A Discussion of our assumptions and their implications

For our theoretical work we assume continuous, noise-free, fully-observed trajectories, which may at first glance appear limiting. Indeed, those assumptions would be limiting if this was a method proposal, where the method only works on such observations. However, for analyzing unidentifiability, they are the right starting point: Without these assumptions, no system would ever be identifiable, i.e., unidentifiability would be trivially given. Only once we have eliminated unidentifiability due to, e.g., undersampling, finite samples, and partial observations, can we make non-trivial claims about the inherent unidentifiability of the underlying systems. We split the discussion of our assumptions into two parts.

1. The assumptions of *continuous* trajectories, *noise-free* observations and *full observability* are assumptions about our ability to observe/measure the system. For these, our setting corresponds to the "best case" idealized setting where (in principle) infinite data (both in terms of sampling frequency as well as in terms of noise instantiations) are available about all relevant parts of the system. As we will discuss one-by-one below, violating any of those assumptions will add additional identifiability issues, which are due to limitations of our observation process (not inherent to the system) that one could in principle overcome. Hence, instead of "our results only apply under those assumptions," our results show unidentifiability "despite these assumptions", even when all solvable sources of unidentifiability due to limited measurement processes have been overcome. In short, these form the appropriate starting point for an identifiability analysis.
2. The assumption of linear systems and sparsity are indeed assumptions about the underlying dynamics. We elaborate on the relevance of these assumptions below.

1a. Continuous trajectories. Any real-world measurement device will have a maximum sampling frequency, so all real-world data is necessarily discrete in time. Discrete observations add trivial unidentifiability issues around aliasing (cf. aliasing in the Nyquist sampling theorem). Hence, the continuous observation case is the best-case scenario - the limit of infinite sampling frequency - where we expect any remaining unidentifiability to be due to the fundamental problem setting, not our insufficient sampling rate.

1b. Noise-free observations. Again, real-world data are almost always noisy. And again, in the presence of noise, system identification becomes inherently more difficult, namely probabilistic: multiple different systems may be compatible with the observed data, each with a different likelihood under the assumed noise model (for example, the standard assumption of i.i.d. additive Gaussian noise in classical machine learning often used to model "measurement errors"). Within this probabilistic framework, our noise-free setting can be interpreted as the best case scenario, where unidentifiability is not just due to suboptimal measurement devices. It can also be interpreted as the asymptotic limit of observing infinitely many trajectories for a single initial condition such that the "noise can be averaged out" to recover the noise-free setting. From this perspective, our results state under which conditions collecting more observations allows (in the asymptotic limit) the identification of the system. Or put differently, we show that there are cases in sparse systems where, even if more data was collected or where the noise level is reduced by other means, the true system remains unidentifiable.

1c. Full observations. Akin to discrete or noisy observations, only having partial observations of the system state will naturally render identifiability more difficult. Unlike discrete observations and noise, partial observations add a fundamental degree of unidentifiability as there is no "asymptotic limit" of more frequent or precise observations to recover our setting. Intuitively, when we allow for the presence of (an arbitrary number of) unobserved state variables, both the full system as well as the "observed part" (viewed as a submatrix of the full system) are inherently unidentifiable. As an extreme (and somewhat pathological) counterexample, one could imagine an exact copy of each observed variable among the latents such that all observed variables may either depend on the unobserved copy or the observed variable, which is indistinguishable from data. Hence, to study the identifiability under partial observations, one would have to provide additional limiting assumptions on the allowed type and nature of confounding. Exploring the palette of possible assumptions is an interesting direction for future work.

2a. Linear systems. There are two main reasons for this assumption: (a) Linear differential equation models are a core pillar of scientific modeling and widely used in practice due to their analytical

tractability, relative simplicity, and inherent interpretability. While the real underlying dynamics of complex systems are arguably rarely perfectly linear, linear models also turn out to be surprisingly useful and predictive (akin to linear regression still being a competitive and reliable tool in many settings). (b) Non-linear ODEs are known to be unidentifiable from single trajectories if one does not impose heavy constraints on the allowed non-linear function family (Scholl et al., 2023). Hence, the previous proclaimed state in the literature was a clear separation: "non-linear systems are unidentifiable while linear systems are almost surely identifiable." Our work adds important nuances to the part about linear systems.

2b. Sparsity. If it were somehow known that real-world systems never evolve according to dynamics below the sparsity-identifiability-threshold, our findings would indeed be of purely theoretical interest. However, the true sparsity of real-world dynamics remains inherently unknown. While it is impossible to determine whether a given unknown system is sparse, the examples in our submission indicate that this may not be unlikely in practice. Our results are thus important in that modeling practitioners should consider them when interpreting their results, e.g., despite a perfectly-fit model, the inferred weights might not reflect the true causal relations between variables due to unidentifiability issues. As a pragmatic step to highlight that high degrees of sparsity may occur in practically relevant settings, we examined published sparse systems, which we discuss in 5.1. In both reported cases, the published systems are on the sparser side of the threshold required for identifiability, which further supports the practical relevance of our findings. Nonetheless, we emphasize that the relevance of our results in terms of how frequently they show up cannot be empirically proven, since the notion of sparsity in our analysis refers to that of the true but unknown dynamics, rather than the model dynamics (models cannot be proven "true"). However, a practical implication for future method development efforts aimed at inferring the "ground truth" networks is for instance that such methods may produce estimates that fit the observed data equally well as the ground truth, yet are different. Importantly, this discrepancy might not stem from the limitations of the developed method itself, but rather from the nature of the underlying dynamics being estimated.

B THEORETICAL RESULTS

B.1 PROOF OF LEMMA 2

Lemma 2. *A sparse-continuous random matrix with sparsity p is globally not identifiable with probability at least $1-(1-p^n)^n - np^n(1-p^n)^{n-1}$ for $n \geq 2$ (and p for $n = 1$).*

Proof. Following Theorem 1, the model Eq. (1) is globally unidentifiable if and only if A has more than one Jordan normal block for one of its eigenvalues, say λ_i, which then has geometric multiplicity $g(\lambda_i) > 1$, i.e., $\operatorname{rank}(A-\lambda_i I) < n-1$. According to Lemma 1, it follows that $P_A(\mathcal{U}) = P_A(\{A \in \mathbb{R}^{n\times n} \mid \operatorname{rank}(A) < n-1\})$. Hence, we are looking for the probability of dependencies among columns in A that drop the rank. For a fixed zero structure $B := (b_{i,j})_{i,j\in[n]} \in \{0,1\}^{n\times n}$ the set on which $X := (x_{i,j})_{i,j\in[n]} \in \mathbb{R}^{n\times n}$ introduces additional linear dependencies among the rows/columns of A has Lebesgue measure zero and thus zero probability under A. Therefore, $P(\operatorname{rank}(A) < n-1) = P(\operatorname{rank}(B) < n-1)$. We focus on the sufficient event that B has multiple zero columns. Since the entries of B are independent, $P(B_{:,i} = 0) = p^n$ for all $i \in [n]$ where $B_{:,i}$ represents the i-th column, it follows that the random variable Z representing the count of zero columns follows a Binomial distribution $Z \sim \operatorname{Bin}(n,q)$ with $q := p^n$. With the probability mass function $P(Z=k) = \binom{n}{k} q^k (1-q)^{n-k}$ we find

$$P(\operatorname{rank}(A) < n-1) \geq P(Z \geq 2) = 1 - P(Z \in \{0,1\}) = 1 - (1-p^n)^n - np^n(1-p^n)^{n-1}\,.$$

□

Lemma 2 provides a non-zero lower bound on the probability of system level unidentifiability for the simple iid Bernoulli sparsity pattern. However, by analogous reasoning the proof extends to a much broader class of structured sparsity patterns that also allow for the existence of hubs, i.e., for heavy tailed degree distributions.

Lemma 5 (General sparsity-pattern lower bound)**.** *Let P_B be a probability measure on $\{0,1\}^{n\times n}$ and let $B \sim P_B$. Consider a sparse-continuous random matrix $A := (B_{ij}X_{ij})_{i,j=1}^n$, where $(X_{ij})_{i,j=1}^n$*

are independent continuous random variables (i.e., their joint distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{n\times n}$) and independent of B. Assume that for some $k \in \{0, 1, \dots, n-2\}$ there exists a set of columns $J \subseteq [n]$ with $|J| \geq k+2$ and a set of rows $R \subseteq [n]$ with $|R| \leq k$ such that the event

$$E_k := \Big\{ B \in \{0,1\}^{n\times n} \;\Big|\; B_{ij} = 0 \text{ for all } j \in J,\ i \notin R \Big\} \tag{5}$$

has positive probability $q := P_B(E_k) > 0$. Then A is globally unidentifiable with probability at least q, i.e., $P(A \in \mathcal{U}) \geq q > 0$.

Proof. The key idea is that every deterministic matrix $M \in \mathbb{R}^{n\times n}$ whose zero pattern belongs to E_k has $\operatorname{rank}(M) \leq n-2$ and is therefore system level unidentifiable. For such matrices, for every $j \in J$ the j-th column $M_{:,j}$ is supported only on rows in R. Hence each $M_{:,j}$ with $j \in J$ lies in an $|R|$-dimensional subspace. We thus found at least $k+2$ columns of M that lie in the same $|R| \leq k$ dimensional subspace, which decreases the rank of M by at least 2 and by Theorem 1, M is globally unidentifiable. Now consider the random matrix $A = (B_{ij}X_{ij})$. Since the above argument only depends on B, we conclude that $\operatorname{rank}(A) \leq n-2$ whenever $B \in E_k$. Hence $P(A \in \mathcal{U} \mid B) = 1$ for all $B \in E_k$ and by total probability

$$P(A \in \mathcal{U}) = \mathbb{E}_B\big[P(A \in \mathcal{U} \mid B)\big] \geq \mathbb{E}_B\big[\chi_{E_k}(B)\, P(A \in \mathcal{U} \mid B)\big] = \mathbb{P}_B(E_k) = q > 0\,.$$

□

The main takeaway is that as long as there may be some variables that are sparsely connected among each other with a sufficient degree of sparsity there will be a non-zero probability of system level identifiability—in contrast to the dense setting.

B.2 Proof of Lemma 3

Lemma 3 (sharp threshold for global unidentifiability)**.** *Let A be a sparse-continuous matrix with n-dependent sparsity level $p(n)$. Then, for $p_c(n) = 1 - \frac{\ln(n)}{n}$ and any function $\omega(n) \to \infty$ we have that if $p(n) = p_c(n) + \frac{\omega(n)}{n}$, then $P(\operatorname{rank}(A) \leq n-2) \to 1$ for $n \to \infty$ and if $p(n) = p_c(n) - \frac{\omega(n)}{n}$, then $P(\operatorname{rank}(A) \leq n-2) \to 0$ for $n \to \infty$. That is, there is a threshold at $p = \ln(n)/n$ decisive for whether A is asymptotically globally unidentifiable with high probability or not.*

Proof. From A form the random bipartite graph $G_{n,n,s}$ with A as the corresponding adjacency matrix. Edges are present independently with probability $s = s(n) = 1 - p(n)$.

For every bipartite graph $G_{n,n,s}$ let

$$m(G) = \max\{|M| : M \text{ is a matching}\}, \qquad d(G) = \max_{S\subseteq[n]} \big(|S| - |N(S)|\big) \tag{6}$$

(Hall deficiency). From Hall-König's theory for graph matching, the rank of the adjacency matrix A

$$m(G) \;\leq\; \operatorname{rank} A \;\leq\; n - d(G). \tag{7}$$

Given

$$s(n) = \frac{\ln n + \alpha(n)}{n},$$

we have that (Frieze & Karoński, 2015) $P(m(G) = n) \to 1$ if $\alpha(n) \to \infty$ and $P(m(G) = n) \to 0$ if $\alpha(n) \to -\infty$.

Case $\alpha(n) \to -\infty$. Let Z be the number of isolated vertices in $G_{n,n,s}$. Without loss of generalization, we will consider isolated vertices from rows. Any fixed vertex is isolated with probability $(1-s)^n$ so

$$Z \sim \text{Bin}\big(n, (1-s)^n\big), \qquad \mathbb{E}[Z] = n(1-s)^n = e^{-\alpha(n)}(1+o(1)), \qquad \operatorname{Var}[Z] = O(\mathbb{E}[Z]).$$

Chebyshev's inequality therefore yields $P(Z \geq 2) \to 1$ for $n \to \infty$. Two isolated vertices form a set S with $|N(S)| \leq |S| - 2$, so $d(G) \geq 2$ and consequently $\Pr[\operatorname{rank} A \leq n-2] \to 1$.

Case $\alpha(n) \to +\infty$. Then $P(m(G) = n) \to 1$, hence $P(\operatorname{rank}(A) \leq n-2) \to 0$. Substituting $s = 1 - p$ concludes the proof. □

B.3 PROOF OF LEMMA 4

Lemma 4. *Let $A, A' \in \mathbb{R}^{n\times n}$ and assume there is an A-invariant subspace V^* such that $(A - A')V^* = \{0\}$. Then, for every $t \geq 0$ and $\boldsymbol{x}_0 \in \mathbb{R}^n$, we have that*

$$\|e^{At}\boldsymbol{x}_0 - e^{A't}\boldsymbol{x}_0\|_2 \leq C(t, A, A')\|A - A'\|_2\, d_A(\boldsymbol{x}_0)$$

with $C(t, A, A') := \int_0^t \|e^{A(t-s)}\|_2\|e^{A's}\|_2 ds$. Further, for any $\varepsilon > 0$, the condition

$$\|(e^{At} - e^{A't})\,\boldsymbol{x}_0\| \leq \varepsilon \quad \textit{for all } 0 \leq t \leq T$$

holds whenever

$$T \leq \frac{1}{\alpha}\, W\Big(\alpha\, \frac{\varepsilon}{\|A - A'\|\, M^2\, d_A(\boldsymbol{x}_0)}\Big),$$

where $W(\cdot)$ denotes the Lambert function and constants $\alpha \in \mathbb{R}, M \geq 1$ such that $\|e^{At}\|, \|e^{A't}\| \leq Me^{\alpha t}$ for all $t \geq 0$.

Proof. Let us assume without loss of generality (otherwise we simply get a loser bound) that V^* is the closest proper A-invariant subspace to $\boldsymbol{x}_0$. We can then decompose $\boldsymbol{x}_0$ as $\boldsymbol{x}_0 = \Pi_{V^*}\boldsymbol{x}_0 + \boldsymbol{w}$ with $\|\boldsymbol{w}\| = d_A(\boldsymbol{x}_0)$ (by definition, all norms are 2-norms). Given the assumption on A and A', we have

$$(e^{At} - e^{A't})\boldsymbol{x}_0 = (e^{At} - e^{A't})(\Pi_{V^*}\boldsymbol{x}_0 + \boldsymbol{w}) = (e^{At} - e^{A't})\boldsymbol{w}\,.$$

From a variation of constants approach and the triangle inequality, we have

$$\|e^{At} - e^{A't}\| = \Big\|\int_0^t e^{A(t-s)}(A - A')e^{A's}ds\Big\| \leq \int_0^t \|e^{A(t-s)}\|\|e^{A's}\|\|A - A'\|ds\,,$$

which ultimately gives the result

$$\|e^{At}\boldsymbol{x}_0 - e^{A't}\boldsymbol{x}_0\| \leq \|e^{At} - e^{A't}\|\,\|\boldsymbol{w}\| \leq \|A - A'\|\, d_A(\boldsymbol{x}_0)\int_0^t \|e^{A(t-s)}\|\|e^{A's}\|\,ds\,.$$

For the second statement, we note that

$$\|(e^{At} - e^{A't})\,\boldsymbol{w}\| \leq \|A - A'\|\, M^2\, \|\boldsymbol{w}\|\left(\int_0^t e^{\alpha t}\,\mathrm{d}s\right) = \|A - A'\|\, M^2\, d_A(\boldsymbol{x}_0)\, t\, e^{\alpha t}\,,$$

which is bounded by ε for $t \in [0, T]$ with

$$Te^{\alpha T} = \frac{\varepsilon}{\|A - A'\|\, M^2\, d_A(\boldsymbol{x}_0)}\,.$$

Using the definition of the Lambert W function $W(\cdot)$ we obtain

$$T = \frac{1}{\alpha}\, W\Big(\alpha\, \frac{\varepsilon}{\|A - A'\|\, M^2\, d_A(\boldsymbol{x}_0)}\Big)\,.$$

□

C EMPIRICAL EVIDENCE OF NON-IDENTIFIABILITY IN GRNS

We elaborate on the practical relevance of non-identifiability by discussing real-world gene regulatory networks (GRNs) that may exhibit structural features leading to non-identifiable dynamics. To illustrate this, we examined the gold-standard *E. coli* network from Marbach et al. (2012) (Supplementary Data 1). The adjacency matrix derived from this network contains several zero columns, indicating genes with no outgoing regulatory edges. This can be checked by analysing the `DREAM5_NetworkInference_GoldStandard_Network3.tsv` data. As discussed in our main text, zero columns in the adjacency matrix are a key source of non-identifiability. A second example provided by Marbach et al. (2012), the *S. cerevisiae* network (Network 4), also contains zero columns and thus similar issues.

For some other networks, such as *P. trichocarpa* and *A. thaliana*, we were not able to obtain the full gold-standard network, and so a definitive identifiability analysis is not possible. However, we

Table 1: Summary of real-world gene regulatory networks and their sparsity levels. The sparsity threshold corresponds to $1 - \frac{\log n}{n}$ from Lemma 3.

Name	**# Nodes** (n)	**# Edges**	**Sparsity** (p)	**Threshold** ($1 - \frac{\log n}{n}$)
A. thaliana	2,864	18,663	0.9977	0.9972
P. trichocarpa	1,690	9,268	0.9968	0.9957

summarize in Table 1 the available statistics from Walker et al. (2022), including the number of nodes, number of edges, sparsity level, and the theoretical sparsity threshold from Lemma 2 of our work.

In both cases, the sparsity level lies below the identifiability threshold, suggesting a high probability of non-identifiability. More broadly, the key point is that identifiability cannot be assessed purely from data. For any given system, even if it is theoretically identifiable, one cannot determine from observed data alone whether the inferred model corresponds to the ground truth. When fitting a linear ODE model to data from a sparse system, we may recover a model that exactly reproduces observations, yet it may be structurally incorrect. Thus, regardless of whether the true system is globally identifiable, it remains unidentifiable from data alone.

Moreover, the datasets discussed above should not be viewed as perfect ground truths. As noted by Walker et al. (2022, see §2.5), these "gold standard" networks are incomplete and potentially inaccurate. As such, both the exact sparsity patterns (*E. coli*) and sparsity levels (*P. trichocarpa*) may be unreliable, and conclusions about identifiability must be treated cautiously.

D Experimental details

D.1 Software

We provide the resources with the corresponding licenses used in this work in Table 2.

Table 2: Overview of resources used in our work.

Name	**Reference**	**License**
Python	(van Rossum & Drake, 2009)	PSF License
PyTorch	(Paszke et al., 2019)	BSD-style license
Numpy	(Harris et al., 2020)	BSD-style license
Pandas	(pandas development team, 2020; Wes McKinney, 2010)	BSD-style license
Matplotlib	(Hunter, 2007)	modified PSF (BSD compatible)
Scikit-learn	(Pedregosa et al., 2011)	BSD 3-Clause
SciPy	(Virtanen et al., 2020)	BSD 3-Clause
SLURM	(Yoo et al., 2003)	modified GNU GPL v2
networkx	(Hagberg et al., 2008)	BSD 3-Clause
JAX	(Bradbury et al., 2018)	Apache-2.0

D.2 Metrics

System-level identifiability metrics. To compute system level identifiability metrics, we perform a batched singular-value decomposition on every system matrix A using `jax.numpy.linalg.svd`. Subsequently, any singular value σ with $|\sigma| < 10^{-6}$ is treated as numerically zero. Eigenvalues are computed with `jax.numpy.linalg.eigvals` and the matrix rank is computed via `jax.numpy.linalg.matrix_rank` with tolerance level set to 10^{-6}.

Trajectory-level identifiability metrics. Smoothed condition number (SCN) analysis begins by constructing the empirical Gram matrix $\hat{\Sigma}_{xx} = YSY^T \in \mathbb{R}^{d\times d}$, where $Y = [x(t_1)\dots x(t_n)]$ collects one simulated trajectory and where the diagonal "smoothing" matrix S contains the numerical quadrature weights (trapezoidal rule by default `jax.numpy.trapz`). The condition number is estimated with `jax.numpy.linalg.cond`.

Normalized Hamming distance. We use the normalized Hamming distance computed on binary input matrices to compare the true system matrix A with the empirically estimated matrix $\hat{A}$. To this end we first binarize A and $\hat{A}$ via $B = \mathbb{I}_\tau(A)$ and $\hat{B} = \mathbb{I}_\tau(\hat{A})$ with threshold $\tau = 10^{-5}$ and where $\mathbb{I}_\tau$ is the indicator function that acts elementwise on matrix entries as

$$\mathbb{I}_\tau(a_{ij}) = \begin{cases} 1 & \text{if } a_{ij} > \tau; \\ 0 & \text{else} \end{cases}$$

The normalized Hamming distance between two matrices $B, \hat{B} \in \mathbb{R}^{n \times n}$ is then defined as $d_{\text{HMD}}(B, \hat{B}) = \frac{1}{n^2} \sum_{i,j} \mathbb{I}_{0.5}(B_{ij} \neq \hat{B}_{ij})$ where $B_{ij} \neq \hat{B}_{ij}$ has to be understood as a boolean comparison which evaluates to one under inequality and zero otherwise.

D.3 Empirical estimators

SINDy. SINDy, short for Sparse Identification of Nonlinear Dynamics (Brunton et al., 2016), is a widely adopted algorithm for system identification. It leverages a user-defined set of basis functions to execute L_2-regularized linear regression, mapping observations of the solution trajectory $\mathbf{x}(t_i)$ to their corresponding temporal derivatives $\dot{\mathbf{x}}(t_i)$. In practical applications, temporal derivatives are often unobservable, and SINDy estimates them through numerical finite difference approximations. We adopt the implementation available in `PySINDy` (de Silva et al., 2020), restrict the basis set to linear functions, and use the default optimization algorithm (sequentially thresholded least squares) which sets any coefficient whose magnitude falls below the user-defined threshold λ to zero. Model and optimizer come with several hyper-parameters out of which we tune the L_2-regularization strength (α), coefficient threshold (λ), finite difference approximation order and maximum number of iterations separately for each sample.

Regularization of SINDy. For each trajectory we select the pruning threshold $\lambda \in \{10^{-6}, \ldots, 10^{-1}\}$ that enforces the sparsity gate: after every ridge-regression step any coefficient whose magnitude falls below λ is hard-set to zero, so increasing λ enforces progressively sparser candidate systems. Complementing this, the ridge weight $\alpha \in \{0.001, 0.05, 0.1\}$ continuously shrinks the surviving coefficients toward the origin; larger values thus promote numerical stability without directly changing the zero pattern. For every trajectory, we select the optimal parameters $(\lambda, \alpha) = \text{argmax} R^2$ where the R^2 score is measured between the observed trajectory and the trajectory obtained by numerically solving the system estimate $\hat{A}$ for the observed initial value. Finally, the identifiability analysis is based on systems with well-fitted trajectories only, which we define as trajectories for which the estimate achieves $R^2 > 0.99$ and MSE $< 10^{-4}$. This regularization sets any coefficient that falls below λ threshold to zero-hence aggressively promoting sparsity, which might lead to problems in low dimensional settings, see Fig. 4. In cases of very high sparsity and low system dimensionality, most coefficients of the true system matrix will be zero. In this case thresholding coefficients to zero biases the model towards a smaller Hamming distance. As the dimensionality increases or the sparsity reduces, this effect vanishes as there are multiple non-zero coefficients.

Neural ODEs. Neural ODEs (**NODE**s) (Chen et al., 2018) use a parameterized function $f_\theta(x(t))$ to approximate the dynamics underlying the observed trajectories. Instead of using finite difference schemes to estimate temporal derivatives, NODEs numerically integrate f_θ to obtain a solution that can be directly compared to the observed trajectory. In practice f_θ is implemented as a neural network; since we focus on linear systems, we use a model with multiple linear layers and no activation functions. To promote sparsity, we incorporate an L1 regularization term into the loss function. The total loss consists of the mean squared error (MSE) of the trajectories, augmented by the regularization parameter λ multiplied by the L1 norm of the network's weights. To optimize the model we use the ode solvers implemented in `torchode` (Lienen & Günnemann, 2022), specifically the `Dopri5` solver in combination with the `IntegralController` for adaptive step size selection, with relative and absolute tolerances set to 1e-3 and 1e-6, respectively. Neural network parameters θ are optimized with PyTorch's `RMSprop` optimizer with a learning rate of 1e-3 to minimize the mean absolute error between the observed and predicted solution trajectory as in Chen et al. (2018). Optimization proceeds for 10000 iterations or until the loss falls below 10^{-5}.

Sparsity-regularization of Neural ODEs. For the LinearNODE experiments we first identify well-fitted trajectories which we define as trajectories for which the empirical estimate (after numerical

integration from the ground truth initial value) achieves $R^2 > 0.99$ or MSE $< 10^{-4}$. The proportion of well-fitted trajectories (among all trajectories) for different system dimensionalities n and sparsity levels p is displayed for different values of regularization weight $\lambda \in \{0, 10^{-1}, 10^{-2}\}$ in Fig. 5. Among the λ-values that yield well-fitted trajectories, we then select the one that most faithfully reproduces the sparsity pattern of the ground-truth system matrix A. Specifically, for every trajectory we count the zero entries in the matrix estimate $\hat{A}$ and in the true A, compute the absolute difference in these counts, and use this sparsity-mismatch score as our selection metric. The optimal λ for a given p is selected as the value that minimizes the average sparsity-mismatch across all well-fitted trajectories. The effect of regularization weight λ, system dimensionality n and sparsity p on the sparsity-mismatch between model estimate $\hat{A}$ and ground truth system matrix A is illustrated in Fig. 6. Our arguably very permissive model selection strategy reflects the idea that we are only interested in well-fitted models (as measured on the trajectory-level) in order to draw conclusions about (empirical) system identifiability rather than about optimization, model architecture or numerical issues.

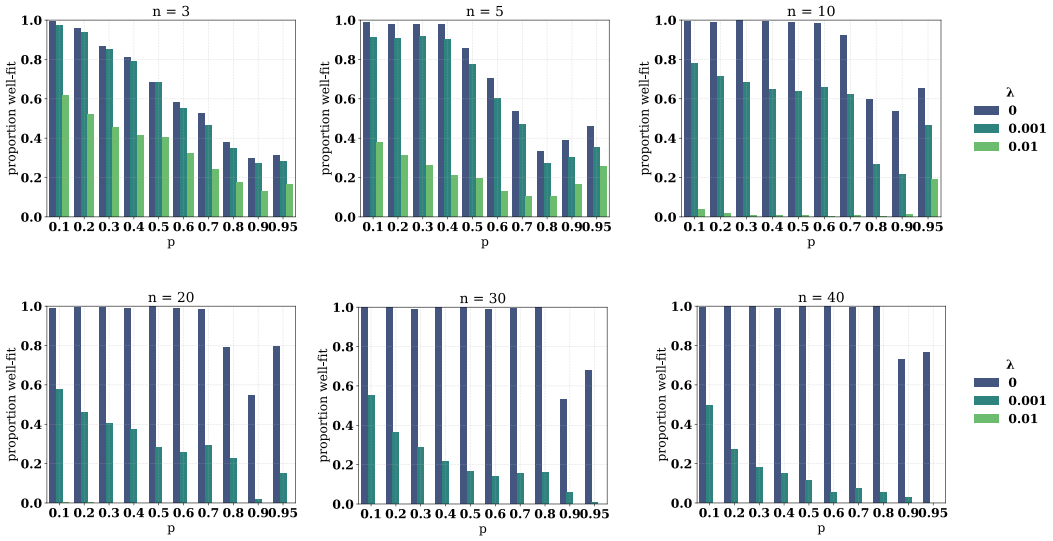

Figure 5: Proportion of trajectories that have been well-reconstructed by Sparse Neural ODEs for different regularization parameters λ and different dimensions n and sparsity levels p. For sparse systems, the model recovers only a smaller fraction of the trajectories.

D.4 Additional results on trajectory-level identifiability metrics

We extend the results on trajectory-level identifiability metrics d_A and SCN to a broad range of system dimensions n and sparsity levels p. Box-plots for the two subgroups $A_{\sigma_{2,\min}}$ and $A_{\sigma_{2,\max}}$ introduced in Section 5.3 are displayed in Fig. 7 and Fig. 8. We observe the consistent trend that subgroup $A_{\sigma_{2,\min}}$, i.e., the subgroup with smaller second smallest singular value σ_2, leads to lower identifiability scores for both metrics (lower value for d_A and higher value for SCN) in line with our theoretical results.

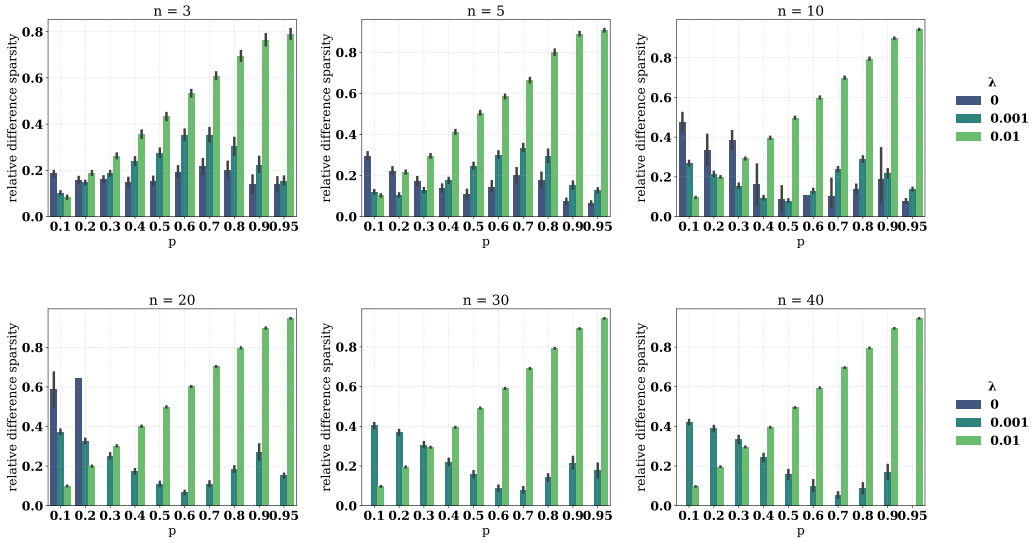

Figure 6: Relative difference in sparsity count (lower the better) between the true and reconstructed system matrices using Sparse Neural ODEs for different regularization parameters λ and different dimensions n and sparsity levels p. Lower regularization recovers dense matrices better, while higher values suit sparse ones.

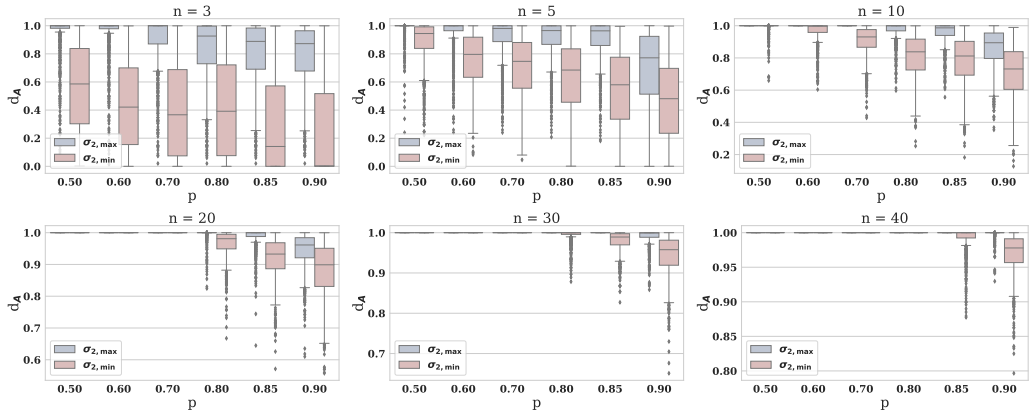

Figure 7: Box-plots of distance-to-unidentifiability d_A for the least and most identifiable groups of systems for different p and n values. Trajectories generated with $A_{\sigma_{2,\min}}$ lead to smaller d_A than those produced with $A_{\sigma_{2,\max}}$.

On the distance of a subspace of dimension n from a random vector on the unit sphere. We now empirically validate the close-to-unidentifiability metric d_A, which measures the distance between an initial condition x_0 and the kernel of the corresponding matrix A. Since we sample the initial conditions uniformly from the unit sphere, we can compare the empirical distribution of d_A to the theoretical expected distance between a random unit vector and a d_0-dimensional subspace of $\mathbb{R}^n$. We partition the trajectories by the null-space dimension $d_0 = \dim(\ker(A))$ of their corresponding generating matrix A and display the resulting distance-to-unidentifiability d_A in Fig. 9. As expected, the (mean) empirical measure closely matches the theoretical expected distance between a random unit vector $\boldsymbol{x}_0$ and a d_0-dimensional subspace of $\mathbb{R}^n$ (Vershynin, 2018), given by

$$\mathbb{E}[d_A(x_0) \mid n, d_0] = \frac{\Gamma(n/2)\Gamma((n-d_0+1)/2)}{\Gamma((n-d_0)/2)\Gamma((n+1)/2)},$$

hence (in expectation) validating the computation of our distance-to-unidentifiability d_A.

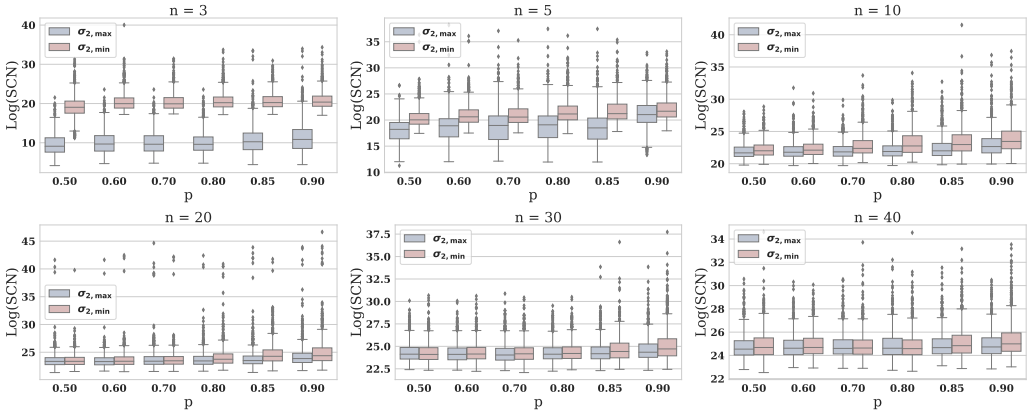

Figure 8: Box-plots of smoothed condition numbers (SCN) in log-scale for the least and most identifiable groups of systems for different p and n values. Trajectories generated with $A_{\sigma_{2,\min}}$ lead to higher SCN than those produced with $A_{\sigma_{2,\max}}$.

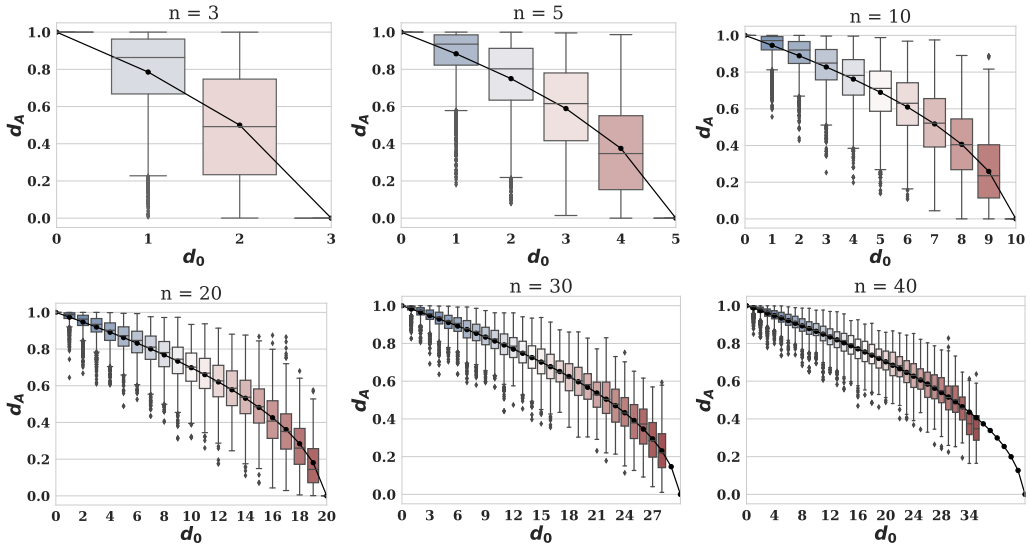

Figure 9: Box-plots of distance-to-unidentifiability d_A different n and different dimensions of $\ker(A)$ ($d_0 = \dim(\ker(A))$) together with the expected distance $\mathbb{E}[d_A(x_0) \mid n, d_0]$ (black line).

D.5 Additional Experiments

A consequence of our findings in Lemma 4 is that the trajectories generated by two candidate matrices A and A' remain indistinguishable over longer time horizons when the initial condition lies closer to an invariant subspace. Here we examine how this behavior manifests in the performance of empirical estimators. To this end, we consider multiple configurations (p, dim), where $p = \{0.1, 0.3, 0.7, 0.8, 0.9, 0.95\}$ and $dim = \{3, 5, 10, 20\}$. For each configuration we sample 10 matrices following the same data generation procedure as Section 5.3. We then generate four initial conditions at varying distances $d_A = \{0.1, 0.5, 1\}$ from the corresponding invariant subspace to assess how the estimators' performances change with increasing d_A. As described in Section D.1 we restrict the performance evaluation to well-fitted trajectories in order to investigate the effect of the distance on identifiability rather than model optimization properties. In Fig. 10, we display the Hamming distance and mean squared error (MSE) of the estimated matrices compared to the ground truth one. In line with our theory, the performance of both the Neural ODE as well as SINDy scales inversely with the distance to the invariant subspace, i.e., the smaller the distance d_A, the larger the estimation error.

A second factor that Lemma 4 predicts to affect empirical performance is the length of the observation interval: since trajectories remain indistinguishable up to time horizon T, reducing the observation window should degrade estimator performance. To investigate this aspect empirically, we use the same data from the previous experiment but restrict input to the estimator to only a fraction of the previous observation window. Our results in Fig. 11 show that indeed, as predicted, performance both in terms of Hamming distance as well as MSE decreases in almost all cases as the observation interval reduces. The exception to this observation occurs at the smallest distance to the invariant subspace, $d_A = 0.1$, where the performance remains constant across different trajectory lengths. As described before, $d_A = 0.1$ also leads to the largest errors overall, hence in this case the initial value is presumably already so close to the invariant subspace that unidentifiability even occurs at the longest observation window.

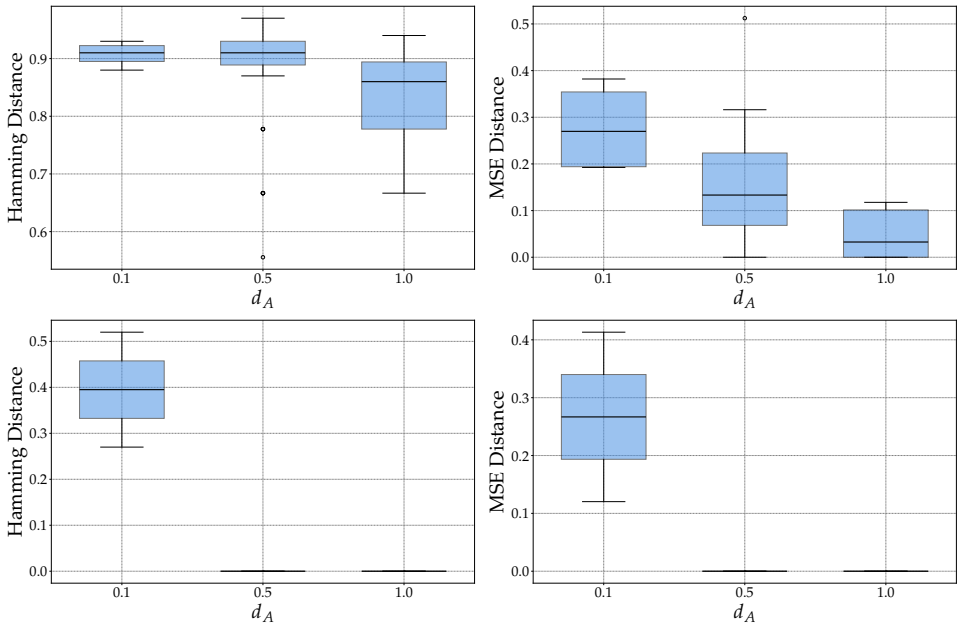

Figure 10: MSE and Hamming distance between the ground truth matrix and the matrix estimated from Neural ODE (top) and SINDy (bottom). The estimators' performance degrades with lower values of d_A.

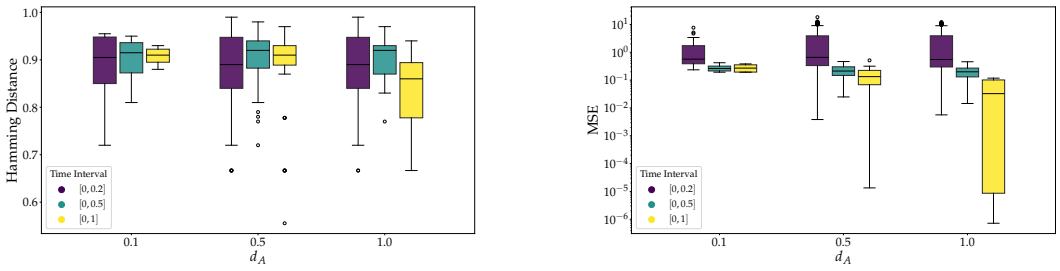

Figure 11: Hamming (left) and MSE (right) distance between the ground truth matrix and the matrix estimated from Neural ODE on trajectories spanning different time intervals.

D.6 EXTENSION TO FURTHER MATRIX MODELS

We focus on further random matrix models and present here the results.

Fixed number of zeros per row ensemble. In this random matrix model we fix the number of zeros per row such that each row contains exactly $d(n) = \lfloor np \rfloor$ zeros. The remaining non-zero coefficients are sampled i.i.d. as $a_{ij} \overset{iid}{\sim} N(0,1)$. We generate 100 system matrices and solve each of the for 100 initial values which are sampled uniformly at random from the unit circle in $\mathbb{R}^n$. Subsequently, we carry out the system-level identifiability and empirical identifiability analyses, following the same procedures described in the main text.

Results for both analyses are in line with the results reported in the main paper: system-level unidentifiability shows are sharp increase as sparsity increases (Fig. 12) and empirical identifiability shows a clear left-right gradient in Hamming distance for both SINDy (Fig. 15) and NODE (Fig. 16), confirming the expected rise in unidentifiability as sparsity increases.

Sparse-Continuous ensemble with no zero rows or columns. In this random matrix model we explicitly exclude matrices with zero rows or zero columns. For this, matrix entries are generated as $a_{ij} = g_{ij} b_{ij}$, where $b_{ij} \overset{iid}{\sim} Ber(p)$ and $g_{ij} \overset{iid}{\sim} N(0,1)$. We again attempt to generate 100 system matrices, however, not all dimensionality n and sparsity p combinations permit any system matrices with no zero rows or columns. (E.g. a 3×3 matrix with sparsity level 0.9 will have ≈ 8 zeros and hence always multiple zero rows and/or columns.) We hence cap the number of attempts to generate a single matrix that fulfills the zero-rows / zero-columns constraints at 100 attempts. For every valid generated system matrix, we sample 100 initial values randomly from the unit circle in $\mathbb{R}^{\ltimes}$ and numerically solve the initial value problem to obtain 100 trajectories per system.

We perform system-level identifiability analysis as well as an empirical identifiability analysis. System-level metrics are provided in Fig. 13. Hamming-distances for different system dimensions n and sparsity levels p are reported in Fig. 15 for SINDy and in Fig. 16 for NODEs. For both estimators the average Hamming distance increases as sparsity rises, confirming the trends observed for other matrix models as well as the theoretical underpinnings.

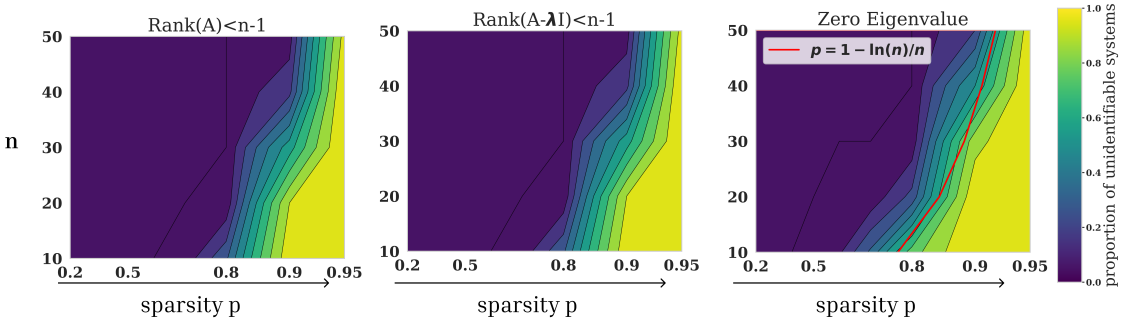

Figure 12: Proportion of matrices satisfying the conditions conditions $\mathrm{rank}(A - \lambda I) < n - 1$ (left), $\exists \lambda \in \mathbb{R} : \mathrm{rank}(A - \lambda I) < n - 1$ (center), and presence of zero eigenvalues (right) at different system dimensions n and sparsity levels p for **fixed number of zeros per row ensemble**.

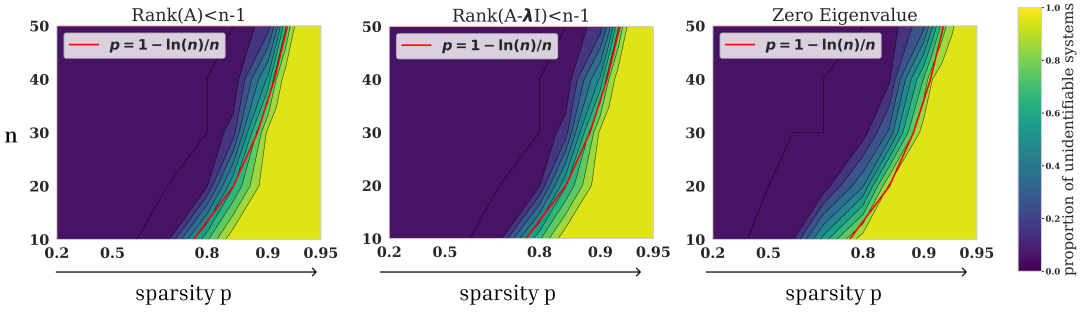

Figure 13: Proportion of matrices satisfying the conditions conditions $\mathrm{rank}(A - \lambda I) < n - 1$ (left), $\exists \lambda \in \mathbb{R} : \mathrm{rank}(A - \lambda I) < n - 1$ (center), and presence of zero eigenvalues (right) at different system dimensions n and sparsity levels p for **sparse-continuous ensemble with no zero rows**.

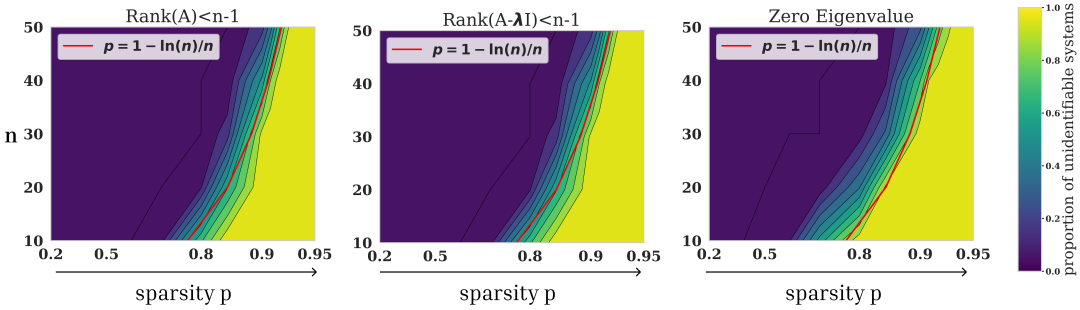

Figure 14: Proportion of matrices satisfying the conditions conditions $\mathrm{rank}(A - \lambda I) < n - 1$ (left), $\exists \lambda \in \mathbb{R} : \mathrm{rank}(A - \lambda I) < n - 1$ (center), and presence of zero eigenvalues (right) at different system dimensions n and sparsity levels p for **sparse-continuous ensemble with no zero columns**.

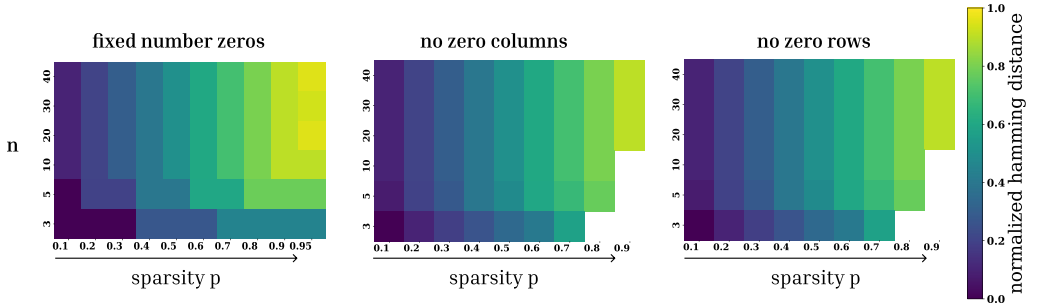

Figure 15: Hamming distance for different generating settings for SINDy on trajectories generated from **fixed number of zeros per row ensemble** (left), **sparse-continuous ensemble with no zero columns** (center), **sparse-continuous ensemble with no zero rows** (right).

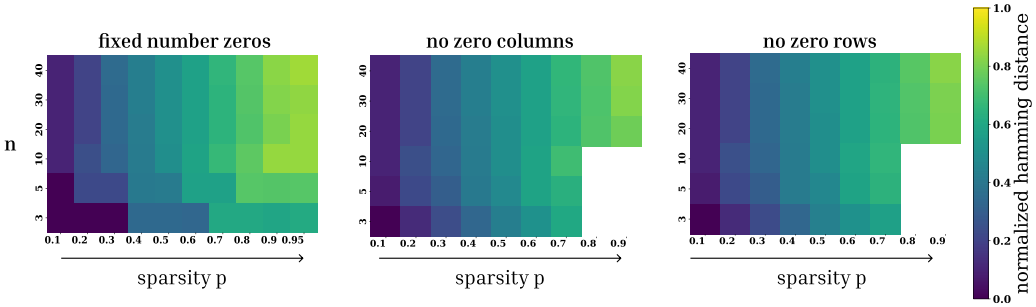

Figure 16: Hamming distance for different generating settings for NODE on trajectories generated from **fixed number of zeros per row ensemble** (left), **sparse-continuous ensemble with no zero columns** (center), **sparse-continuous ensemble with no zero rows** (right).