GEASS: NEURAL CAUSAL FEATURE SELECTION FOR HIGH-DIMENSIONAL BIOLOGICAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying nonlinear causal relationships in high-dimensional biological data is an important task. However, current neural network based causality detection approaches for such data suffer from poor interpretability and cannot scale well to high dimensional regime. Here we present GEASS (Granger fEAture Selection of Spatiotemporal data), which identifies sparse Granger causality mechanisms of high dimensional spatiotemporal data by a single neural network. GEASS maximizes sparsity-regularized multi-dimensional transfer entropy with a theoretical guarantee of recovering features with spatial/temporal Granger causal relationships. The sparsity regularization is achieved by a novel combinatorial stochastic gate layer to select sparse non-overlapping feature subsets. We demonstrate the efficacy of GEASS in several synthetic datasets and real biological data from single-cell RNA sequencing and spatial transcriptomics.

1 INTRODUCTION

Advances in single-cell omics research enable full characterizations of high-dimensional gene dynamics in biological systems on a either temporal or spatial scale. An example for the former is single-cell RNA sequencing (scRNA-seq) trajectories, where cells are sampled from a dynamical biological process, sequenced, and ordered based on either real sampled time or inferred pseudo-time (Cannoodt et al., 2016; Saelens et al., 2019). Gene dynamics along the specified cell order encodes information of causal regulation for the underlying biological process. An example for the latter is single-cell level spatial transcriptomics (e.g. SeqFISH+ (Eng et al., 2019), Merfish (Fang et al., 2022)), in which cells from a tissue slice are sequenced with their spatial coordinates preserved (Moses and Pachter, 2022; Rao et al., 2021; Palla et al., 2022). Spatial profiling allows investigations of the cellular interplay, corresponding to conditional gene expression change caused by neighborhood phenotypic states. However, despite the potential significance, data-driven causal discovery for such data remains largely unexplored, especially for the spatial omics data.

Identifications of causal regulatory patterns in such data can be reformulated into the general task of causal feature selection in observational data with intrinsic structures, e.g. spatial data or temporal data. Identifications of causal interactions in time-series has lead to valuable findings in multiple disciplines, including but not limited to, economy, climate science, and biology (Hoover, 2006; Kamiński et al., 2001; Runge et al., 2019a).

Learning directed causal relationships in temporal / spatial data is feasible as time and space both induce asymmetric dependencies. In the case of time-series data, a feature in the future cannot have effect on past values of other features. For data with spatial structure, a similar definition of causal dependency can be established based on the irreversibility of the spatial local diffusion operator.

The concept of Granger causality is proposed in order to uncover the assymetric causal dependency (Granger, 1969; Shojaie and Fox, 2022). In time-series data, this would translate to identifying one variable's causal relationship with other variables based on how well the historical observations of other variables can predict the variable's present value. The application of Granger causality in a spatial context corresponds to predicting significant relationships between neighboring observations of other variables and the specified variable (Mielke et al., 2020), which is a key insight used in recent works aimed to discover cellular interaction patterns in spatial omics data (Fischer et al., 2021; Valdés-Sosa et al., 18).

In the nonlinear regime, information-theoretic measures such as directed information, transfer entropy (Schreiber, 2000), and partial transfer entropy (Staniek and Lehnertz, 2008), are used as a counterpart of linear Granger causality. Moreover, some works consider modeling conditional independence in time-series data to identify the underlying causal graph (Entner and Hoyer, 2010; Malinsky and Spirtes, 2018; Moneta et al., 2011; Runge et al., 2019a; Pfister et al., 2019; Mastakouri et al., 2021). Two examples are VarLINGAM (Hyvärinen et al., 2010) and PCMCI (Runge et al., 2019b), which are generalizations of LINGAM (Shimizu et al., 2006) and PC (Spirtes et al., 2000) respectively. Finally, multiple recent works have proposed to use neural network approaches to model the nonlinear Granger causality, including MLP, LSTM, and neural-ODE based approaches, resulting in improved prediction power for nonlinear time-series dynamics (Li et al., 2017; Tank et al., 2021; Nauta et al., 2019; Yin and Barucca, 2022; Bellot et al., 2021).

Despite the success of these methods in various systems of interest, multiple challenges limit their use in previously discussed high-dimensional biological datasets.

- Although linear methods (LINGAM, linear Granger causality) have succeeded in various settings and can potentially scale to high feature numbers, these methods may completely fail when the feature dependency in data is highly complex and nonlinear.
- As the number of conditional independence generally scales exponentially with feature size, applying CI-test based methods in data with high feature dimensionality is not realistic. Meanwhile, Granger-causality based methods build prediction models with the number equal to feature size. Therefore the time complexity of solving the stacked model is of polynomial level with respect to the feature size.
- Edge sparsity is usually assumed by previous methods to maximize interpretability of the identified causal graph. However, In biological data, there exists a large proportion of nuisance features (e.g. constitutively expressed genes), therefore the feature sparsity and edge sparsity need to be both included by an ideal model. This is of particular interest as false discovery rate (FDR) control is favorable in biological data analysis.
- While a large number of methods are designed for causal discovery in time-series data, only a limited number of present works aim for causal discovery in general graph-structured data. Time-series based methods cannot be applied on data with multi-branch trajectory dynamics or spatial structures.

Our contribution. In this work, we present GEASS (Granger fEAture Selection of Spatiotemporal data), which identifies causally interacting features of high dimensional temporal / spatial data by a single neural network. GEASS considers sparsity of underlying causal mechanisms instead of link sparsity, thus selects most significant corresponding features for downstream causal discovery. Our contributions are three-folds.

1. Instead of direct causal discovery in data, we formulate the task as two steps of causal feature selection and causal graph identification. We provide a novel solution of causal feature selection problem in general graph-structured data by the use of multi-dimensional transfer entropy maximization with theoretical guarantees.
2. In order to solve our proposed optimization problem, we design a novel combinatorial stochastic gate layer to select non-overlapping sparse feature sets with a newly designed initialization procedure.
3. We demonstrate the power of our method by benchmarking it on both temporal data and spatial data of multiple settings. Our method gives accurate and robust causal feature identification and reveals novel biology in real datasets.

1.1 Related works

Neural Granger causality

Despite the large body of work based on linear Granger causal discovery, neural Granger causality still remains an active area of research. Various neural network architectures, such as MLP, sequential model, and attention-based architecture (Tank et al., 2021; Nauta et al., 2019; Khanna and Tan, 2019; Sun et al., 2021), have been proposed for nonlinear Granger causality discovery. A recent work uses

the information of proxy variable to learn latent confounder for Granger causality by a dual-decoder neural network (Yin and Barucca, 2022). One recent biology-oriented work extends the definition of Granger causality to DAGs, where the use of a linear graph neural network is proposed to model underlying Granger causality (Wu et al., 2021). Meanwhile, a neural-ODE based approach has been proposed to reformulate the Granger causality problem in terms of local dependence graph identification (Bellot et al., 2021).

Causal feature selection

The task of causal feature selection has been considered by multiple groups. Most works in this category uses constraint-based methods to identify each feature's causal relation with all other features, equivalent of identifying the whole causal graph structure, including VARLINGAM, tsFCI, SVAR-FCI, and PCMCI (Hyvärinen et al., 2010; Entner and Hoyer, 2010; Malinsky and Spirtes, 2018; Moneta et al., 2011; Runge et al., 2019a). Meanwhile, seqICP focus on identifying the direct or indirect cause for each feature assuming sufficient interventions in the dataset (Pfister et al., 2019). SyPI tackles the causal feature selection problem without the assumption of causal sufficiency and avoids issues in multihypothesis testing by construction of the correct conditional set (Mastakouri et al., 2021). Finally, Guo et al. (2022) considers dual correction of causal feature selection to control both false positive rates and false negative rates.

2 MULTI-DIMENSIONAL TRANSFER ENTROPY

In order to tackle the issue that a neural network may overfit each model therefore overestimates causal interaction numbers, we need a prediction-free loss function that directly indicates causal signficance. In this work, we propose a novel function, multi-dimensional transfer entropy, based on transfer entropy as a metric of causal interaction significance.

Transfer entropy is a information-theoretic measure of cross dependence. (Schreiber, 2000) Consider two time series X_t and Y_t for $t \in 1, ..., T$. Transfer entropy aims to model nonlinear dependency across X_t and Y_t while accounting for possible autocorrelations. It has been theoretically shown that transfer entropy value is equivalent to Granger causality statistical significance for Gaussian data (Barnett et al., 2009). The transfer entropy can be formulated as: $\mathrm{TE}_h(X, Y) = I(X_t; Y_{t+h}|Y_t)$. Denoting the time/space forward operator by W, the transfer entropy can be generalized as:

$$\mathrm{TE}_W(X, Y) = I(X; WY|Y).$$

In this work, we propose a novel function, multi-dimensional transfer entropy to extend the bivariate transfer entropy to higher-dimensional settings. The key insight is to consider two feature subsets $S_1, S_2 \subset \{1, ..., p\}$, such that the time-lagged mutual information over two multivariate random variables X_{S_1}, X_{S_2} is maximized.

Definition 2.1. Assume $X \in R^{n \times p}$ to be the data of interest with a forward operator W. $S_1, S_2 \subset \{1, ..., p\}$ each represents a set of feature subset indices. The multi-dimensional transfer entropy (mTE) is defined as:

$$\mathrm{mTE}^*_W(X) = \max_{S_1, S_2} \mathrm{mTE}_W(X_{S_1}, X_{S_2}) = I(X_{S_1}; WX_{S_2}) - I(X_{S_1}; X_{S_2}).$$

And the estimator $(\hat{S}_1, \hat{S}_2)$ can be defined as the element with smallest cardinal in the solution set $\mathrm{argmax}_{S_1, S_2} \mathrm{mTE}_W(X_{S_1}, X_{S_2})$.

Note the constructed mTE function requires strictly stronger dependence and includes no false positives, as shown by the proposition below (Proof can be seen at Appendix A.1):

Proposition 2.2. $\forall S_1, S_2 \subset \{1, ..., p\}, \mathrm{mTE}_W(X_{S_1}, X_{S_2}) > 0 \Rightarrow \mathrm{TE}_W(X_{S_1}, X_{S_2}) > 0.$

The purpose of mTE construction is to enable neural optimization as discussed in the next section. Under some mild assumptions, we are able to provide theoretical justification for mTE maximization for time-series setting. The complete proof can be seen in Appendix A.3.

Assumptions:

A1 **Causal Markov assumption** for the causal graph.

A2 **Faithfulness** for the causal graph.

A3 **Causal sufficiency** for the causal graph.

A4 **Ergodicity and Stationarity** of the stochastic process defined by the causal graph, meaning the ensemble average equals time average, and the functional relationships encoded by the causal graph do not change by time (or location).

A5 **DAG causal graph:** We assume $X = [s_1, ..., s_l, t_1, ..., t_m]$, where s_k are nuisance features that may correlate with t_i, and t_i are causally interacting signals and form a connected directed acyclic graph (DAG). An illustration corresponding to the assumption can be seen in Figure 1.

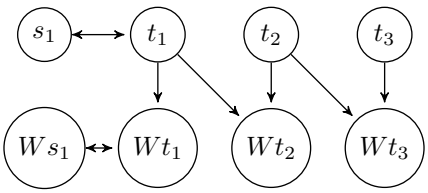


Figure 1: Causal graph for a single forward step by assumption 5.

A6 **Interaction regularity:** Given two interacting features $A, B \in X, i \in \{0, 1\}$ with $A \subset Pa(W^i B) - B$ and any other feature set X^c, the mutual information difference is larger than zero:

$$\forall j, k, l(l \neq i), I(W^j A, W^{j+i} B | X^c) > I(W^k A, W^{k+l} B | X^c).$$

Remark 2.3. Here our only additional assumption from prevalent literatures is A6, which aims to filter out features with spurious causations and regularize the algorithmic complexity of causal interactions, thus enabling information-theoretic analysis. A6 has direct connections with the concept of conditional transfer entropy, which aims to filter out indirect causal relationships via conditioning (Faes et al., 2016; Shahsavari Baboukani et al., 2020). Further discussions can be seen at Appendix A.2.

Theorem 2.4. *Given assumption A1-A6,* $X^*_{\hat{S}_1 \cup \hat{S}_2} \subset (t_1, ..., t_m)$. *Moreover, each feature in* $\hat{S}_1 \cup \hat{S}_2$ *is connected to other features in the set.*

3 Neural optimization of multi-dimensional transfer entropy

With theorem 3.1 stated below, we are able to give a theoretical gaurantee of the neural optimization of mTE. The proof can be seen at Appendix A.4.

Theorem 3.1. *Assume assumptions A1-A6 holds. Then for small enough* $\lambda > 0$*:*

$$\min_{f,g,h,S_1,S_2} -(I(f(X \odot \mathbb{1}_{S_1}); h(Wg(X \odot \mathbb{1}_{S_2}))) - I(f(X \odot \mathbb{1}_{S_1}); g(X \odot \mathbb{1}_{S_2}))) + \lambda |S_1 \cup S_2|$$

We have 1. the optimal h *is a one-to-one mapping; 2. Assume* f, g *define one-to-one mappings on* $X \odot \mathbb{1}_{S_1}, X \odot \mathbb{1}_{S_2}$ *respectively. Then,* $X^*_{S_1 \cup S_2} \subset (t_1, ..., t_m)$. *Moreover, each feature in* $S_1 \cup S_2$ *is connected to other features in the set.*

Remark 3.2. The estimation of mutual information by neural networks is an active field itself (Belghazi et al., 2018; Hjelm et al., 2018; McAllester and Stratos, 2020; Zhang et al., 2019). In contrast, by this theorem, we show that an accurate estimation of the transfer entropy (such as in (Zhang et al., 2019)) may not be needed as optimizing the **upper bound** of the modified transfer entropy automatically gives the best feature subset selection.

Remark 3.3. Compared with mTE, the original transfer entropy is less robust to selections of mappings f, g, h. Intuitively, we can construct non-one-to-one gs such that $I(f, Wg|g) = I(f, Wg, g) - I(f, g) > 0$ strictly for any features. This is further illustrated by the empirical performance comparison in the next section.

Given Theorem 3.1, we are able to construct a neural network for optimizing the proposed loss function. However, the estimation of mutual information is not directly tractable. In this case, because mutual information is invariant by one-to-one transforms, we can restrict the function class as:

$$\min_{f,g,h \sim Gaussian, S_1, S_2} -(I(f(X \odot \mathbb{1}_{S_1}); h(Wg(X \odot \mathbb{1}_{S_2}))) - I(f(X \odot \mathbb{1}_{S_1}); g(X \odot \mathbb{1}_{S_2}))) + \lambda |S_1 \cup S_2|$$

Here $f, g, h \sim Gaussian$ indicate flows transforming the original feature distributions into Gaussian distributions. We are able to formulate the target for neural network optimization by the explicit formula for mutual information between Gaussians: $I(X,Y) = \frac{1}{2}\log\frac{\det\Sigma_X \det\Sigma_Y}{\det\Sigma_{[X,Y]}}$. The Gaussian regularization can be applied either by regularizing over the discrepancy between embedding distributions and Gaussian distributions or by applying a adversarial training procedure. In this work, we use the former approach, constructing mean and variances for each embedding component as learnable parameters and minimize the cross entropy between target distributions and the parametrized Gaussian distributions.

3.1 COMBINATORIAL STOCHASTIC GATES

In order to solve the optimization problem, we need to learn two sparse sets S_1, S_2. Here we use a stochastic gate based approach (Yamada et al., 2020; Lindenbaum et al., 2021). In order to explicitly construct S_1 and S_2 by stochastic gates, we define two random vectors T^1 and T^2 ranging in $[0,1]$ with lengths equal to the feature number, with each element independently sampled from STG distribution defined as: $T_d^i = \max(0, \min(1, \mu_d^i + \epsilon_d^i))$, where $\epsilon_d^i \sim N(0, \sigma^2)$ with fixed variance and μ_d^i is a parameter trainable by the reparametrization trick (Miller et al., 2017; Figurnov et al., 2018).

In this case, denote $\mathbb{1}_A$ as the indicator function, the loss function can be reformulated as:

$$\mathbb{E}_X\mathbb{E}_{T^1,T^2} - (I(f(X_{S_1}); h(Wg(X_{S_2}))) - I(f(X_{S_1}); g(X_{S_2}))) + \lambda_1||T^1||_0 + \lambda_2||\mathbb{1}_{T^2\in[0,1]}||_0,$$
$$s.t. \quad X_{S_1} = X \odot T^1 \odot T^2, \quad X_{S_2} = X \odot T^1 \odot (1 - T^2). \quad (*)$$

In our construction, T^1 controls the sparsity of feature selection, while T^2 corresponds to overlapping level between S_1 and S_2. Defining the Gaussian error function as $\mathrm{erf}()$, the regularization term for the first layer is of form:

$$\sum_{d=1}^{p}\mathbb{P}(T_d^1 \geq 0) = \sum_{i=1}^{p}(\frac{1}{2} - \frac{1}{2}\mathrm{erf}(\frac{\mu_d^1}{\sqrt{2}\sigma})).$$

The regularization term for the second layer can be expressed as:

$$\sum_{d=1}^{p}\mathbb{P}(T_d^2 \in [0,1]) = \sum_{d=1}^{p}\mathbb{P}(T_d^2 \geq 0) - \mathbb{P}(T_d^2 \geq 1) = \frac{1}{2}\sum_{d=1}^{p}(\mathrm{erf}(\frac{\mu_d^2}{\sqrt{2}\sigma}) - \mathrm{erf}(\frac{\mu_d^2 - 1}{\sqrt{2}\sigma})).$$

We are able to show strong consistency for our stochastic-gate based feature selection as the theorem below (Proof can be seen at Appendix A.6):

Theorem 3.4. *Given assumptions A1-A6 and assumption on f, g described in theorem 3.1, denote for the optimal solution of (*), a sample of stochastic gate is given by T^1, T^2 and ground truth active gates as $T_{S_1^*}, T_{S_2^*}$, then we have for small enough $\lambda_1 > 0$ and $\lambda_2 > 0$,*

$$\mathbb{P}(\lim_{N\to\infty}\lim_{\epsilon\to 0}\mathbb{1}_{T^1>\epsilon, T^2<\epsilon} = T_{S_1^*}) = 1; \mathbb{P}(\lim_{N\to\infty}\lim_{\epsilon\to 0}\mathbb{1}_{T^1>\epsilon, T^2>\epsilon} = T_{S_2^*}) = 1.$$

In practice, we also have observed the method's solution highly depends on the stochastic gate initialization. Here we provide an heuristic initialization scheme that shows superior empirical performance. Details of the initialization scheme can be seen in Appendix B.

3.2 PROPOSED NETWORK ARCHITECTURE

Our proposed network architecture is summarized in Figure 2. For an input dataset $X \in \mathbb{R}^{n\times p}$ and its corresponding graph adjacency matrix $W \in \mathbb{R}^{n\times n}$, we first pass each feature through two sequential stochastic gate layers (STG layer 1: w, STG layer 2: z). L0 penalty is conducted on the first STG layer, while the second STG layer is regularized such that large μ_d in magnitude is favored (0-1 loss).

After passing each feature, denote $\hat{z}_i = 1 - z_i$, we have two intermediate embeddings defined by $x_i \odot w_i \odot z_i$ and $x_i \odot w_i \odot \hat{z}_i$ respectively. Then these two embeddings are passed through two MLPs (f, g in theorem 3.1) to generate two output embeddings A, B. Finally we apply the graph adjacency matrix W and another MLP layer (MLP3, h) on B to generate embedding B*. We estimate the mTE via the mutual information formula of joint Gaussian distributions and add the cross-entropy penalty between the concatenated (A,B,B*) and a Gaussian distribution with learnable parameters.

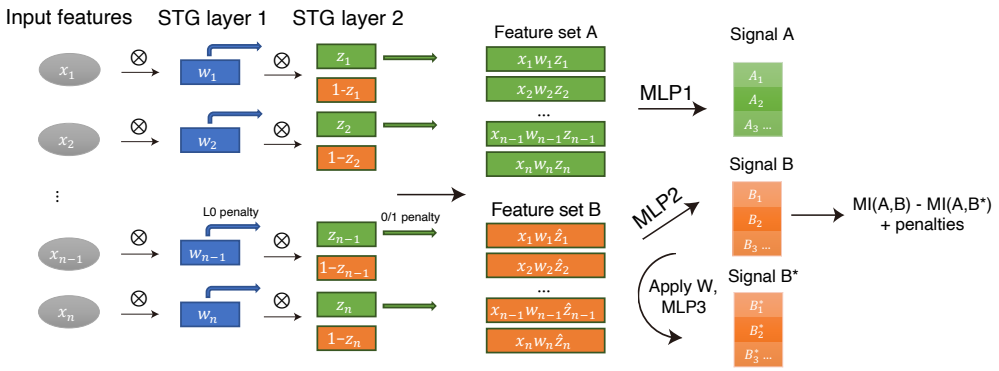


Figure 2: Overview of GEASS network architecture.

3.3 Output interpretation

Upon the algorithm convergence, GEASS provides both outputs of active features and embeddings produced by causally interacting features. In this paper, we emphasize the use of the former as the latter embedding output may be complex and nonlinear, potentially requiring additional architectures to maximize its interpretability.

By the construction of GEASS, we are able to get two separate sparse feature subsets as source features and sink features. These features may be used as inputs to further proper causal analysis, such as LPCMCI (Gerhardus and Runge, 2020), which despite its statistical power in depicting possible lags, identifying latent confounders, and allowing nonlinear tests, can only work on data with intermediate feature size. Also, these feature themselves may be used in downstream unsupervised analysis to improve generalizations of prediction models.

4 Experiments

4.1 Gaussian time-series with possible nonlinearity

In order to benchmark the method in time-series data, we consider two settings: 1. Minor effect of latent processes, with autocorrelation present; 2. Significant effect of latent processes, with autocorrelation present. Both settings are modelled by Gaussian structural processes with an underlying causal graph. Further details can be seen in Appendix C.1.

We test the false discovery rate (FDR) and F1 score between ground truth interacting features and recovered features as two metrics for high-dimensional data causal discovery. We compare GEASS with two categories of methods, namely conditional independence based (CI-based) methods and Granger causality based (GC-based) methods respectively. The first method category includes VAR-LINGAM (Hyvärinen et al., 2010), PCMCI (Runge et al., 2019b), and LPCMCI (Gerhardus and Runge, 2020). Among them, despite the statistical power, LPCMCI is not included in our experiment as it fails to converge in given time in our preliminary experiments. The second method category includes a neural-network based generalized vector autoregression model GVAR Granger (Marcinkevičs and Vogt, 2021), and Grid-net which generalizes the definition of Granger causality to Directed Acyclic Graph (DAG) (Wu et al., 2021); moreover we include two state-of-the-art approaches, DCM and NGM implemented in (Bellot et al., 2021) that use neural ODE to model nonlinear dependence graph.

Table 1 shows our benchmarking result. Among the alternative methods, GVAR and GrID-net fail in all settings as they are not designed for causal feature selection. VAR-LINGAM achieves high accuracy in linear settings while fails in nonlinear settings. In contrast, PCMCI fails when latent processes contribute to both true causally interacting features and nuisance features, creating spurious correlation. Empirically we also observe that DCM and NGM achieves comparable performance when the dynamics are linear but performs worse in the nonlinear setting, where the dynamics are

more irregular. Finally, GEASS consistently gives accurate causal feature identifications (high F1) and low false discovery rate (low FDR) in all settings considered.

Table 1: Comparison of methods for Gaussian linear / nonlinear time-series data with different feature numbers and different nuisance feature settings.

	Weak confounding interactions				Strong confounding interactions			
	Linear		nonlinear		Linear		nonlinear	
Methods	FDR	F1 score	FDR	F1 score	FDR	F1 score	FDR	F1 score
LINGAM (CI)	.00 (.00)	.94 (.04)	.83 (.00)	.17 (.00)	.00 (.00)	.94 (.04)	.83 (.00)	.17 (.00)
PCMCI (CI)	.17 (.01)	.81 (.04)	.12 (.08)	.85 (.05)	1.0 (.00)	.00 (.00)	.63 (.23)	.36 (.22)
GVAR (GC)	.94 (.00)	.11 (.00)	.94 (.00)	.11 (.00)	.94 (.00)	.11 (.00)	.94 (.00)	.11 (.00)
GrID-net (GC)	1.0 (.00)	.00 (.00)	1.0 (.00)	.00 (.00)	1.0 (.00)	.00 (.00)	1.0 (.00)	.00 (.00)
DCM (GC)	.12 (.20)	.88 (.20)	.65 (.12)	.35 (.12)	.18 (.09)	.82 (.09)	.93 (.11)	.07 (.11)
NGM (GC)	.07 (.08)	.88 (.04)	.48 (.17)	.50 (.17)	.00 (.00)	.91 (.00)	.62 (.25)	.38 (.25)
GEASS (Ours)	.05 (.15)	.97 (.10)	.03 (.06)	.92 (.05)	.03 (.07)	.90 (.04)	.00 (.00)	.91 (.00)

Furthermore, we evaluate different methods' scalability with respect to the feature size. (Experimental details can be seen at Appendix C.1.2). As described before, we anticipate high computational complexity of both conditional independence based methods and neural network based methods with respect to the feature size, which prohibits further use of these methods for high-dimensional biological data analysis, where the feature number is at the scale of $10^3 - 10^4$. Meanwhile, GEASS constructs a single neural network with parameters approximately proportional to p, thus largely reducing the complexity in the high-dimensional regime. We benchmark PCMCI, GVAR, GrID-net, NGM, GEASS, and an additional combination of GEASS with a downstream CI-test based causal graph identification method LPCMCI. Our experimental result shows the superior performance of GEASS as well as GEASS+LPCMCI in time complexity, consistent with our qualitative analysis (Figure 3).

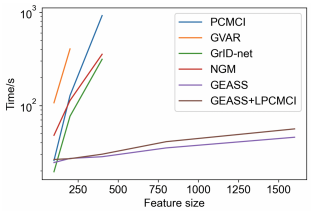


Figure 3: Running time comparison for methods on time-series data with different feature sizes.

4.2 Simulated spatial omics data with cell type confounder

In order to jointly consider spatial confounders and corresponding autocorrelation patterns that are potentially enriched in specific niches, we consider the case of spatial omics data, where autocorrelation is modeled by higher likelihood of same type of cells in the neighborhood, and confounder is modelled by coherent shift of global gene expression for each cell type. We first simulate scRNA-seq data, then the synthetic scRNA-seq data are assigned to a fixed size grid with cell type labels simulated by Ising model simulation. We then add artificial genes that are spatially correlated with neighboring cell's given gene set. Finally the dataset is normalized and log1p transformed as the standard pipeline in Scanpy (Wolf et al., 2018).

The majority of methods are not available as their focus is on time-series data. Therefore in order to perform our benchmarking study, we compare GEASS with Lasso Granger, as well as our implemented L1-regularized version of NCEM, an approach proposed to detect interactions in spatial omics data (Fischer et al., 2021). Finally, we also implemented a method that maximizes over original transfer entropy to select causal features (TE).

As shown in table 2, original LASSO cannot identify causal features because of strong correlation between features. L1-NCEM ameliorates the issue by conditioning on the cell type label in regression. TE outperforms linear methods yet generate a number of false positives, as it may overfit to learn spurious causations as discussed in remark 3.3. Finally, GEASS consistently outperforms over other methods in identifying causal features of data as shown by both high F1 score and low FDR.

Table 2: Comparison of methods for simulated spatial transcriptomics data.

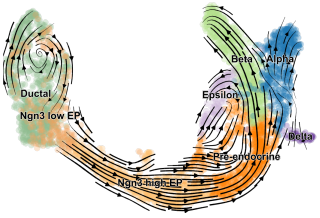
	Linear		nonlinear	
Methods	FDR	F1 score	FDR	F1 score
Lasso	0.950±0.055	0.050±0.055	0.970±0.040	0.030±0.040
L1-NCEM	0.380±0.138	0.620±0.138	0.535±0.134	0.465±0.134
TE	0.190±0.127	0.767±0.070	0.141±0.087	0.761±0.060
GEASS (Ours)	**0.095±0.128**	**0.787±0.088**	**0.000±0.000**	**0.775±0.110**

4.3 scRNA-SEQ PANCREATIC ENDOCRINOGENESIS TRAJECTORY

We test GEASS on the pancreatic endocrinogenesis trajectory data, which is a standard dataset for scRNA-seq trajectory inference task (Bergen et al., 2020; Bastidas-Ponce et al., 2019). The pancreas trajectory data contains 3696 cells and 27998 genes. After preprocessing, lowly-expressed genes are filtered out as the standard pipeline in `scVelo` (Bergen et al., 2020), with remaining 2000 genes for further analysis. We aim to use GEASS to identify causally-related genes along the developmental trajectory to reveal underlying biology. (See Appendix C.3 for experimental details).

scRNA-seq data provides a snapshot of cell population distribution therefore time-series based analysis methods cannot be directly applied. However, due to GEASS's flexible setting in forward operator W, we are able to define the time flow by RNA velocity analysis. RNA velocity analysis uses the additional information of intron RNAs to infer the underlying dynamics of gene expression change. Thus, we are able to define a velocity kernel matrix W_{velo}, which provides weighted adjacency relationships of cells based on velocity direction and cell phenotypal proximity.

GEASS identifies 50 causally-related features with high biological relevance. For example, the gene list includes the key transcriptional regulator *Neurog3*, which is required for the specification of a common precursor of the 4 pancreatic terminal states (uni, 2021). For further quantitative validation, we take advantage of the additional information given by RNA velocity, as significant RNA velocity is a necessary condition for causal dynamics of successor genes along the trajectory. Therefore, we benchmark different gene selection methods' performance on selecting dynamical genes. Specifically, we evaluate the mean velocity likelihood over GEASS causal genes with high-expressed genes (HEG), highly-variable genes (HVG), and genes having high correlation with inferred latent time (HCG). Our result suggests here GEASS achieves the best performance in selecting dynamical genes.



	Mean RNA velocity likelihood
HEG	0.0528
HVG	0.1753
HCG	0.1889
GEASS	**0.2366**

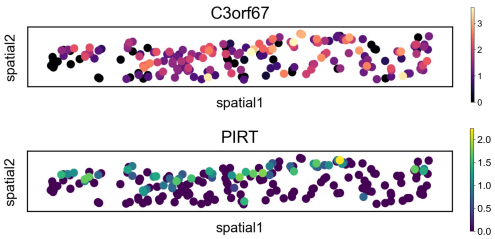
Figure 4: Visualization of pancreas trajectory dataset and comparisons of gene selection criterions.

4.4 MERFISH HUMAN CORTEX SINGLE-CELL LEVEL SPATIAL TRANSCRIPTOMICS

Spatial transcriptomics represent a wide category of method that can achieve spatial profiling of gene expression in tissues (Moses and Pachter, 2022; Rao et al., 2021; Palla et al., 2022). By the additional information of spatial locations, such measurements enable deeper understandings of cellular interactions (Palla et al., 2022; Jerby-Arnon and Regev, 2022; Fischer et al., 2021). However, current computational methods revealing interaction modules (Jerby-Arnon and Regev, 2022) or niche effects (Fischer et al., 2021) for spatial omics data lacks causal interpretation. Applying GEASS, we aim to reveal underlying causal intercellular patterns to fully utilize the potential of spatial omics data for biological discovery.

Here we use GEASS on a recent published MERFISH data measuring spatially-resolved gene expression of human cortex (Fang et al., 2022). MERFISH is a multiplexed single-molecule imaging-based technology that enables measurements of spatially-resolved single-cell gene expression in tissues (Chen et al., 2015; Fang et al., 2022). The data comprises of 3044 cells and 4000 genes; each cell is annotated as one of the eight cell types: excitatory neurons (EXC), inhibitory neurons (INC), astrocytes (ASC), microglial cells (MGC), oligodendrocytes (OGC), oligodendrocyte progenitor cells (OPC), endothelial cells (ENDO), and mural cells (MURAL) as shown by the first panel of figure 6 in Appendix D. We selected 1000 highly variable genes by Scanpy preprocessing pipeline. After preprocessing, our GEASS analysis selects 9 genes, namely *FILIP1, SLC17A7, MYH11, RP11-10j21.2, PIRT, C3orf67, TRDMT1, RGS8, SPTLC2*. An overview of the gene expression is available in the rest panels of figure 6. Among these genes, *Myh11*, *RP11-10j21.2*, and *TRDMT1* are enriched at the spatial regions with neighboring endothelial and mural cells, corresponding to underlying vascular structures (marked by ellipses in the first panel of figure 6). The finding is consistent with the MERFISH images, which reveals the spatial selective proximity of oligodendrocyte and microglial cell bodies to vascular structures, while astrocytes contact blood vessels with their end feet (Fang et al., 2022). Therefore, these identified marker genes of vascular structure encode underlying meaningful cellular interaction, supporting the validity of GEASS.

In order to show GEASS's ability to facilitate biological findings, we focus on two GEASS identified genes, *C3orf67* and *PIRT*. These two genes are enriched at the same spatial locations, both expressing prominently in astrocytes. Further statistical analysis reveals that the model of *PIRT* affecting neighboring *C3orf67* is favored. This is consistent with the predicted role of *PIRT* in enabling transmembrane transporter binding activity and being involved in phosphatidylinositol-mediated signaling (Safran et al., 2021), and previous works highlighting the gene's role in pain sensing (Kim et al., 2008; Wang et al., 2021). As the role of C3orf67 in human cortex remains unclear, this revealed causal link may lead to further discoveries.



	Pearson r	P value
(C3orf67, PIRT)	0.173	0.014*
(C3orf67, PIRT$_{nb}$)	0.139	0.049*
(PIRT, C3orf67$_{nb}$)	0.201	0.004**

Figure 5: Analysis of genes *C3orf67* and *PIRT* highlighting the possible causal mechanism of *PIRT* $\rightarrow$ *C3orf67*.

5 Conclusions

In this work, we present GEASS, a causal feature selection method based on information-theoretic tools and neural networks. GEASS is able to scale to high dimensions and identify sparse interacting features. We provide both theoretical gaurantee and empirical validation of GEASS on synthetic and real biological data. Our results show GEASS can be integrated into high-dimensional spatiotemporal data analysis pipeline to provide unique insights for further findings.

Limitations. GEASS is a method designed for nonlinear causal feature selection. GEASS does not provide a causal graph itself as it optimizes a latent embedding corresponding to different causal mechanisms. Therefore, in applications where a causal graph output is favored, constraint-based methods may need to be applied after GEASS. Moreover, when underlying causal graph has a large number of vertices, the sparsity assumption is violated and GEASS is not gauranteed to work. Also, further efforts may be taken to incorporate lag selections for GEASS.

Broader impact. We anticipate wide use of GEASS in high-dimensional graph-structured data, especially for high-dimensional biological data such as single cell trajectories and spatial omics measurements. Applying GEASS along with causal graph identification methods to a wider range of real biological data may greatly facilitate downstream biological discovery.

REFERENCES

Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. Computational methods for trajectory inference from single-cell transcriptomics. *European journal of immunology*, 46(11):2496–2506, 2016.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239, 2019.

Rongxin Fang, Chenglong Xia, Jennie L Close, Meng Zhang, Jiang He, Zhengkai Huang, Aaron R Halpern, Brian Long, Jeremy A Miller, Ed S Lein, et al. Conservation and divergence of cortical cell organization in human and mouse revealed by merfish. *Science*, 377(6601):56–62, 2022.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022.

Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

Giovanni Palla, David S Fischer, Aviv Regev, and Fabian J Theis. Spatial components of molecular tissue biology. *Nature Biotechnology*, 40(3):308–318, 2022.

Kevin D Hoover. Causality in economics and econometrics. 2006.

Maciej Kamiński, Mingzhou Ding, Wilson A Truccolo, and Steven L Bressler. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological cybernetics*, 85(2):145–157, 2001.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019a.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319, 2022.

Konrad P Mielke, Tom Claassen, J Huijbregts, Aafke M Schipper, and Tom M Heskes. Discovering cause-effect relationships in spatial systems with a known direction based on observational data. In *International Conference on Probabilistic Graphical Models*, pages 305–316. PMLR, 2020.

David S Fischer, Anna C Schaar, and Fabian J Theis. Learning cell communication from spatial graphs of cells. *BioRxiv*, 2021.

PA Valdés-Sosa, JM Bornot-Sánchez, M Vega-Hernández, L Melie-García, A Lage-Castellanos, and E Canales-Rodríguez. granger causality on spatial manifolds: applications to neuroimaging. handbook of time series analysis: recent theoretical developments and applications, 18.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Matthäus Staniek and Klaus Lehnertz. Symbolic transfer entropy. *Physical review letters*, 100(15): 158101, 2008.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128, 2010.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*, pages 23–47. PMLR, 2018.

Alessio Moneta, Nadine Chlaß, Doris Entner, and Patrik Hoyer. Causal search in structural vector autoregressive models. In *NIPS Mini-Symposium on Causality in Time Series*, pages 95–114. PMLR, 2011.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

Atalanti A Mastakouri, Bernhard Schölkopf, and Dominik Janzing. Necessary and sufficient conditions for causal feature selection in time series with latent common causes. In *International Conference on Machine Learning*, pages 7502–7511. PMLR, 2021.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019b.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

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

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Zexuan Yin and Paolo Barucca. Deep recurrent modelling of granger causality with latent confounding. *arXiv preprint arXiv:2202.11286*, 2022.

Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2021.

Saurabh Khanna and Vincent YF Tan. Economy statistical recurrent units for inferring nonlinear granger causality. *arXiv preprint arXiv:1911.09879*, 2019.

Xiangyu Sun, Guiliang Liu, Pascal Poupart, and Oliver Schulte. Nts-notears: Learning nonparametric temporal dags with time-series data and prior knowledge. *arXiv preprint arXiv:2109.04286*, 2021.

Alexander P Wu, Rohit Singh, and Bonnie Berger. Granger causal inference on dags identifies genomic loci regulating transcription. In *International Conference on Learning Representations*, 2021.

Xianjie Guo, Kui Yu, Lin Liu, Fuyuan Cao, and Jiuyong Li. Causal feature selection with dual correction. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

Luca Faes, Daniele Marinazzo, Giandomenico Nollo, and Alberto Porta. An information-theoretic framework to map the spatiotemporal dynamics of the scalp electroencephalogram. *IEEE Transactions on Biomedical Engineering*, 63(12):2488–2496, 2016.

Payam Shahsavari Baboukani, Carina Graversen, Emina Alickovic, and Jan Østergaard. Estimating conditional transfer entropy in time series using mutual information and nonlinear prediction. *Entropy*, 22(10):1124, 2020.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

Jingjing Zhang, Osvaldo Simeone, Zoran Cvetkovic, Eugenio Abela, and Mark Richardson. Itene: Intrinsic transfer entropy neural estimator. *arXiv preprint arXiv:1912.07277*, 2019.

Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In *International Conference on Machine Learning*, pages 10648–10659. PMLR, 2020.

Ofir Lindenbaum, Moshe Salhov, Amir Averbuch, and Yuval Kluger. L0-sparse canonical correlation analysis. In *International Conference on Learning Representations*, 2021.

Andrew Miller, Nick Foti, Alexander D'Amour, and Ryan P Adams. Reducing reparameterization gradient variance. *Advances in Neural Information Processing Systems*, 30, 2017.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12615–12625. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/94e70705efae423efda1088614128d0b-Paper.pdf`.

Ričards Marcinkevičs and Julia E Vogt. Interpretable models for granger causality using self-explaining neural networks. *arXiv preprint arXiv:2101.07600*, 2021.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12): 1408–1414, 2020.

Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J Theis, et al. Comprehensive single cell mrna profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12):dev173849, 2019.

Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, 49(D1):D480–D489, 2021.

Livnat Jerby-Arnon and Aviv Regev. Dialogue maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nature Biotechnology*, pages 1–11, 2022.

Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

Marilyn Safran, Naomi Rosen, Michal Twik, Ruth BarShir, Tsippi Iny Stein, Dvir Dahary, Simon Fishilevich, and Doron Lancet. The genecards suite. In *Practical guide to life science databases*, pages 27–56. Springer, 2021.

Andrew Y Kim, Zongxiang Tang, Qin Liu, Kush N Patel, David Maag, Yixun Geng, and Xinzhong Dong. Pirt, a phosphoinositide-binding protein, functions as a regulatory subunit of trpv1. *Cell*, 133(3):475–485, 2008.

Wanzhi Wang, Miaomiao Kong, Yu Dou, Shanghai Xue, Yang Liu, Yinghao Zhang, Weiwei Chen, Yanqing Li, Xiaolong Dai, Jianghui Meng, et al. Selective expression of a snare-cleaving protease in peripheral sensory neurons attenuates pain-related gene transcription and neuropeptide release. *International Journal of Molecular Sciences*, 22(16):8826, 2021.

Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, 8, 2019.

APPENDIX

A PROOFS

A.1 PROOF OF PROPOSITION 2.2.

Proposition 2.2. $\forall S_1, S_2 \subset \{1, ..., p\}, \mathrm{mTE}_W(X_{S_1}, X_{S_2}) > 0 \Rightarrow \mathrm{TE}_W(X_{S_1}, X_{S_2}) > 0$.

Proof. We have

$$\mathrm{TE}_W(X_{S_1}, X_{S_2}) = I(X_{S_1}; WX_{S_2}|X_{S_2})$$

$$= I(X_{S_1}; WX_{S_2}, X_{S_2}) - I(X_{S_1}; X_{S_2})$$

$$= I(X_{S_1}; WX_{S_2}) - I(X_{S_1}; X_{S_2}) + I(X_{S_1}; X_{S_2}|WX_{S_2}).$$

Therefore we have $\mathrm{TE}_W(X_{S_1}, X_{S_2}) \geq \mathrm{mTE}_W(X_{S_1}, X_{S_2})$ holds, therefore $\mathrm{mTE}_W(X_{S_1}, X_{S_2}) > 0 \Rightarrow \mathrm{TE}_W(X_{S_1}, X_{S_2}) > 0$.

A.2 DISCUSSION OF ASSUMPTION A6.

Our assumption A6 is based on the concept of conditional mutual entropy, which aims to filter out possible indirect causal relationships.

Here is a simple example to see why TE/mTE can have problems with indirect causal interactions: consider the relationship: $s_t \to w_t \to v_{t+1}$, then we would have: $I(s_t, v_{t+1}) > 0$ although there are no direct causal relationship between s and v.

The issue can be resolved by considering the conditional mutual information $I(s_t, v_{t+1}|w_t)$, which equals 0. This inspires the concept of conditional transfer entropy:

Definition (Conditional transfer entropy) (Shahsavari Baboukani et al., 2020). Assume X and Y are the features of interest and the conditioning features are Z. Denote $-$ as $[1, 2, ..., t]$, then we have

$$\mathrm{cTE}_{Z_t}(X, Y) = I(Y_{t+1}, X_-|Y_-, Z_-).$$

The classical formulation of conditional transfer entropy is widely used in high-dimensional observational data to learn direct causal dependencies (Faes et al., 2016; Shahsavari Baboukani et al., 2020). It implicitly assumes that, there is direct causal relationship between X and Y if $\forall Z, \mathrm{cTE}_{Z_t}(X, Y) > 0$. Here, we are further assuming that this also holds for cmTE, which we define as an analogy to cTE in the Markovian model setting:

Definition (Conditional mTE). Assume X and Y are the feature sets of interest and the conditioning features are Z. Then we have

$$\mathrm{cmTE}_{Z_t}(X, Y, j) = I(X_j, Y_{j+1}|Z_t) - \max_{k;l\neq 1} I(X_k, Y_{k+l}|Z_t).$$

We can see not only our definition is similar to cTE in form, the key idea of the conditional transfer entropy is still preserved: If X_j causally affects Y_{j+1} through Z_t, or dependency between X_j and Y_{j+1} is because of spurious correlation to Z_t, then we have $\mathrm{cmTE}_{Z_t}(X, Y, j) \leq 0$. Meanwhile, the assumption A6 also implies strong interaction as discussed by proposition 2.2. In summary, our A6 can be reformulated as $\forall Z_t, j, \mathrm{cmTE}_{Z_t}(X, Y, j) > 0$ for ground truth interacting X, Y and can be similarly defined for nuisance features.

A.3 PROOF OF THEOREM 2.4.

Theorem 2.4. *Given assumption A1-A6, $X^*_{\hat{S}_1 \cup \hat{S}_2} \subset (t_1, ..., t_m)$. Moreover, each feature in $\hat{S}_1 \cup \hat{S}_2$ is connected to other features in the set.*

Proof. Step 1. First we prove $S_1 \cap S_2 = \emptyset$. If not, assume the overlapping signal to be p, then for arbitrary S_1, S_2, assume the corresponding sources are A and B respectively. Then we have

$$I(WA, Wp; B, p) - I(A, p; B, p) - I(WA, Wp; B) - I(A, p; B)$$
$$= I(WA, Wp; p|B) - I(A, p; p|B) = I(WA, Wp; p|B) - I(p; p|B) < 0.$$

Therefore removing p would increase the value of mTE, leading to a contradiction.

Step 2. Now we prove nuisance signals cannot be in either $\hat{S}_1$ or $\hat{S}_2$.

Otherwise, first we assume S is in $\hat{S}_2$, as S only interacts with variables at the same time point, S can only interact with WA via indirect links through a subset of interacting features. Assume this feature set as $Pa(WA)^S$, then with A6, we have

$$I(WA; S, B, Pa(WA)^S - B) - I(A; S, B, Pa(WA)^S - B) - I(WA; S, B) + I(A; S, B)$$
$$= I(WA; Pa(WA)^S - B|S, B) - I(A, Pa(WA)^S - B|S, B)$$
$$= I(WA; Pa(WA)^S - B|S, B) - I(A, Pa(WA)^S - B|S, B) > 0.$$

Moreover, we have

$$I(WA; B, Pa(WA)^S - B) - I(A; B, Pa(WA)^S - B) - I(WA; S, B, Pa(WA)^S - B)$$
$$+ I(A; S, B, Pa(WA)^S - B)$$
$$= -I(WA; S|B, Pa(WA)^S - B) + I(A; S|B, Pa(WA)^S - B) = 0 + I(A; S|B, Pa(WA)^S - B) \geq 0.$$

Therefore we have

$$mTE(A; B, Pa(WA)^S - B) \geq mTE(A; S, B, Pa(WA)^S - B) > mTE(A; S, B),$$

thus leading to a contradiction.

Then we are able to assume S is in $\hat{S}_1$, and B does not include any nuisance features. Then as S only interacts with variables at the same time point, WS can only interact with B via indirect links through a subset of interacting features. Assume the whole intermediate feature set for B as $Pa(WS)^B$, then we first note that we have $Pa(WS)^B \cap WS = \emptyset$. then with A6, we have

$$I(WA, WS, Pa(WS)^B - WS; B) - I(A, S, Pa(S)^B - S; B) - I(WA, WS; B) + I(A, S; B)$$
$$= I(Pa(WS)^B - WS; B|WA, WS) - I(Pa(WS)^B - WS; WB|WA, WS) > 0.$$

Moreover, we have

$$I(WA, Pa(WS)^B - WS; B) - I(A, Pa(S)^B - S; B) - I(WA, WS, Pa(WS)^B - WS; B)$$
$$+ I(A, S, Pa(S)^B - S; B)$$
$$= -I(WS; B|WA, Pa(WS)^B - WS) + I(S; B|A, Pa(S)^B - S) = 0 + I(S; B|A, Pa(S)^B - S) \geq 0.$$

Therefore we have

$$mTE(A, Pa(S)^B - S; B) \geq mTE(A, Pa(S)^B - S, S; B) > mTE(A, S; B),$$

thus leading to a contradiction.

Step 3. Moreover, if there exists a component in $\hat{S}_1 \cup \hat{S}_2$ not connected to any other feature components in the set, assume the feature to be q. Then, in this case with A1-4, the feature q is independent of any other features in $\hat{S}_1 \cup \hat{S}_2$. From step 1 it can be deduced that q cannot be in both $\hat{S}_1, \hat{S}_2$. Therefore in this case, we have $\mathrm{mTE}(\hat{S}_1 - q, \hat{S}_2 - q) = \mathrm{mTE}(\hat{S}_1, \hat{S}_2)$ thus leading to the contradiction of finding an $(\hat{S}_1, \hat{S}_2)$ with smaller $|\hat{S}_1 \cup \hat{S}_2|$.

A.4 PROOF OF THEOREM 3.1.

Theorem 3.1. *Assume assumptions A1-A6 holds. Then for small enough $\lambda > 0$:*

$$\min_{f,g,h,S_1,S_2} -(I(f(X \odot \mathbb{1}_{S_1}); h(Wg(X \odot \mathbb{1}_{S_2}))) - I(f(X \odot \mathbb{1}_{S_1}); g(X \odot \mathbb{1}_{S_2}))) + \lambda|S_1 \cup S_2|$$

We have 1. the optimal h is a one-to-one mapping; 2. Assume f, g define one-to-one mappings on $X \odot \mathbb{1}_{S_1}, X \odot \mathbb{1}_{S_2}$ respectively. Then, $X^*_{S_1 \cup S_2} \subset (t_1, ..., t_m)$. Moreover, each feature in $S_1 \cup S_2$ is connected to other features in the set.

$$\min_{f,g,h,S_1,S_2} -(I(f(X \odot \mathbb{1}_{S_1}); h(Wg(X \odot \mathbb{1}_{S_2}))) - I(f(X \odot \mathbb{1}_{S_1}); g(X \odot \mathbb{1}_{S_2}))) + \lambda|S_1 \cup S_2|$$

Proof. Step 1. function h must be injective in the optimal solution as non-one-to-one mapping reduces mutual information (Standard property of mutual information).

Step 2. Given the assumption that f, g define injective mappings on $\mathbb{1}_{S_1}, \mathbb{1}_{S_2}$ respectively, and G and W are commutable as $G(WX_t) = G(X_{t+1}) = WG(X_t)$, and one-to-one transformation does not change mutual information, we have

$$\begin{aligned} &I(f(X \odot \mathbb{1}_{S_1}); Wg(X \odot \mathbb{1}_{S_2})) - I(f(X \odot \mathbb{1}_{S_1}); g(X \odot \mathbb{1}_{S_2})) \\ &= I(X \odot \mathbb{1}_{S_1}; WX \odot \mathbb{1}_{S_2}) - I(X \odot \mathbb{1}_{S_1}; X \odot \mathbb{1}_{S_2}) \end{aligned}$$

Therefore using theorem 2.4, in the case of DAG causal graph, we have optimal $f(X)$ and $g(X)$ define one-to-one mappings to non-overlapping subsets of $\{t_1, ..., t_m\}$.

Step 3. With our definition of optimal S_1, S_2, the minimal gap between the optimal solution and any other solution with smaller $|S_1 \cup S_2|$ size is larger than zero. Denote the minimal gap as δ, then for these other solutions with smaller $|S_1 \cup S_2|$, we have

$$\begin{aligned} &\min_{f,g,h,S_1,S_2} -(I(f(X \odot \mathbb{1}_{S_1}); h(Wg(X \odot \mathbb{1}_{S_2}))) - I(f(X \odot \mathbb{1}_{S_1}); g(X \odot \mathbb{1}_{S_2}))) + \lambda|S_1 \cup S_2| \\ &\geq -\mathrm{mTE}^* + \delta + 0. \end{aligned}$$

Meanwhile, for solutions with larger $|S_1 \cup S_2|$ size, we have

$$\begin{aligned} &\min_{f,g,h,S_1,S_2} -(I(f(X \odot \mathbb{1}_{S_1}); h(Wg(X \odot \mathbb{1}_{S_2}))) - I(f(X \odot \mathbb{1}_{S_1}); g(X \odot \mathbb{1}_{S_2}))) + \lambda|S_1 \cup S_2| \\ &\geq -\mathrm{mTE}^* + \lambda(|S_1 \cup S_2| + 1). \end{aligned}$$

Meanwhile, for the desired optimal S_1, S_2, we have

$$\begin{aligned} &\min_{f,g,h,S_1,S_2} -(I(f(X \odot \mathbb{1}_{S_1}); h(Wg(X \odot \mathbb{1}_{S_2}))) - I(f(X \odot \mathbb{1}_{S_1}); g(X \odot \mathbb{1}_{S_2}))) + \lambda|S_1 \cup S_2| \\ &= -\mathrm{mTE}^* + \lambda|S_1 \cup S_2|. \end{aligned}$$

Therefore, when taking $\lambda \in (0, \frac{\delta}{|S_1 \cup S_2|})$, the desired optimal S_1, S_2 by mTE is the optimal output of the constructed optimization problem.

A.5 PROOF OF THEOREM 3.4.

Theorem 3.4. *Given assumptions A1-A6 and assumption on f, g described in theorem 3.1, denote for the optimal solution of (*), a sample of stochastic gate is given by T^1, T^2 and ground truth active gates as $T_{S_1^*}, T_{S_2^*}$, then we have for small enough $\lambda_1 > 0$ and $\lambda_2 > 0$,*

$$\mathbb{P}(\lim_{N \to \infty} \lim_{\epsilon \to 0} \mathbb{1}_{T^1 > \epsilon, T^2 < \epsilon} = T_{S_1^*}) = 1; \mathbb{P}(\lim_{N \to \infty} \lim_{\epsilon \to 0} \mathbb{1}_{T^1 > \epsilon, T^2 > \epsilon} = T_{S_2^*}) = 1.$$

$$\begin{aligned} &\mathbb{E}_X \mathbb{E}_{T^1, T^2} - (I(f(X_{S_1}); h(Wg(X_{S_2}))) - I(f(X_{S_1}); g(X_{S_2}))) + \lambda_1 ||T^1||_0 + \lambda_2 ||\mathbb{1}_{T^2 \in [0,1]}||_0, \\ &s.t. \quad X_{S_1} = X \odot T^1 \odot T^2, \quad X_{S_2} = X \odot T^1 \odot (1 - T^2). \quad (*) \end{aligned}$$

Proof. Step 1. Given f, g, h projects input distributions into Gaussian distributions, by convergence of Gaussian covariance matrices, we have:

Empirical mTE

$$= \hat{I}(f(X \odot T^1 \odot T^2); h(Wg(X \odot T^1 \odot (1-T^2)))) - \hat{I}(f(X \odot T^1 \odot T^2); g(X \odot T^1 \odot (1-T^2)))$$

$$= I(f(X \odot T^1 \odot T^2); h(Wg(X \odot T^1 \odot (1-T^2)))) - I(f(X \odot T^1 \odot T^2); g(X \odot T^1 \odot (1-T^2))) + o(1);$$

$$P(\lim_{N\to\infty} \text{Empirical mTE} = \text{mTE}) = 1.$$

Step 2. Upon step 1, applying Glivenko–Cantelli theorem on the stochastic gate terms, we have (*) equals

$$= I(f(X \odot T^1 \odot T^2); h(Wg(X \odot T^1 \odot (1-T^2)))) - I(f(X \odot T^1 \odot T^2); g(X \odot T^1 \odot (1-T^2)))$$

$$+\lambda_1(\sum_d P(T_d^1 > 0) + o(1)) + \lambda_2(\sum_d P(T_d^2 \in (0,1)) + o(1)) + o(1).$$

$$= I(X \odot T^1 \odot T^2; WX \odot T^1 \odot (1-T^2)) - I(X \odot T^1 \odot T^2; X \odot T^1 \odot (1-T^2))$$

$$+\lambda_1(\sum_d P(T_d^1 > 0) + o(1)) + \lambda_2(\sum_d P(T_d^2 \in (0,1)) + o(1)) + o(1).$$

Therefore, as $N \to \infty$, applying step 1 of theorem 2.4, for any $\lambda_2 > 0$, we have $\forall d, P(T_d^2 = 0) = 0/1, P(T_d^2 = 0) = 1/0$ holds, in which case the penalty for T^2 is zero.

Applying above, noting the stochastics are i.i.d sampled, independent with X, we have

$$I(X \odot T^1 \odot T^2; X \odot T^1 \odot (1-T^2))$$

$$= I(X \odot T^1 \odot T^2; X \odot T^1 \odot (1-T^2) | T^2 \odot T^1) + I(T^2 \odot T^1; X \odot T^1 \odot (1-T^2))$$

$$-I(T^2 \odot T^1; X \odot T^1 \odot (1-T^2) | X \odot T^1 \odot T^2)$$

With the non-overlapness of $T^1 \odot T^2$ and $T^1 \odot (1-T^2)$, the second and third term above equals zero. Therefore we have

$$I(X \odot T^1 \odot T^2; X \odot T^1 \odot (1-T^2)) = I(X \odot T^1 \odot T^2; X \odot T^1 \odot (1-T^2) | T^2 \odot T^1)$$

Use the same steps, we eventually have

$$I(X \odot T^1 \odot T^2; X \odot T^1 \odot (1-T^2)) = I(X \odot T^1 \odot T^2; X \odot T^1 \odot (1-T^2) | T^2 \odot T^1, (1-T^2) \odot T^1)$$

$$= I(X \odot (T^1 \odot T^2 > 0); X \odot (T^1 \odot (1-T^2) > 0) | (T^2 \odot T^1 > 0), ((1-T^2) \odot T^1 > 0)).$$

Also for the first term of mTE we similar have

$$I(X \odot T^1 \odot T^2; WX \odot T^1 \odot (1-T^2)) = I(X \odot T^1 \odot T^2; WX \odot T^1 \odot (1-T^2) | T^2 \odot T^1, (1-T^2) \odot T^1)$$

$$= I(X \odot (T^1 \odot T^2 > 0); WX \odot (T^1 \odot (1-T^2) > 0) | (T^2 \odot T^1 > 0), ((1-T^2) \odot T^1 > 0)).$$

Then we eventually have the optimization problem as

$$\mathbb{E}_{T^1,T^2} - I(X \odot (T^1 \odot T^2 > 0); WX \odot (T^1 \odot (1-T^2) > 0) | (T^2 \odot T^1 > 0), ((1-T^2) \odot T^1 > 0))$$

$$+I(X \odot (T^1 \odot T^2 > 0); X \odot (T^1 \odot (1-T^2) > 0) | (T^2 \odot T^1 > 0), ((1-T^2) \odot T^1 > 0)) + \lambda_1 ||T^1||_0$$

From theorem 3.1, with λ_1 small enough, we have above

$$\geq \min_{T^1,T^2} -I(X \odot (T^1 \odot T^2 > 0); WX \odot (T^1 \odot (1-T^2) > 0) | (T^2 \odot T^1 > 0), ((1-T^2) \odot T^1 > 0))$$

$$+I(X \odot (T^1 \odot T^2 > 0); X \odot (T^1 \odot (1-T^2) > 0) | (T^2 \odot T^1 > 0), ((1-T^2) \odot T^1 > 0)) + \lambda_1 ||T^1||_0$$

$$= -\text{mTE}^* + \lambda_1 |S_1 \cup S_2|,$$

which can be achieved when $P((T_d^1 \odot T_d^2 > 0) = 1/0$ when $d \in / \notin S_1$, and $P((T_d^1 \odot (1-T_d^2) > 0) = 1/0$ when $d \in / \notin S_2$.

Note above holds when $N \to \infty$ with probability 1, where we have

$$\lim_{\epsilon\to 0} \mathbb{1}_{T^1>\epsilon, T^2<\epsilon} = T_{S_1}^*; \lim_{\epsilon\to 0} \mathbb{1}_{T^1>\epsilon, T^2>\epsilon} = T_{S_2}^*.$$

Therefore we have proven strong consistency of the stochastic-gate based estimator.

B GATE INITIALIZATION

Our proposed initialization scheme is based on analysis of the linear case. Assume

$$f(X_{S_1}) = Xa, g(X_{S_2}) = Xb,$$

where $a, b \in \mathbb{R}^p$ represents two feature loadings. Then:

1. a, b should be non-overlapping, therefore we expect $|a^T b|$ to be small.
2. We should have $f(X) \approx Wg(X)$ to maximize the mTE.

Based on the two criteria above, we formulate an optimization problem as

$$\min ||a^T b||, \quad s.t.\ f(X) = Wg(X).$$

The constraint can be formulated into a regression problem $WXb = Xa$, therefore a natural solution is given by $a = X^{\dagger}WXb = (X^T X)^{-1}X^T WXb$. In this case, $||a^T b|| = ||b^T (X^T X)^{-1}X^T WXb|| = ||b||^2_{(X^T X)^{-1}X^T WX}$. Given b is normalized, it can be shown that the optimal b corresponds to the eigenvector with least absolute eigenvalue of matrix $(X^T X)^{-1}X^T WX$.

After getting a, b, we select a quantile threshold over $a/(a + b)$ to initialize the second stochastic gate layer. The first stochastic gate layer is initialized with uniform weights.

C EXPERIMENTAL DETAILS

C.1 TIME-SERIES BENCHMARKING STUDY

In the study the causal processes is simulated with Python package `Tigramite`. Among the total 100 features, there are 6 interacting features $\{1, 2, 3, 4, 5, 6\}$. The causal links are: 1->2 with time lag 2, 2->3 with time lag 1, 5->4 with time lag 1, 1->5 with time lag 1, 3->6 with time lag 3. These features also have autocorrelations with time lags ranging from 1 to 3. There is also a latent confounder modeled by `Tigramite` interacting with feature 0 and feature 2. In the case of strong latent process, the latent confounder also have effects on other 43 features. All other features (93/50) not mentioned above are nuisance features with white noise dynamics. The forward operator is defined by 5-neighbor lower triangular matrix.

C.1.1 ALGORITHM IMPLEMENTATION

If not particularly mentioned, default settings of the algorithms are used throughout.

- **VAR-LINGAM.** The VAR-LINGAM algorithm is implemented in the Python package `LINGAM`, available at https://github.com/cdt15/lingam. VAR-LINGAM gives an weighted matrix as output. Therefore in our benchmarking study, we choose the most significant edge corresponding features with the number matching the sparsity level.
- **PCMCI.** The PCMCI algorithm is implemented in the Python package `Tigramite`, which gives an weighted matrix as output. We choose the most significant edge corresponding features with the number matching the sparsity level.
- **GVAR.** The GVAR algorithm is implemented at https://github.com/i6092467/GVAR. The sparsity parameter is set to be 1. We use the stable training option in GVAR, which trains the first and second half of the time series respectively to optimize over edge selection sparsity level then train on the whole time series, giving a binary output and no threshold selection is needed.
- **Grid-net.** The Grid-net algorithm is implemented at https://github.com/alexw16/gridnet. The parameter set: order=5, hidden_layer_size = 10, end_epoch=50, batch_size = 50, lmbd=1 is used throughout our study. After the training finishes, we choose the most significant edge corresponding features with the number matching the sparsity level.
- **DCM, NGM.** The two algorithms are both implemented at https://github.com/alexisbellot/Graphical-modelling-continuous-time. For DCM, the default setting is used, and we use hidden dim = 10 for NGM. After both training finishes, we choose the most significant edge corresponding features with the number matching the sparsity level.

- **GEASS.** We use the same training parameters in all time-series settings, with the key sparsity regularization parameter λ_1 set with 0.04/0.05 based on a validation set, and the rest parameter settings are consistent with default.

C.1.2 SCALABILITY ANALYSIS

We test PCMCI, GVAR, GrID-net, NGM, GEASS, GEASS+LPCMCI's running time with consistent settings described in the above section. (LPCMCI's setting is consistent with PCMCI's setting). We use the same data generation pipeline and select the set of the total feature numbers as $[100, 200, 400, 800, 1600]$.

C.2 SIMULATED SPATIAL OMICS DATA BENCHMARKING STUDY

In the study the spatial omics data is simulated with Python package `Scsim` (Kotliar et al., 2019). 1000 genes are simulated in total, while 990 genes are cell-type-specificly expressed. The rest 10 genes each has a functional relationship (linear/nonlinear) with one cell-type-specific genes plus the noise term in order to model the cell-type-specific interactions. The data is then normalized and log-transformed according to the standard Scanpy pipeline. (Wolf et al., 2018) The forward operator is defined by 4-neighbor adjacency matrix.

C.2.1 ALGORITHM IMPLEMENTATION

If not particularly mentioned, default settings of the algorithms are used throughout.

- **Lasso Granger.** The Lasso algorithm is implemented by Scipy with tuned α (0.12) to match the sparsity level.
- **NCEM.** NCEM (Linear) is a linear graph neural network, which in the grid case corresponds to a standard linear regression based on neighbors and the cell type label. Based on the original work, we implemented our equivalent version by Lasso regression with $\alpha = 0.019$ to match the sparsity level.
- **GEASS.** We use the same training parameters in all settings, with the key sparsity regularization parameter λ_1 set with 0.02 based on a validation set, and the latent dimension number is set to be 64.
- **TE.** To give a fair comparison, we use the same architecture as GEASS except for the loss function is changed. We use the same training parameters in all time-series settings, with the key sparsity regularization parameter λ_1 set with 0.05 based on a validation set, and the latent dimension number is set to be 64 consistent with GEASS.

C.3 SCRNA-SEQ PANCREAS TRAJECTORY

The data preprocessing is consistent with the `scVelo` tutorial: https://scvelo.readthedocs.io/ VelocityBasics/. (Bergen et al., 2020) The parameter set: $\lambda_1 = 0.06, \lambda_2 = 0.1$. Here because the gene regulatory network is fully connected and activated in cascade along the developmental trajectory, we consider the opposite initialization with b be the largest eigenvalues corresponding eigenvectors of the matrix $(X^T X)^{-1} X^T W X$.

C.4 MERFISH SPATIAL TRANSCRIPTOMICS DATA

The data is downloaded from Dryad and preprocessed with the standard Scanpy pipeline. (Wolf et al., 2018) The forward operator is defined by 5-neighbor adjacency matrix. The GEASS parameter set is consistent with those used in the spatial omics benchmarking.

D ADDITIONAL EXPERIMENTAL RESULTS

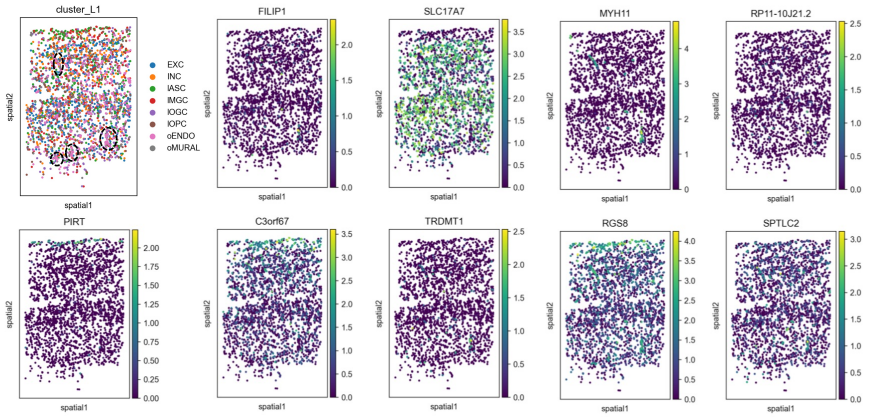


Figure 6: Spatial profiling of MERFISH human cortex slice, colored by cell type annotation and GEASS identified gene expressions. The ellipses in the first panel represent examples of vascular structures.