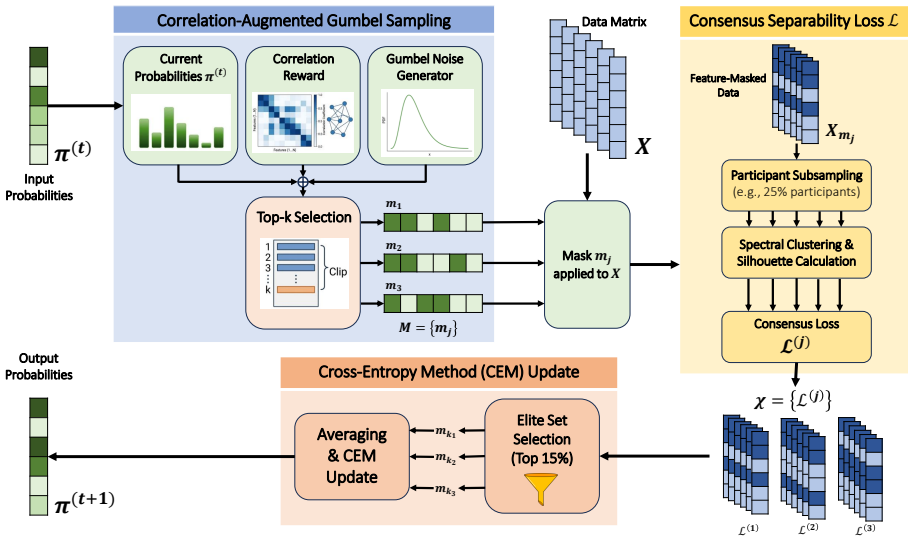


Figure 1: Overview of CLASP. The algorithm iteratively optimizes a feature selection probability vector $\pi^{(\mathbf{t})}$. In the forward pass, feature subsets $\mathbf{m}_j$ are drawn via a correlation-augmented Gumbel sampling procedure that encourages structurally cohesive networks. The quality of each feature mask is evaluated through the consensus separability loss based on the silhouette score across data subsamples. A CEM update is used to refine sampling probabilities for the next iteration.

Sparse Clustering: Sparse clustering methods combine the tasks of dataset stratification and feature selection. Classical approaches achieve this by adding a sparsity penalty to existing clustering algorithms (Witten & Tibshirani (2010)) and generative model formulations (Bouveyron & Brunet-Saumard (2014)). While conceptually appealing, these methods have limited success in biomedical regimes due the low SNR and $p \gg n$ problem (Duò et al. (2020); Kiselev et al. (2019)). In fact, minor dataset variations and hyperparameter tuning yield drastically different feature subsets, suggesting that the methods are influenced by cohort-specific artifacts rather than true biological signal.

Deep clustering methods attempt to bypass traditional feature selection by jointly learning latent representations and cluster assignments in end-to-end fashion(Min et al. (2018)). Sparse variants often rely on autoencoder bottlenecks or attention mechanisms to filter irrelevant dimensions. However, modern deep networks are overparameterized and prone to memorizing spurious sample-specific variations rather than generalizable phenotypes (Grinsztajn et al. (2022)). Hence, these methods can yield unstable results on biomedical datasets (Marrone et al. (2019), Bosma et al. (2024)).

Taken together, current methods do not adequately fulfill the need for stable, consensus-driven cluster representations in noisy, low-sample data. CLASP addresses this gap by shifting the paradigm from static penalty-based optimization to consensus-driven cross-subset separability.

3 CLASP FOR STABLE SEPARABILITY-DRIVEN UFS

Our CLASP framework combines the goals of UFS and sparse clustering in an iterative procedure that remains efficient in high-dimensional regimes. Unlike existing UFS methods, our strategy is to identify a subset of features that consistently produces a separable partitioning of the dataset, while preventing the overfitting of CLASP to cohort-specific noise. Formally, this is done by maximizing the average silhouette score of the clusterings obtained from different subsamples of the dataset. We stochastically optimize the feature inclusion probabilities, while accounting for the network structure of the feature set. Fig. 1 provides a graphical overview of our framework.

3.1 MATHEMATICAL FORMULATION AND OBJECTIVE FUNCTION

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be our data matrix, where n is the number of samples, and p is the number of features. From $\mathbf{X}$, we can obtain a feature correlation matrix $\mathbf{R} \in \mathbb{R}^{p \times p}$, where each element $\mathbf{R}_{ij}$

Algorithm 1 Stochastic optimization routine for CLASP. Given a data matrix, target feature count, and number of clusters, CLASP optimizes the feature selection probabilities and outputs a binary mask identifying a joint feature set that yields a separable partition consistent across subsamples.

Require: Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, feature count z, number of clusters K
Hyperparameters: Mask sample size M, elite fraction γ, correlation weight λ_{corr}, smoothing rate α, patience T
Ensure: Optimal binary feature mask $\mathbf{m}^* \in \{0,1\}^p$
Compute empirical correlation matrix $\mathbf{R} \leftarrow \text{Corr}(\mathbf{X}) \in \mathbb{R}^{p \times p}$
Initialize sampling distribution $\boldsymbol{\pi}^{(0)} \leftarrow \frac{z}{p}\mathbf{1}_p$
repeat
  for $j = 1$ **to** M **do**
    Sample noise vector $\mathbf{g} \sim \text{Gumbel}(0, \mathbf{I}_p)$
    Generate mask $\mathbf{m}_j \leftarrow \text{top-}z\left(\log(\boldsymbol{\pi}^{(t)}) + \lambda_{\text{corr}}|\mathbf{R}|\boldsymbol{\pi}^{(t)} + \mathbf{g}\right)$
    Evaluate separability loss $\mathcal{L}(\mathbf{m}_j)$ via Eq. 1
  end for
  Form elite set $\mathcal{E} \leftarrow$ top-γ proportion of $\{\mathbf{m}_j\}_{j=1}^M$ maximizing $\mathcal{L}$
  Update distribution: $\boldsymbol{\pi}^{(t+1)} \leftarrow (1-\alpha)\boldsymbol{\pi}^{(t)} + \frac{\alpha}{|\mathcal{E}|}\sum_{\mathbf{m}\in\mathcal{E}}\mathbf{m}$
until top-$z(\boldsymbol{\pi}^{(t)})$ mask remains unchanged for T consecutive iterations
return Final feature mask $\mathbf{m}^* \leftarrow \text{top-}z(\boldsymbol{\pi}^{(t)})$

is computed as the Pearson correlation of features i and j across data samples. CLASP outputs a feature selection probability vector $\boldsymbol{\pi} \in \mathbb{R}^{p\times 1}$ such that the subset of highest-weight features can reliably partition the samples of $\mathbf{X}$ into K clusters, where K is user- or application-defined.

The feature selection probabilities are stochastically optimized via a novel separability loss designed to prevent overfitting to cohort-specific noise. Let $\mathbf{m} \in \{0,1\}^{p\times 1}$ be a binary feature mask obtained from the current selection probabilities. In parallel, we define a set of data subsamples $S = \{s_1, ..., s_C\}$ and mask each ones using $\mathbf{m}$ to obtain a collection of sparse data subsamples $\{\mathbf{X}_{s_c}\}_{c=1}^C$. We use an off-the-shelf clustering algorithm to partition each subsample $\mathbf{X}_{\mathbf{s_c}}$ into K clusters. The objective function to optimize the feature selection probabilities is based on the silhouette scores of the masked data clustering. Mathematically, for each data point $\mathbf{x} \in \mathbf{X}_{s_c}$, let $a(\mathbf{x})$ and $b(\mathbf{x})$ denote the mean Euclidean distances to samples in the same cluster as $\mathbf{x}$ and to samples in the next-nearest cluster, respectively. The objective function to maximize cluster distinctiveness is

$$\mathcal{L}(\mathbf{m}) = \frac{1}{|s_c|}\sum_{\mathbf{x}\in\mathbf{X}_{s_c}} \frac{b(\mathbf{x}) - a(\mathbf{x})}{\max(a(\mathbf{x}), b(\mathbf{x}))}, \tag{1}$$

where $|s_c|$ is the number of data samples in the set s_c. The silhouette scores rewards data partitions that have low within-cluster variance and high between-cluster variance. At a high level, Eq. (1) evaluates the quality of the feature mask $\mathbf{m}$ by its ability to reliably partition the data subsamples into highly separable clusters. The underlying stochastic nature of Eq. (1) encourages generalizable solutions and prevents cohort-specific noise from dominating the optimization.

3.2 Optimization for the Non-Differentiable Objective Function

The objective function $\mathcal{L}(\mathbf{m})$ in Eq. (1) is non-differentiable due to the discrete masking procedure, and the mask search space is combinatorial in the number of features p. Therefore, we define a stochastic optimization routine to iteratively update the feature selection probability vector $\boldsymbol{\pi}_i^{(t)}$ at sequential iterations t. Once converged, we can use this vector to identify the subset of top features.

In lieu of a gradient step, we construct an empirical distribution of binary masks that is influenced by three factors: (1) the current selection probabilities $\boldsymbol{\pi}_i^{(t)}$ at iteration t, (2) the feature correlation structure implicit in $\mathbf{X}$, and (3) the cluster separability quantified by Eq. (1). This empirical distribution, termed the "elite set" is used to obtain $\boldsymbol{\pi}_i^{(t)}$ through a CEM update.

First, we generate M candidate masks via a correlation-augmented Gumbel sampling scheme:

$$\mathbf{m}_j = \text{top-}z\left\{\log(\boldsymbol{\pi}^{(t)}) + \lambda_{\text{corr}} \cdot |\mathbf{R}|\,\boldsymbol{\pi}^{(t)} + \mathbf{g}_j\right\} \qquad j = 1, \ldots, M, \tag{2}$$

where the function top-$z\{\cdot\}$ sets the z highest entries to 1 and zeros out the remaining entries, $\mathbf{R}$ is the feature correlation matrix defined in Section 3.1, λ_{corr} is a hyperparameter that controls the relative influence of the correlation structure, and the Gumbel noise $\mathbf{g}_j \sim \text{Gumbel}(0, 1)$ injects stochasticity. The logarithmic first term maps selection probabilities into log-odds to correctly calibrate the Gumbel noise, while the linear second term provides a direct proportional shift that rewards features based on their similarity to the currently active set without forcing a multiplicative relationship in the probability space. Intuitively, the $\log(\pi_i)$ term encourages the current set of important features to be selected, while the correlation reward is designed to mimic the structure of the dataset while exponentially reducing the effective search space for this combinatorial problem.

Sampling via Eq. (2) induces a discrete probability distribution $\mathcal{P}_\lambda(\mathbf{m} \mid \boldsymbol{\pi}, \lambda_{\text{corr}})$ over the space of valid masks $\mathcal{M} = \{\mathbf{m} \in \{0, 1\}^p : \|\mathbf{m}\|_1 = z\}$. We prove that the correlation-augmented distribution $\mathcal{P}_\lambda$ strictly contracts the effective search space relative to the unaugmented baseline distribution $\mathcal{P}_0$ ($\lambda_{\text{corr}} = 0$). Because the correlation reward shifts the logits non-uniformly according to feature network structure, it decreases Shannon entropy such that the effective search volume satisfies $\Omega_{\text{eff}}(\lambda_{\text{corr}}) = \Omega_{\text{eff}}(0) \cdot \exp\left(-D_{\text{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0)\right)$ (see proof in Appendix A.1). This exponential contraction significantly reduces the sample budget M required to evaluate elite candidate masks.

Each sampled mask $\mathbf{m}_j$ is applied to subsamples of the data in order to compute the silhouette objectives $\mathcal{L}(\mathbf{m}_j)$. The elite set contains the masks that achieve the highest values of $\mathcal{L}(\cdot)$. Formally,

$$\mathcal{E} = \{\mathbf{m}_j : \text{rank}(\mathcal{L}(\mathbf{m}_j)) \geq (1 - \gamma)M\}, \tag{3}$$

where γ is the thresholding percentage, and $\text{rank}(\cdot)$ is a ranking operator. Effectively, Eq. (3) isolates the best-performing subset masks to serve as the target distribution for our probability updates.

Once the elite set $\mathcal{E}$ has been constructed, we update the feature selection probabilities using a gradient-free CEM (De Boer et al. (2005)) that consists of two steps:

$$\pi_i^{(new)} = \frac{1}{|\mathcal{E}|} \sum_{\mathbf{m} \in \mathcal{E}} [\mathbf{m}]_i \qquad \pi_i^{(t+1)} = (1 - \alpha)\pi_i^{(t)} + \alpha \pi_i^{(new)} \tag{4}$$

As seen, Eq. (4) calculates the empirical frequency of each feature within the elite set and smoothly interpolates between it and the previous iteration probabilities using a learning rate α. The CEM updates the individual feature probabilities based on how well they partition the data across random subsamples. Feature sets that achieve high separability will be associated with larger entries of $\boldsymbol{\pi}^{(t+1)}$ and will therefore be sampled more often in subsequent iterations.

The iterative procedure for CLASP is detailed in Algorithm 1. The proof of asymptotic convergence to the global optimum for this optimization routine can be found in Appendix A.1 and considers the sequence of solutions obtained by monotonically changing the loss threshold of the elite set.

3.3 Implementation Details

Initialization and Stoppage: We initialize entries of $\boldsymbol{\pi}^{(0)}$ uniformly to z/p, where z is the number of desired top features, and p is the total number of features in the data matrix $\mathbf{X}$. The algorithm terminates when the top-z features in $\boldsymbol{\pi}^{(t)}$ are consistent across a fixed number of iterations.

Hyperparameters: All methods are evaluated using nested 5-fold cross-validation, with CLASP hyperparameters tuned via simulation experiments. We fix the user-defined target feature size $z = 100$, and select mask budget $M = 400$, subsample count $|S| = 50$, elite proportion $\gamma = 0.15$, smoothing rate $\alpha = 0.2$, and correlation weight $\lambda_{\text{corr}} = 0.3$. Empirically, we observe that CLASP is robust to the choices of hyperparameters over a wide range (Appendix A.3). We employ spectral clustering to compute Eq. (1), though CLASP is agnostic to the underlying clustering algorithm.

Computational Complexity: Computing the feature correlation matrix $\mathbf{R}$ is of complexity $\mathcal{O}(np^2)$. Each iteration of the CLASP optimization involves correlation-augmented Gumbel sampling, computing Eq. (1) across s data subsamples, and sorting the results. In the $p \gg n$ regime, these operations have complexity $\mathcal{O}(p^2 + Mp \log p)$, as guided by the sampling procedure. We implement CLASP and the baselines in Python 3.12.11 using NumPy 2.3.3 and Scikit-Learn 1.7.2. For computational efficiency, we parallelize the M independent feature mask evaluations across 28 CPU cores using Python Joblib 1.5.3. Correlation matrices $\mathbf{R}$ are precomputed per-fold to avoid redundant operations. Wall-clock runtimes for the seven biomedical datasets are in Appendix A.4.

3.4 Baseline Comparison Methods

We benchmark the performance of CLASP against a representative set of UFS methods spanning classical, spectral, and modern stochastic paradigms. These include (1) **Filter and Spectral Methods** consisting of the Laplacian Score (He et al. (2005)), Spectral Feature Selection (SPEC) (Zhao & Liu (2007)), Multi-Cluster Feature Selection (MCFS) (Cai et al. (2010)), and Independent Component Analysis (ICA) (Oja & Yuan (2006)); and (2) **Optimization and Deep Baselines** consisting of Concrete Autoencoders (CAE) (Balın et al. (2019)) and GroupFS (Lifshitz et al. (2026)).

Given our high-dimensional regime of $p \gg n$, we exclude baseline methods that require multi-run wrapper evaluations or dense matrix operations, as these scale poorly with the feature dimension p. Therefore, methods such as Nonnegative Discriminative Feature Selection (NDFS) (Li et al. (2012)) and Sample Feature Selection with Minowski-Weighted K-means (SFS-MWK++) (Zhang & de Amorim (2026)) were excluded. Together, the retained baselines provide a robust, representative comparative landscape spanning graph-based, spectral, and stochastic methods.

4 Experimental Results

We conduct a rigorous simulation study to verify the performance gain of CLASP over current UFS methods in the presence of correlated and high-variance noise. Our real-world experiments on seven publicly available datasets demonstrate improvements of CLASP in feature selection stability, out-of-sample cluster separability, and clustering stability. These datasets are diverse in modality and application, ranging from neuroimaging studies of brain disorders to genetics studies of cancer.

4.1 Simulation Study Under Varying Noise Conditions

Data Generation: Let $y_i \in \{1, \ldots, K\}$ denote the cluster assignment for data sample i. The input vector $\mathbf{x}_i \in \mathbb{R}^{p\times 1}$ is sampled according to $\mathbf{x}_i | y_i \sim \mathcal{N}(\boldsymbol{\mu}_{y_i}, \boldsymbol{\Sigma})$.

The covariance matrix $\boldsymbol{\Sigma}$ has the following block-diagonal structure:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{signal} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{dist} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma^2 \mathbf{I}_{p-2d} \end{bmatrix} \tag{5}$$

The true signal block $\boldsymbol{\Sigma}_{signal} \in \mathbb{R}^{d\times d}$ describes the d informative features, which have unit variance and pairwise correlation of ρ. The "distractor" noise block $\boldsymbol{\Sigma}_{dist} \in \mathbb{R}^{d\times d}$ describes the next d features. These features have unit variance pairwise correlation of γ but are *uncorrelated with the cluster label*. Finally, the remaining features represent independent noise with variance σ^2. The mean vector $\boldsymbol{\mu}_{y_i}$ enforces a cluster-specific mean shift of $\pm\frac{\delta}{2}$ strictly within the d signal features and enforces a mean of zero for the remaining distractor and noise features. By sweeping ρ, γ, and σ, we can evaluate the algorithmic capacity to isolate the true feature sets amid structured noise.

Performance Comparison: We evaluate CLASP and the baselines on the ground-truth feature recovery (FR) averaged over 5 independent trials per setting across 64 configurations of the simulated data. We analyzed performance across a grid of data regimes by varying the internal correlations of the true signal (ρ) and the distractor noise (γ), while fixing independent noise variance at $\sigma^2 = 2.0$. In all cases, we fix the number of features $p = 2000$, the number of data samples to $n = 200$, and the number of ground-truth clusters to $K = 2$. Fig. 2 illustrates the performance gain in FR over the best-performing baseline (ΔFR) alongside detailed results for all methods for a fixed $\rho = 0.5$.

As seen, CLASP demonstrates near-perfect feature recovery across the entire parameter grid. In regimes where the true signal dominates the noise ($\rho > \gamma$), the baseline methods achieve comparable accuracy, resulting in a performance delta near zero. However, when the data is corrupted by severe structural noise ($\gamma \geq \rho$), the baseline methods collapse. In the most adversarial regimes tested, CLASP retains a perfect recovery rate, which translates to up to 95% performance improvement over the next-best baseline. These results underscore the advantages of out separability-driven subsampling routine in CLASP, which can successfully isolate the true separable feature set.

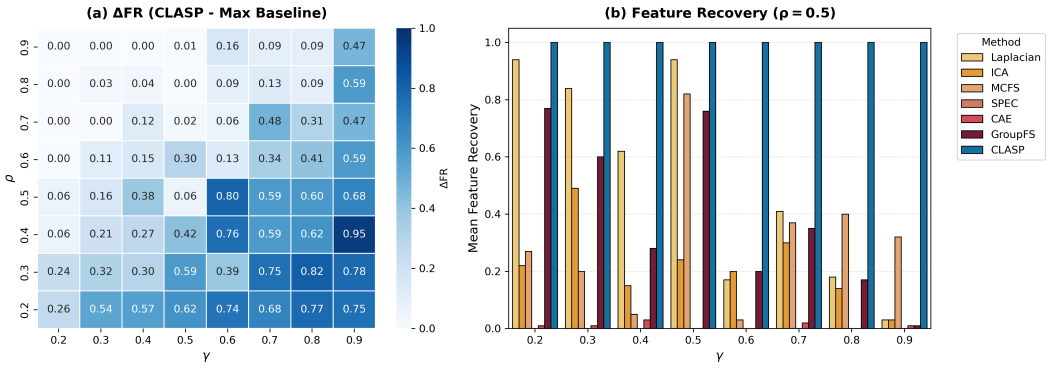


Figure 2: Feature recovery (FR) in simulated data sweeping inter-feature correlations ρ (true signal) and γ (noise) at constant $\sigma^2 = 2.0$. **(a):** Difference in FR percentage between CLASP and the top baseline. **(b):** FR of CLASP and baseline methods across γ for a fixed $\rho = 0.5$.

Table 1: Summary of the seven biomedical datasets.

Dataset	**Application**	**Data Type**	**Samples** (n)	**Features** (p)
ADHD-200 NYU	ADHD Classification	fMRI Connectivity	174	6670
ALLAML	Leukemia Classification	Genetics Microarray	72	7129
CPTAC-BRCA	Breast Cancer	Mass Spectrometry	125	12882
TADPOLE D1	Alzheimer's Progression	Neuroimaging Features	255	656
NCI-60 GI50	Drug Sensitivity	Molecular Drug Response	60	15154
ME CFS	Chronic Fatigue	Metabolomics	52	1790
GLI 85	Glioma Classification	Genetics Microarray	85	22283

4.2 Real-World Analysis of Biomedical Data

Datasets and Evaluation Setup: We perform UFS on seven biomedical datasets spanning distinct data modalities: ADHD-200, NYU site (consortium (2012)), ALLAML (Golub et al. (1999)), CPTAC-BRCA (National Cancer Institute Clinical Proteomic Tumor Analysis Consortium et al. (2020)), TADPOLE D1 (Marinescu et al. (2018)), NCI-60 GI50 (Reinhold et al. (2012)), ME CFS (Germain et al. (2020)), and GLI 85 (Freije et al. (2004)). Table 1 summarizes the dataset characteristics. In all cases, the number of features is much larger than the numbers of samples.

We use 5-fold cross validation to create dataset perturbations and use CLASP and the baselines to learn a feature subset on each training fold. We sweep the target number of clusters ($K = 2, \ldots, 5$) on each dataset to show the performance variation. We evaluate the methods across three distinct objectives: the consistency of the features selected, the separability of the resulting clusters, and the robustness of those solutions to variations in the input dataset.

Feature Selection Stability: We use Nogueira's stability index (Nogueira et al. (2018)) to quantify the robustness of each method across the 5 training folds. This metric quantifies the degree of overlap among the selected feature sets for a given method, while simultaneously correcting for chance agreement based on the data dimensionality and the selected feature set size.

Fig. 3 compares the Nogueira's stability index across methods. As seen, CLASP consistently outperforms all baseline methods regardless of the target number of clusters K. On average, CLASP achieves a feature stability of 0.798, reflecting a 34% improvement over the next best-performing baseline. In this way, CLASP takes a crucial step towards addressing the reproducibility crisis in biomedical data science. The stable features it identifies may be better aligned with underlying biological mechanisms that transient features that depend on a particular data sampling.

Clustering Quality: We quantify how well the selected feature sets partition the data by using them to cluster the held-out validation data of each training fold and computing the Davies-Bouldin

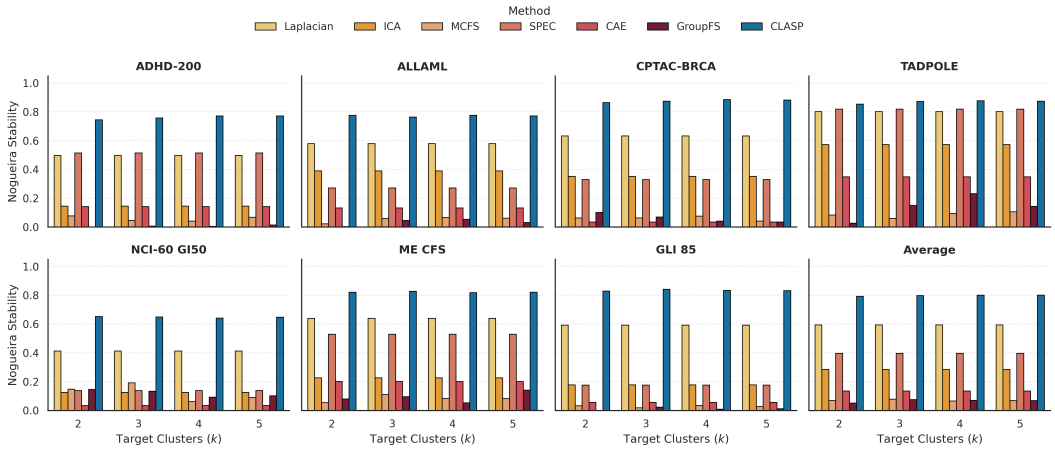


Figure 3: Nogueira stability index for all methods as a function of the number of target clusters K. The results consistently indicate that CLASP achieves the highest feature selection stability compared to all baseline methods across all evaluated conditions.

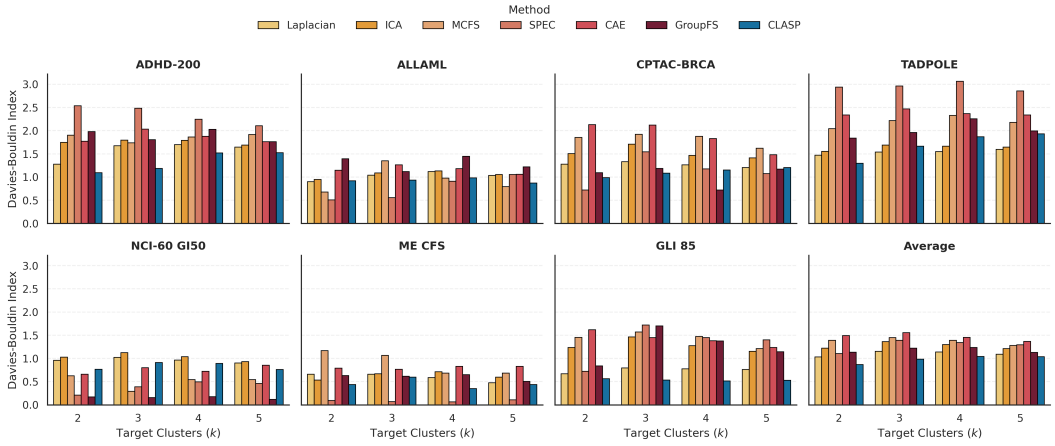


Figure 4: Davies-Bouldin Index (DBI) scores evaluating the clustering quality for all methods as a function of the number of target clusters K. Lower scores indicate better cluster separability, suggesting that CLASP generally learns either the best or comparable feature sets.

Index (DBI) on the clustering result (Davies & Bouldin (1979)). DBI evaluates cluster separability as the ratio of internal cluster dispersion to the distances between cluster centroids. A lower DBI indicates better separability, denoting clusters that are simultaneously dense and far apart.

Fig. 4 reports DBI across methods and target number of clusters for each dataset. CLASP consistently yields highly competitive DBI scores and outperforms the baselines in the majority of evaluated regimes. On average, CLASP achieves a DBI of 0.985, reducing the clustering error index by 11% compared to the strongest baseline. Together with the stability results, this indicates that CLASP successfully isolates reproducible feature subspaces that produce genuinely cohesive partitions, rather than rigidly selecting stable but non-informative artifacts.

Clustering Stability: Lastly, we evaluate the algorithmic robustness of the clustering outputs using the Adjusted Rand Index (ARI). We compute the ARI by comparing the clustering solutions across random subsets of 80% of the validation data. This analysis mimics the real-world sampling

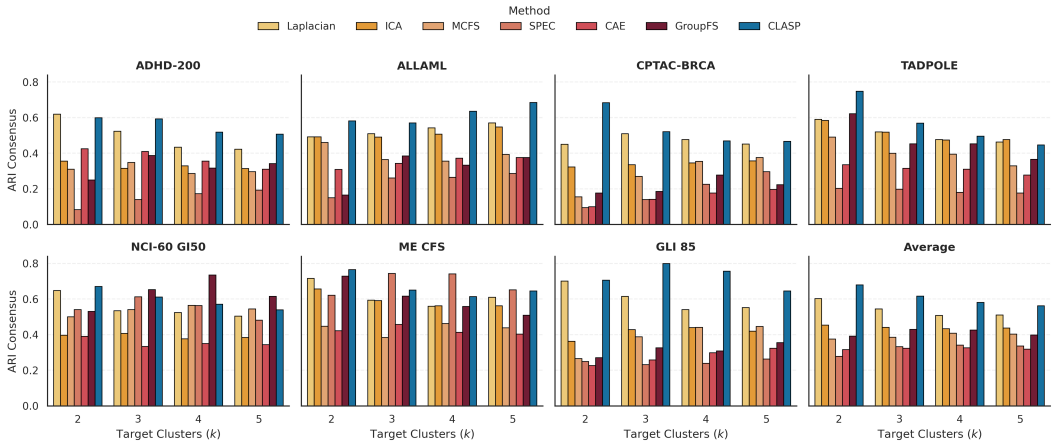


Figure 5: Adjusted Rand Index (ARI) consensus scores measuring the stability of clustering solutions under data perturbations across target cluster sizes K. CLASP consistently achieves either comparable or the most robust clustering partitions, as compared to the baseline methods.

variations that might be found in a biomedical dataset and quantifies how consistently the selected features partition unseen data across perturbations of the new cohort.

Fig. 5 shows the ARI across methods and datasets for $K = 2, \ldots, 5$. As seen, CLASP consistently produces the most resilient clustering configurations when subjected to data perturbation. On average across the evaluated datasets, CLASP achieves an ARI consensus score of 0.610, marking a 13% improvement over the most competitive baseline method. By maintaining high clustering stability alongside exceptional feature stability and cluster separability, CLASP ensures that the discovered clinical phenotypes and biomarkers are robust. Taken together, the results in this section underscore the practical utility of CLASP for reliable, real-world biomarker discovery.

5 Conclusion

High-dimensional biomedical environments ($p \gg n$) frequently yield unstable feature representations, which is a major bottleneck to clinical translation. To address this challenge, we presented Correlation-Leveraged Across-sample Separability Prioritization (CLASP), an unsupervised feature selection framework engineered to isolate reproducible dataset-level partitions by maximizing cluster separability across random subsamples of the data. CLASP leverages a CEM stochastic optimization with a correlation-augmented Gumbel sampling routine. Synthetic evaluations show the robustness of CLASP under heavy, correlated noise; benchmark evaluations across seven biomedical datasets established a 34% improvement in feature selection stability (Nogueira index), a 11% decrease in DBI, and a 13% increase in clustering stability (ARI across data subsamples).

By shifting the feature selection paradigm from static penalization to cross-subset consensus, CLASP provides a tool for mitigating the ongoing reproducibility crisis in biological data science. Rather than overfitting to cohort-specific acquisitions and nuisance variance, the algorithm isolates cohesive subspaces, offering researchers a foundation for downstream biomarker discovery. Future work includes integrating biological domain knowledge to prioritize clinically actionable feature sets, developing a continuous relaxation formulation that can be integrated into gradient-based optimization routines, and extending the framework to multi-modal and longitudinal regimes.

Reproducibility Statement

To ensure the full reproducibility of our work, all source code, experimental scripts, and preprocessed datasets are available in an anonymous repository at https://anonymous.4open.science/r/CLASP-E087/README.md. Algorithmic implementation choices are detailed in Section 3.3, with sensitivity analyses provided in Appendix A.3. Synthetic

data generation protocols and biomedical dataset characteristics are specified in Sections 4.1 and 4.2, respectively. Furthermore, the mathematical proofs for our theoretical claims regarding asymptotic convergence and search space entropy reduction are provided in Appendix A.1 and A.2.

Ethics Statement

This work exclusively uses pre-existing, anonymized, open-source biomedical datasets. We adhered to the licensing and privacy standards set by their original data providers and by our institution. This work does not involve any interaction with human subjects and poses no physical or privacy risks.

AI Disclosure

In this work, we used generative AI tools to provide feedback on research experiments and for general code completion. We have NOT used generative AI tools to create synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to make slight modifications to Fig. 1, create and edit software code, draft parts of the research paper, and brainstorming. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

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

Joeran Sander Bosma, Dré Peeters, Natália Alves, Anindo Saha, Zaigham Saghir, Colin Jacobs, and Henkjan Huisman. Reproducibility of training deep learning models for medical image analysis. In *Medical Imaging with Deep Learning*, pp. 1269–1287. PMLR, 2024.

Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 333–342, 2010.

Zhiwen Cao, Xijiong Xie, Feixiang Sun, and Jiabei Qian. Consensus cluster structure guided multi-view unsupervised feature selection. *Knowledge-Based Systems*, 271:110578, 2023.

Kelly D Cobey, Sanam Ebrahimzadeh, Matthew J Page, Robert T Thibault, Phi-Yen Nguyen, Farah Abu-Dalfa, and David Moher. Biomedical researchers' perspectives on the reproducibility of research. *PLoS biology*, 22(11):e3002870, 2024.

ADHD-200 consortium. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6:62, 2012.

Ian Covert, Uygar Sumbul, and Su-In Lee. Deep unsupervised feature selection. 2020.

Thiago F Covões and Eduardo R Hruschka. Towards improving cluster-based feature selection with a simplified silhouette filter. *Information Sciences*, 181(18):3766–3782, 2011.

David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7:1141, 2020.

Rajesh Dwivedi, Aruna Tiwari, Neha Bharill, Milind Ratnaparkhe, and Alok Kumar Tiwari. A taxonomy of unsupervised feature selection methods including their pros, cons, and challenges: R. dwivedi et al. *The Journal of Supercomputing*, 80(16):24212–24240, 2024.

William A Freije, F Edmundo Castro-Vargas, Zixing Fang, Steve Horvath, Timothy Cloughesy, Linda M Liau, Paul S Mischel, and Stanley F Nelson. Gene expression profiling of gliomas strongly predicts survival. *Cancer research*, 64(18):6503–6510, 2004.

Arnaud Germain, Dinesh K Barupal, Susan M Levine, and Maureen R Hanson. Comprehensive circulatory metabolomics in me/cfs reveals disrupted metabolism of acyl lipids and steroids. *Metabolites*, 10(1):34, 2020.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35: 507–520, 2022.

Md Mehedi Hassan, Swarnali Mollick, and Farhana Yasmin. An unsupervised cluster-based feature grouping model for early diabetes detection. *Healthcare Analytics*, 2:100112, 2022.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. *Advances in neural information processing systems*, 18, 2005.

Xiaofei He, Ming Ji, Chiyuan Zhang, and Hujun Bao. A variance minimization criterion to feature selection using laplacian regularization. *IEEE transactions on pattern analysis and machine intelligence*, 33(10):2013–2025, 2011.

Katsuya Hotta, Chunzhi Gu, and Chao Zhang. Consensus-aware sparse subspace clustering via coreset selection. *Pattern Recognition*, pp. 113097, 2026.

Pei Huang and Xiaowei Yang. Unsupervised feature selection via adaptive graph and dependency score. *Pattern Recognition*, 127:108622, 2022.

Iris AM Huijben, Wouter Kool, Max B Paulus, and Ruud JG Van Sloun. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1353–1371, 2022.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

Phi Le, Xingyue Gong, Leah Ung, Hai Yang, Bridget P Keenan, Li Zhang, and Tao He. A robust ensemble feature selection approach to prioritize genes associated with survival outcome in high-dimensional gene expression data. *Frontiers in systems biology*, 4:1355595, 2024.

Guojie Li, Zhiwen Yu, Kaixiang Yang, Mianfen Lin, and C. L. Philip Chen. Exploring feature selection with limited labels: A comprehensive survey of semi-supervised and unsupervised approaches. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6124–6144, 2024. doi: 10.1109/TKDE.2024.3397878.

Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, pp. 1026–1032, 2012.

Shira Lifshitz, Ofir Lindenbaum, Gal Mishne, Ron Meir, and Hadas Benisty. Unsupervised feature selection through group discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23496–23504, 2026.

Arun Vignesh Malarkkan, Xinyuan Wang, Kunpeng Liu, Denghui Zhang, and Yanjie Fu. Causally-guided diffusion for stable feature selection. *arXiv preprint arXiv:2603.20930*, 2026.

Razvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Stefan Klein, Daniel C Alexander, et al. Tadpole challenge: prediction of longitudinal evolution in alzheimer's disease. *arXiv preprint arXiv:1805.03909*, 2018.

Stefano Marrone, Stefano Olivieri, Gabriele Piantadosi, and Carlo Sansone. Reproducibility of deep cnn for biomedical image processing across frameworks and architectures. In *2019 27th european signal processing conference (eusipco)*, pp. 1–5. IEEE, 2019.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Stefan Michiels, Serge Koscielny, and Catherine Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365(9458):488–492, 2005.

Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE access*, 6:39501–39514, 2018.

Boaz Nadler. On consistency and sparsity for principal components analysis in high dimensions discussion. *Journal of the American Statistical Association*, 104(486):694–697, 2009.

W National Cancer Institute Clinical Proteomic Tumor Analysis Consortium et al. The clinical proteomic tumor analysis consortium breast invasive carcinoma collection (cptac-brca). *The Cancer Imaging Archive*, 2020.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018. URL http://jmlr.org/papers/v18/17-514.html.

Erkki Oja and Zhijian Yuan. The fastica algorithm revisited: Convergence analysis. *IEEE transactions on Neural Networks*, 17(6):1370–1381, 2006.

Wei Perng and Stella Aslibekyan. Find the needle in the haystack, then find it again: Replication and validation in the 'omics era. *Metabolites*, 10(7), 2020. ISSN 2218-1989. doi: 10.3390/metabo10070286. URL https://www.mdpi.com/2218-1989/10/7/286.

Jörg Rahnenführer, Riccardo De Bin, Axel Benner, Federico Ambrogi, Lara Lusa, Anne-Laure Boulesteix, Eugenia Migliavacca, Harald Binder, Stefan Michiels, Willi Sauerbrei, et al. Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. *BMC medicine*, 21(1):182, 2023.

William C Reinhold, Margot Sunshine, Hongfang Liu, Sudhir Varma, Kurt W Kohn, Joel Morris, James Doroshow, and Yves Pommier. Cellminer: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set. *Cancer research*, 72 (14):3499–3511, 2012.

Uri Shaham, Ofir Lindenbaum, Jonathan Svirsky, and Yuval Kluger. Deep unsupervised feature selection by discarding nuisance and correlated features. *Neural Networks*, 152:34–43, 2022.

Chang Tang, Xiao Zheng, Wei Zhang, Xinwang Liu, Xinzhong Zhu, and En Zhu. Unsupervised feature selection via multiple graph fusion and feature weight learning. *Science China Information Sciences*, 66(5):152101, 2023.

Jian-Bao Wang, Li-Jun Zheng, Qing-Jiu Cao, Yu-Feng Wang, Li Sun, Yu-Feng Zang, and Hang Zhang. Inconsistency in abnormal brain activity across cohorts of adhd-200 in children with attention deficit hyperactivity disorder. *Frontiers in Neuroscience*, 11:320, 2017.

Jianjun Wang et al. Biostatistical challenges in high-dimensional data analysis: Strategies and innovations. *Computational Molecular Biology*, 14, 2024.

Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

Xianchao Xiu, Chenyi Huang, Pan Shang, and Wanquan Liu. Bi-sparse unsupervised feature selection. *IEEE Transactions on Image Processing*, 2025.

Sijie Yang, Fei Zhu, Xinghong Ling, Quan Liu, and Peiyao Zhao. Intelligent health care: Applications of deep learning in computational medicine. *Frontiers in Genetics*, Volume 12 - 2021, 2021. ISSN 1664-8021. doi: 10.3389/fgene.2021.607471. URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2021.607471.

Zhong Yuan, Hongmei Chen, Pengfei Zhang, Jihong Wan, and Tianrui Li. A novel unsupervised approach to heterogeneous feature selection based on fuzzy mutual information. *IEEE Transactions on fuzzy systems*, 30(9):3395–3409, 2021.

Xudong Zhang and Renato Cordeiro de Amorim. Scalable unsupervised feature selection via weight stability. *Information Sciences*, pp. 123807, 2026.

Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 1151–1157, 2007.

Xiaoyan Zhu, Yu Wang, Yingbin Li, Yonghui Tan, Guangtao Wang, and Qinbao Song. A new unsupervised feature selection algorithm using similarity-based feature clustering. *Computational Intelligence*, 35(1):2–22, 2019.

Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M. Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2018.09.012. URL https://www.sciencedirect.com/science/article/pii/S1566253518304482.

A Appendix

A.1 Search Space Entropy Reduction via Correlation Augmentation

While standard CEM guarantees asymptotic convergence ($t \to \infty$), evaluating the full combinatorial space $\binom{p}{z}$ in finite time is computationally challenging when $p \gg n$. Here, we formally prove that introducing the empirical correlation matrix $\mathbf{R}$ via λ_{corr} strictly prunes the search space by reducing the Shannon entropy of the initial sampling distribution.

Let $\mathcal{P}(\mathbf{m} \mid \lambda_{\text{corr}})$ denote the probability distribution over feature masks $\mathbf{m} \in \mathcal{M}$ generated at initialization ($t = 0$), where inclusion probabilities are initialized uniformly as $\pi_i^{(0)} = \frac{z}{p}$. To formalize the search space entropy reduction, we state the following structural assumption:

1. **Non-Uniform Network Degree:** The underlying empirical correlation matrix $\mathbf{R} \in \mathbb{R}^{p \times p}$ contains non-uniform row sums. That is, there exist at least two features i, j such that $c_i \neq c_j$, where $c_i = \sum_{k=1}^{p} \mathbf{R}_{ik}$ represents the overall correlation centrality of feature i.

Theorem 1 (Search Space Entropy Reduction)**.** *Under Assumption **1**, for any scalar penalty weight $\lambda_{\text{corr}} > 0$, the Shannon entropy H of the initial sampling distribution is strictly less than the unaugmented baseline ($\lambda_{\text{corr}} = 0$):*

$$H(\mathcal{P}(\cdot \mid \lambda_{\text{corr}})) < H(\mathcal{P}(\cdot \mid 0))$$

Proof. At $t = 0$, the unaugmented baseline ($\lambda_{\text{corr}} = 0$) evaluates Equation 2 with uniform logits $\log(\boldsymbol{\pi}^{(0)}) = \log(z/p)\mathbf{1}$. Sampling the top-$z$ values using independent and identically distributed Gumbel noise $\mathbf{g}$ yields an equal probability for every valid mask $\mathbf{m} \in \mathcal{M}$:

$$\mathcal{P}(\mathbf{m} \mid 0) = \frac{1}{\binom{p}{z}} \quad \forall \mathbf{m} \in \mathcal{M}$$

The baseline distribution $\mathcal{P}(\cdot \mid 0)$ is uniform over $\mathcal{M}$, achieving maximum Shannon entropy:

$$H_0 = H(\mathcal{P}(\cdot \mid 0)) = \ln \binom{p}{z}$$

When $\lambda_{\text{corr}} > 0$, the sampling logits shift by $\lambda_{\text{corr}}(\mathbf{R}\boldsymbol{\pi}^{(0)})$. Substituting $\pi_k^{(0)} = \frac{z}{p}$, the logit for feature i becomes:

$$y_i = \ln\left(\frac{z}{p}\right) + \lambda_{\text{corr}} \frac{z}{p} c_i + g_i$$

Under the Gumbel-Max formulation, the marginal sampling probability $p_i(\lambda_{\text{corr}})$ for feature i is strictly monotonically increasing with respect to its logit location shift c_i.

By Assumption **1**, the row sum c_i is non-constant across features ($c_i \neq c_j$). Consequently, the marginal selection probabilities diverge from uniformity ($p_i \neq p_j$), making $\mathcal{P}(\cdot \mid \lambda_{\text{corr}})$ a non-uniform probability distribution over $\mathcal{M}$.

We quantify the divergence from uniformity using the Kullback-Leibler (KL) divergence between $\mathcal{P}_\lambda = \mathcal{P}(\cdot \mid \lambda_{\text{corr}})$ and the uniform distribution $\mathcal{P}_0 = \mathcal{P}(\cdot \mid 0)$:

$$D_{\text{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0) = \sum_{\mathbf{m} \in \mathcal{M}} \mathcal{P}_\lambda(\mathbf{m}) \ln\left(\frac{\mathcal{P}_\lambda(\mathbf{m})}{\mathcal{P}_0(\mathbf{m})}\right)$$

Expanding the logarithm with $\mathcal{P}_0(\mathbf{m}) = 1/\binom{p}{z}$:

$$D_{\text{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0) = \sum_{\mathbf{m} \in \mathcal{M}} \mathcal{P}_\lambda(\mathbf{m}) \ln \mathcal{P}_\lambda(\mathbf{m}) + \ln \binom{p}{z} \sum_{\mathbf{m} \in \mathcal{M}} \mathcal{P}_\lambda(\mathbf{m})$$

$$D_{\text{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0) = -H(\mathcal{P}_\lambda) + H_0$$

Rearranging gives:

$$H(\mathcal{P}_\lambda) = H_0 - D_{\text{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0)$$

By Gibbs' inequality, $D_{\mathrm{KL}}(P \parallel Q) \geq 0$ with equality if and only if $P = Q$. Because $\mathcal{P}_\lambda \neq \mathcal{P}_0$ under Assumption **1**, $D_{\mathrm{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0) > 0$, proving:

$$H(\mathcal{P}(\cdot \mid \lambda_{\mathrm{corr}})) < H_0$$

Finally, we quantify the physical contraction of the search space. In information theory, the effective size of a state space governed by a distribution $\mathcal{P}$ is characterized by its effective support size defined as $\Omega_{\mathrm{eff}} \equiv \exp(H(\mathcal{P}))$. For the uniform baseline ($\lambda_{\mathrm{corr}} = 0$), this recovers the full volume:

$$\Omega_{\mathrm{eff}}(0) = \exp(H_0) = \exp\left(\ln \binom{p}{z}\right) = \binom{p}{z}$$

For $\lambda_{\mathrm{corr}} > 0$, taking the exponential of both sides of the entropy relation $H(\mathcal{P}_\lambda) = H_0 - D_{\mathrm{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0)$ yields the following exponentiation:

$$\exp\left(H(\mathcal{P}_\lambda)\right) = \exp\left(H_0 - D_{\mathrm{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0)\right)$$

Applying exponent rules, we separate the joint term into a multiplicative contraction factor:

$$\Omega_{\mathrm{eff}}(\lambda_{\mathrm{corr}}) = \Omega_{\mathrm{eff}}(0) \cdot \exp\left(-D_{\mathrm{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0)\right)$$

Because $D_{\mathrm{KL}}(\mathcal{P}_\lambda \parallel \mathcal{P}_0) > 0$ under Assumption **1**, its negative exponentiation is less than 1. This proves that correlation-augmented sampling yields an exponential reduction in the effective volume of the search space relative to the unaugmented combinatorial baseline. Consequently, this exponential volume contraction guarantees that a sample budget $M \ll \binom{p}{z}$ can produce reliable empirical quantile estimates for constructing the elite set $\mathcal{E}$. □

A.2 Proof of Asymptotic Convergence for CEM Feature Selection

Having established in Appendix A.1 that correlation-augmented sampling exponentially contracts the effective search volume per iteration, we now prove that the iterative parameter updates $\boldsymbol{\pi}^{(t)}$ in CLASP achieve asymptotic convergence to the globally optimal feature mask $\mathbf{m}^*$.

Let $\mathcal{M} = \{\mathbf{m} \in \{0,1\}^p \mid \sum_{i=1}^p m_i = z\}$ denote the finite state space of all valid binary feature masks. The objective is to maximize the Consensus Separability Loss $\mathcal{L}(\mathbf{m})$ over $\mathcal{M}$. To formalize asymptotic convergence, we state the following structural assumptions:

1. **Finite Search Space:** The domain $\mathcal{M}$ is discrete and finite, with cardinal volume $|\mathcal{M}| = \binom{p}{z} < \infty$.
2. **Unique Global Optimum:** There exists a unique optimal mask $\mathbf{m}^* \in \mathcal{M}$ such that $\mathcal{L}(\mathbf{m}^*) > \mathcal{L}(\mathbf{m})$ for all $\mathbf{m} \in \mathcal{M} \setminus \{\mathbf{m}^*\}$.
3. **Strict Positivity of Sampling Support:** For all iterations t, the probability of generating any arbitrary mask $\mathbf{m} \in \mathcal{M}$ under the parametrized sampling distribution is strictly positive, $P(\mathbf{m}; \boldsymbol{\pi}^{(t)}) > 0$. This is satisfied via the unbounded support of the additive Gumbel noise $\mathbf{g} \sim \mathrm{Gumbel}(0, 1)$.
4. **Asymptotic Sample Limits:** The number of sampled feature masks per iteration $M \to \infty$, enabling exact calculation of the empirical γ-quantile loss threshold.
5. **Unsmoothed Asymptotic Update:** The learning rate satisfies $\alpha \to 1$ as $t \to \infty$.

Theorem 2 (Asymptotic Convergence)**.** *Under Assumptions **1–5**, as $t \to \infty$, the parameter vector $\boldsymbol{\pi}^{(t)}$ converges almost surely to the indicator vector of the global optimum $\mathbf{m}^*$:*

$$\lim_{t \to \infty} \boldsymbol{\pi}^{(t)} = \mathbf{m}^* \quad \textit{a.s.}$$

Proof. Let L_t denote the elite loss threshold at iteration t, defined as the loss achieved at the $(1-\gamma)$-quantile of sampled masks. Under Assumption **4** ($M \to \infty$), the stochastic sequence $\{L_t\}_{t=0}^{\infty}$ generated by selecting the top γ-proportion of samples forms a deterministic non-decreasing sequence:

$$L_0 \leq L_1 \leq L_2 \leq \cdots \leq L_t \leq L^*$$

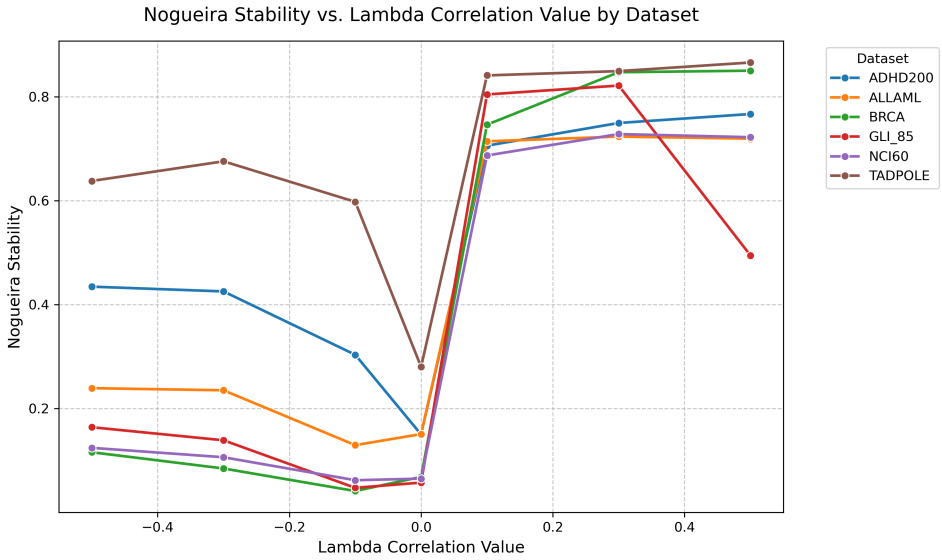


Figure 6: Nogueira stability of selected features across folds for different values of λ_{corr}. The performance is significantly improved for positive values of the hyperparameteres. The results are robust to values of λ_{corr} over a wide range satisfying $\lambda_{corr} > 0.1$.

where $L^* = \mathcal{L}(\mathbf{m}^*)$.

Because $\mathcal{M}$ is finite (**1**), the loss function is bounded above by L^*. By the Monotone Convergence Theorem, the sequence $\{L_t\}$ converges to a limit $L_\infty \leq L^*$. Because the Gumbel distribution provides full support over $\mathbb{R}^p$ (**3**), every mask in $\mathcal{M}$ has a non-zero probability of selection at every iteration. Consequently, the upper quantile threshold must eventually reach the global maximum:

$$\lim_{t\to\infty} L_t = L^* = \mathcal{L}(\mathbf{m}^*)$$

As $L_t \to L^*$, the elite set $\mathcal{E}^{(t)} = \{\mathbf{m_j} : \mathcal{L}(\mathbf{m_j} \geq L_t\}$ shrinks to include only masks that attain the global maximum. By Assumption **2** (uniqueness of $\mathbf{m}^*$), there exists a finite iteration T such that for all $t \geq T$:

$$\mathcal{E}^{(t)} = \{\mathbf{m}^*\}$$

Recalling the CEM parameter update equation with $\alpha = 1$ (**5**):

$$\pi_i^{(t+1)} = \frac{1}{|\mathcal{E}^{(t)}|} \sum_{\mathbf{m}\in\mathcal{E}^{(t)}} m_i$$

For $t \geq T$, substituting $\mathcal{E}^{(t)} = \{\mathbf{m}^*\}$ directly yields:

$$\pi_i^{(t+1)} = m_i^* \quad \forall i \in \{1, \ldots, p\}$$

Thus, $\lim_{t\to\infty} \boldsymbol{\pi}^{(t)} = \mathbf{m}^*$, completing the proof. □

A.3 Hyperparameter Sensitivity for CLASP

A.3.1 Empirical Ablation of λ_{corr} in Biomedical Datasets

To corroborate that the reduced search space provided by the correlation reward improves the stability of our results, we assess the Nogueira stability of selected feature sets at different values of λ_{corr}. The results can be seen in Fig. 6. The best average performance occurs with a positive $\lambda_{corr} = 0.3$, which yields an average Nogueira stability of 0.786. We also note that the Nogueria's index is stable across a range of λ_{corr} values satifying $\lambda_{corr} > 0.1$.

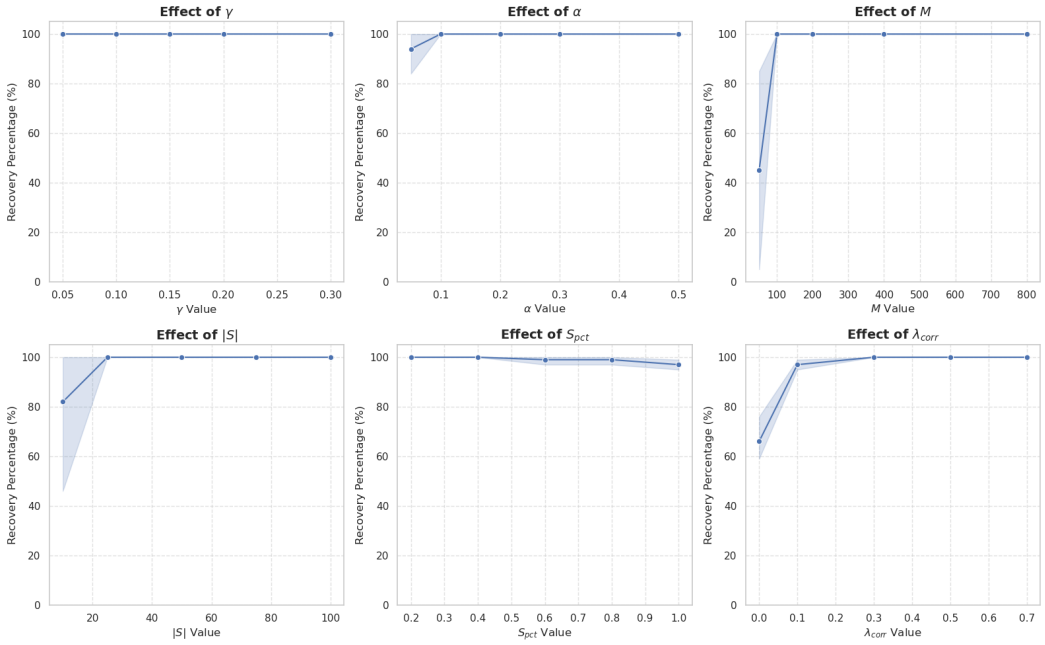


Figure 7: Hyperparameter sensitivity analysis of CLASP on synthetic data ($n = 200, p = 2000, d = 20, \delta = 2.0$). Each subplot displays the mean feature recovery percentage and shaded confidence intervals, while sweeping one parameter and holding the others at their defaults ($\gamma = 0.15, \alpha = 0.2, N = 400, |S| = 50$, selected participant percentage $S_{pct} = 0.2, \lambda_{corr} = 0.3$). CLASP demonstrates broad robustness across the hyperparameter space, provided minimal thresholds for sample budget ($M \geq 100$) and correlation weight ($\lambda_{corr} \geq 0.1$) are met.

A.3.2 Performance of CLASP Across Different Hyperparameter Values

We conduct a hyperparameter sweep using simulated data ($n = 200, p = 2000, d = 20, \delta = 2.0$, 5 repeats) to evaluate the robustness of our CLASP optimization. Fig. 7 displays the results for multiple hyperparameters. As seen, CLASP demonstrates high stability across the core CEM. Feature recovery remains perfect across all evaluated elite fractions ($\gamma \in [0.05, 0.3]$) and is robust to learning rates $\alpha \geq 0.1$. While performance degrades if mask sample budget M is excessively constrained ($M = 50$), CLASP consistently achieves 100% recovery for $M \geq 100$. This analysis confirms that correlation-augmented sampling significantly reduces the effective search space, requiring only a fractional mask budget to consistently locate the optimal feature set.

The hyperparameter sweeps further validate the necessity of our cross-subsample consensus loss and correlation prior. Feature recovery drops to roughly 65% when the correlation weight λ_{corr} is removed, but rapidly converges to 100% for $\lambda_{corr} \geq 0.3$, corroborating that the prior effectively guides the initial search. Furthermore, evaluating the algorithm across varying participant sampling fractions S_{pct} reveals that lower sampling ratios (20% to 40%) yield optimal performance, while perfect recovery is achieved with 25 participant subsamples and greater.

A.4 Wall-Clock Runtime of CLASP Across Biomedical Datasets

Table 2 looks at the runtime of CLASP across the seven presented biomedical datasets, assessing the feasibility of the method to find the selected features in a tractable amount of time. For computational efficiency, we parallelize the M independent feature mask evaluations across 28 CPU cores using Python Joblib 1.5.3. Correlation matrices $\mathbf{R}$ are precomputed to avoid redundant operations. We select $K = 2$ clusters and $z = 100$ target features. Runtimes remain practical, ranging from 6 minutes 16 seconds to 22 minutes 47 seconds depending on the dataset. For comparison, we also

Table 2: Statistics of CLASP and GroupFS runtime on the seven biomedical datasets.

Dataset	**Instances** n	**Features** p	**CLASP Runtime in mins:secs**	**GroupFS Runtime in mins:secs**
ADHD-200 NYU	174	6670	8m 58.83s	2m 6.55s
ALLAML	72	7129	7m 44.52	2m 3.25s
CPTAC-BRCA	125	12882	6m 8.30s	12m 7.27s
TADPOLE D1	255	656	8m 28.91s	0m 1.57s
NCI-60 GI50	60	15154	22m 47.20s	96m 3.17s
ME CFS	52	1790	6m 16.80s	0m 2.41s
GLI 85	85	22283	19m 42.75s	62m 16.01s

report the runtimes for GroupFS, a stochastic baseline, which demonstrates that CLASP achieves comparable execution times.