Determining minimal output sets that ensure structural identifiability

D. Joubert ⬡*, J. D. Stigter, J. Molenaar ⬡

Wageningen University and Research, Biometris, Department of Mathematical and Statistical Methods, Wageningen, The Netherlands

* dominique.joubert@wur.nl

Abstract

The process of inferring parameter values from experimental data can be a cumbersome task. In addition, the collection of experimental data can be time consuming and costly. This paper covers both these issues by addressing the following question: "Which experimental outputs should be measured to ensure that unique model parameters can be calculated?". Stated formally, we examine the topic of minimal output sets that guarantee a model's structural identifiability. To that end, we introduce an algorithm that guides a researcher as to which model outputs to measure. Our algorithm consists of an iterative structural identifiability analysis and can determine multiple minimal output sets of a model. This choice in different output sets offers researchers flexibility during experimental design. Our method can determine minimal output sets of large differential equation models within short computational times.

Introduction

Mathematical models are powerful tools that enable the scientific community to understand processes otherwise immeasurable by predicting outcomes of numerous physical properties. The field of systems biology often utilises ordinary differential equations to model dynamic systems. These models can comprise large systems of differential equations that contain vast numbers of unknown parameters [1]. Despite improvements in the quality of experimental sensors and therefore both the quality and quantity of experimental data, the process of parameter estimation remains cumbersome. This may be due to noisy data or due to the inherent structure of the model (structural unidentifiability) [2]. A structurally unidentifiable model implies that certain parameters are totally correlated, also referred to as 'aliased', and have confidence intervals that span the interval $(-\infty, \infty)$. Uncertainty in inferred parameter values calls into question the validity of the entire model and therefore it is imperative to address these uncertainties upfront by conducting identifiability analyses.

We will focus on ensuring structural identifiability and since this property can be analysed before conducting experiments, our analysis can be utilised in preliminary experimental design. An experimental researcher may wish to know: "Which of the pre-defined model outputs do I at least need to measure to ensure that I can infer unique parameter values?". The

answer is addressed by the topic of minimal output sets, where a minimal output set is defined as: *Measuring a minimal set of model outputs ensures that a model is structurally identifiable.* Due to its complexity, the topic of minimal output sets has received little attention [3]. Scientists often rely on intuitive experimental design, which may easily result in redundant or insufficient experimental measurements.

In this paper we present an algorithm to determine minimal output sets by identifying sets of totally correlated parameters using an iterative structural identifiability analysis. This algorithm offers insight into which states should be measured, thereby aiding intuitive experimental design. A particular model may have multiple minimal output sets. This offers great flexibility to the experimental researcher as he/she can decide which output set to measure taking factors such as time, cost and physical constraints into account.

This structural identifiability issue has been considered in a few previous papers [3–6], which we will briefly describe. The first paper, published in 2009, introduces a minimisation algorithm to determine which parameters are identifiable [4]. Three simple examples are included and due to its computational complexity the author states that defining minimal output sets for medium sized models is still too hard using this algorithm. In a paper published in 2012, the authors present an algorithm tasked with identifying symmetries, i.e. sets of totally correlated parameters, in a system of differential equations [3]. Once these symmetries have been identified, the states and parameters that destroy these symmetries are included into minimal output sets. Minimal output sets of the well-known NF-κB and JAK/STAT models are determined assuming that all model parameters and states can potentially be measured. The final step in their algorithm is doing a symbolic computation to test for structural identifiability and identify any remaining symmetries. Other papers address observability [5, 6]. Identifiability can be regarded as a special case of observability [7]. In [5], the authors introduce a graphical method and illustrate its key concepts using non-linear models. They construct a directed graph from the so-called adjacency matrix and inspect it to identify strongly connected components and more specifically root strongly connected components. Two nodes are classified as strongly connected if they are reachable from each other [8]. Root strongly connected components are strongly connected components with no outgoing edges. Minimum output sets are identified from the different elements in these root strongly connected components. A different approach is followed by Letellier and co-authors [6]. They use a symbolically computed Jacobi matrix to compute the output sets that ensure observability. An interesting extension of minimal output sets in the preliminary experimental design phase, could be to determine these sets taking measurement noise into account, thereby establishing practical identifiability. To this end, Docherty and co-authors present a graphical method to identify such sets [9].

Our minimal output set algorithm is different from the existing techniques as it numerically identifies sets of unidentifiable parameters. Through a number of computational experiments, we provide evidence (but not a complete mathematical proof) that our proposed algorithm has the following attributes:

- It can calculate the minimal output sets of large models.

- It can easily be adjusted to allow for cases in which only a limited subset of predefined outputs are measurable. This is illustrated in example 7 in the results and discussion section.

- Non-rational models can also be analysed as shown in example 8 in the results and discussion section.

The numerical findings are validated in a second step using symbolic computations as explained in [10]. This paper is divided into the following sections: Section 2 covers the

underlying theory and concepts of our algorithm. Section 3 showcases the algorithm using 8 examples and the final section contains concluding remarks.

Materials and methods

Background theory

Many dynamic systems biology phenomena are described in terms of differential equation models. These models can often be written in the standard state-space form [11]:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{\theta}), \tag{1}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{2}$$

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}(t), \boldsymbol{\theta}). \tag{3}$$

State variables are contained in a vector $\boldsymbol{x}(t)$ (dim $\boldsymbol{x} = n$), model parameters are contained in vector $\boldsymbol{\theta}$, (dim $\boldsymbol{\theta} = p$) and the output signals or measured variables are contained in vector $\boldsymbol{y}(t)$ (dim $\boldsymbol{y} = m$). Function $\boldsymbol{f}$ denotes a dynamic model structure and $\boldsymbol{h}$ is the output or observation function. Our approach allows for functions $\boldsymbol{f}$ and $\boldsymbol{h}$ to be either rational or non-rational. Unknown initial conditions of model states in vector $\boldsymbol{x}_0$, can be regarded as additional unknown parameters and can accordingly be included into $\boldsymbol{\theta}$. If all initial conditions are unknown, $\boldsymbol{\theta}$ contains $n + p$ elements.

The identifiability analysis method used in this paper was first proposed in [10]. In essence, this method relies on the singular value decomposition (SVD) of an output sensitivity matrix. Reid introduced the concept of sensitivity based identifiability analysis for linear models [12]. In his paper, he defines a sensitivity matrix as $\boldsymbol{S} = \partial \boldsymbol{y}/\partial \boldsymbol{\theta}$, with its elements describing the sensitivities of the model output with respect to model parameters. These partial derivatives are evaluated for nominal parameter values $\boldsymbol{\theta}_0$. Let $\Delta\boldsymbol{\theta}$ denote a small perturbation of the nominal vector $\boldsymbol{\theta}_0$, so $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}$. This perturbation will result in a corresponding perturbation in the model output, $\boldsymbol{y}(\boldsymbol{\theta}) = \boldsymbol{y}(\boldsymbol{\theta}_0) + \Delta\boldsymbol{y}$. A first order Taylor series approximation can be used to relate these perturbations [13, 14]:

$$\boldsymbol{y}(\boldsymbol{\theta}) - \boldsymbol{y}(\boldsymbol{\theta}_0) \approx \boldsymbol{S} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \quad \text{or} \quad \Delta\boldsymbol{y} \approx \boldsymbol{S} \cdot \Delta\boldsymbol{\theta}. \tag{4}$$

To solve $\Delta\boldsymbol{\theta}$ from the measured $\Delta\boldsymbol{y}$ uniquely, $\boldsymbol{S}^T \boldsymbol{S}$ should be non-singular [15–17] and therefore $\boldsymbol{S}$ should be of full rank [18, 19]. For non-linear models, the individual sensitivities are obtained by deriving the model (1)–(3) with respect to $\boldsymbol{\theta}$, thereby obtaining the system:

$$\frac{d}{dt}\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}\right) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}}, \tag{5}$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}. \tag{6}$$

To obtain the output sensitivity matrix $\boldsymbol{S}$, the matrix function $\partial \boldsymbol{y}/\partial \boldsymbol{\theta}$ is evaluated over a discretised finite time grid, $[t_0, \ldots, t_N]$ and the obtained matrices at each time point are concatenated [10]. It is advantageous to normalise $\boldsymbol{S}$ to adjust for sensitivities measured in different units [20]. We emphasise that working with normalised matrix elements is numerically

attractive but not essential. The normalised matrix S_{norm} is given as:

$$S_{norm} = \begin{pmatrix} \frac{\theta_1}{y_1(t_0)}\frac{\partial y_1}{\partial \theta_1}(t_0) & \cdots & \frac{\theta_{p+n}}{y_1(t_0)}\frac{\partial y_1}{\partial \theta_{p+n}}(t_0) \\ \vdots & \ddots & \vdots \\ \frac{\theta_1}{y_m(t_0)}\frac{\partial y_m}{\partial \theta_1}(t_0) & \cdots & \frac{\theta_{p+n}}{y_m(t_0)}\frac{\partial y_m}{\partial \theta_{p+n}}(t_0) \\ \vdots & & \vdots \\ \frac{\theta_1}{y_1(t_N)}\frac{\partial y_1}{\partial \theta_1}(t_N) & \cdots & \frac{\theta_{p+n}}{y_1(t_N)}\frac{\partial y_1}{\partial \theta_{p+n}}(t_N) \\ \vdots & \ddots & \vdots \\ \frac{\theta_1}{y_m(t_N)}\frac{\partial y_m}{\partial \theta_1}(t_N) & \cdots & \frac{\theta_{p+n}}{y_m(t_N)}\frac{\partial y_m}{\partial \theta_{p+n}}(t_N) \end{pmatrix}. \tag{7}$$

If all the initial conditions of model states are unknown, matrix S_{norm} (and also S) has dimensions $M \times (n + p)$, with $M = m \times (N + 1)$. To determine the rank of S_{norm} (and also S), the numerical rank test using a SVD reads as [15]:

$$S_{norm} = U\Sigma V^T. \tag{8}$$

The 2 matrices of importance are, the diagonal matrix, Σ (dim $M \times (n + p)$), and V (dim $(n + p) \times (n + p)$). The singular values in Σ, σ_i, $i = 1, \ldots, n + p$, are used to determine whether or not S_{norm} (or S) is of full rank. The rank of S_{norm} (or S) is the number of non-zero singular values and this can be expressed as follows [15]:

$$\text{If } \sigma_1 \geq \ldots \geq \sigma_q > \sigma_{q+1} = \ldots = \sigma_{n+p} = 0, \text{ then rank } (S_{norm}) = q. \tag{9}$$

In practice, singular values are never exactly vanishing due to numerical rounding errors. That is why one uses as practical definition: zero-valued singular values are values that fall beyond a significant gap in the spectrum of singular values [21]. In this paper we consider a gap larger than 3 decades on the log scale as significant. Once structural unidentifiability has been established, the non-zero entries of the singular vectors of matrix V, related to vanishing singular values beyond this gap, allude to which model parameters and initial conditions may be unidentifiable. The singular values and the unidentifiable parameters are graphically illustrated in a so-called identifiability signature [22].

To illustrate our approach, we use the NF-κB model, also analysed in Section 3. It has 15 states and 28 model parameters and if all the initial conditions of the individual model states are considered to be unknown, it has a total of 43 parameters [3]. Measuring $y_{max} = \{x_1, \ldots, x_{15}\}$ as model output, we observe no gap in the singular values (See Fig 1). This confirms that there are no vanishing singular values and therefore the sensitivity matrix, S_{norm}, is of full rank and the model is structurally identifiable for this particular choice of output sensors.

However, if we omit state x_4 from the output, y_{max}, we observe from Fig 2 that matrix S_{norm} is now rank deficient. This is apparent from the clear gap in the singular values and the vanishing singular value of $\sigma_{43} = 7.8 \times 10^{-16}$.

We can now examine the columns of V, corresponding to vanishing singular values, for suggestions as to which model parameters may be unidentifiable. Fig 2 reveals only 1 vanishing singular value and therefore it suffices to consider only the last column vector, v_{43},
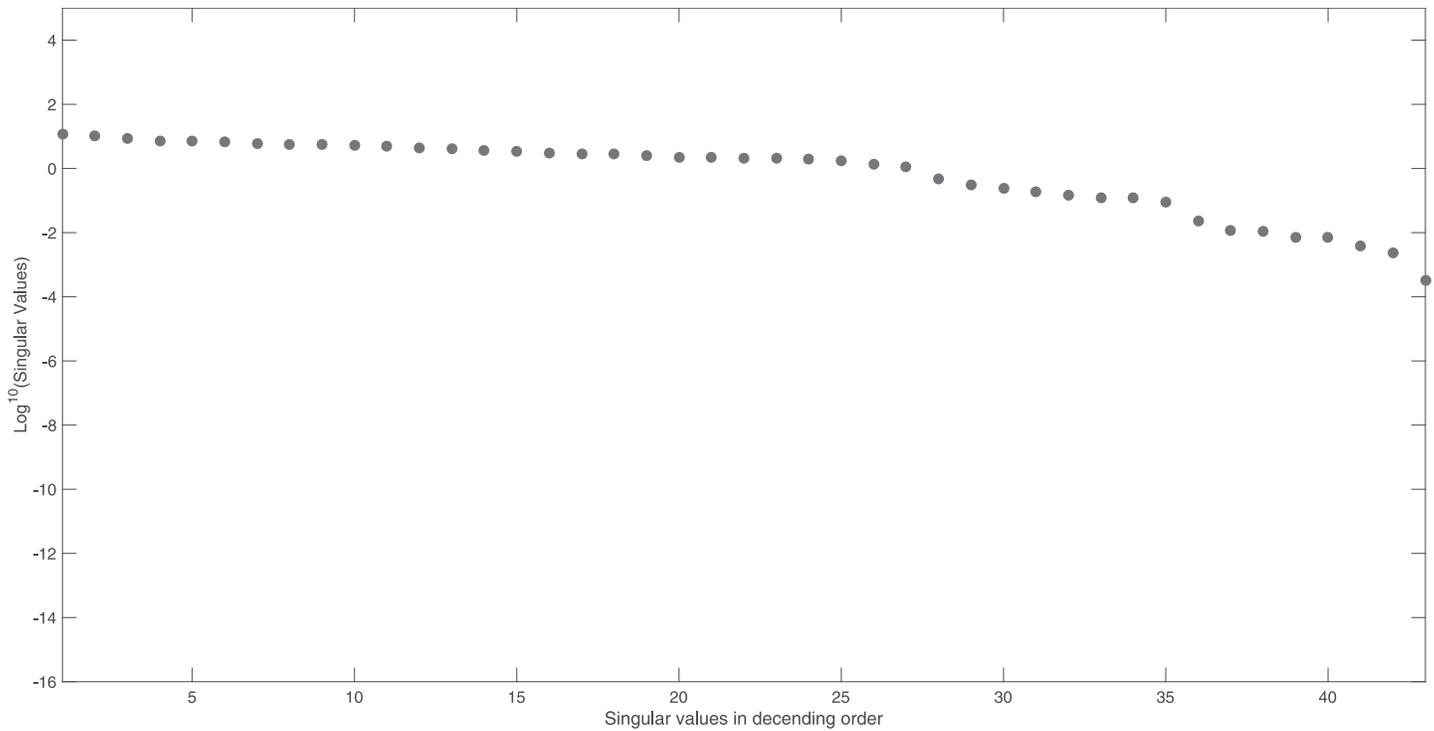
Fig 1. NF-κB model: Singular values of the output sensitivity matrix, S_{norm}, if we measure all states, $\{x_1, \ldots, x_{15}\}$, as model output. Singular values, arranged in descending order, reveal no gap. This suggests that the sensitivity matrix is of full rank and therefore the model is structurally identifiable.
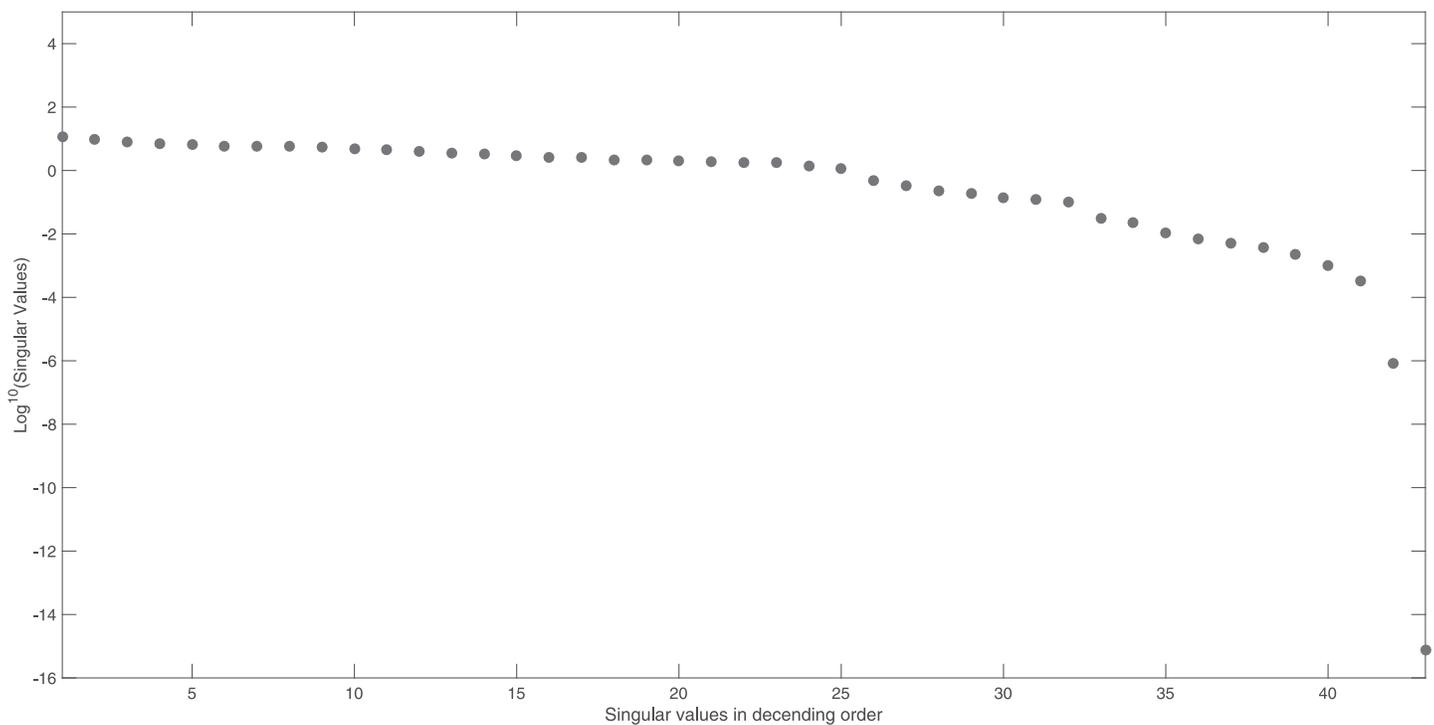
https://doi.org/10.1371/journal.pone.0207334.g001

Fig 2. NF-κB model: Singular values of the output sensitivity matrix, S_{norm}, if we measure all states apart from x_4. Singular values, arranged in descending order, reveal a clear gap with $\sigma_{43} = 7.8 \times 10^{-16}$. This indicates that the sensitivity matrix is rank deficient and so the model is structurally unidentifiable.

https://doi.org/10.1371/journal.pone.0207334.g002
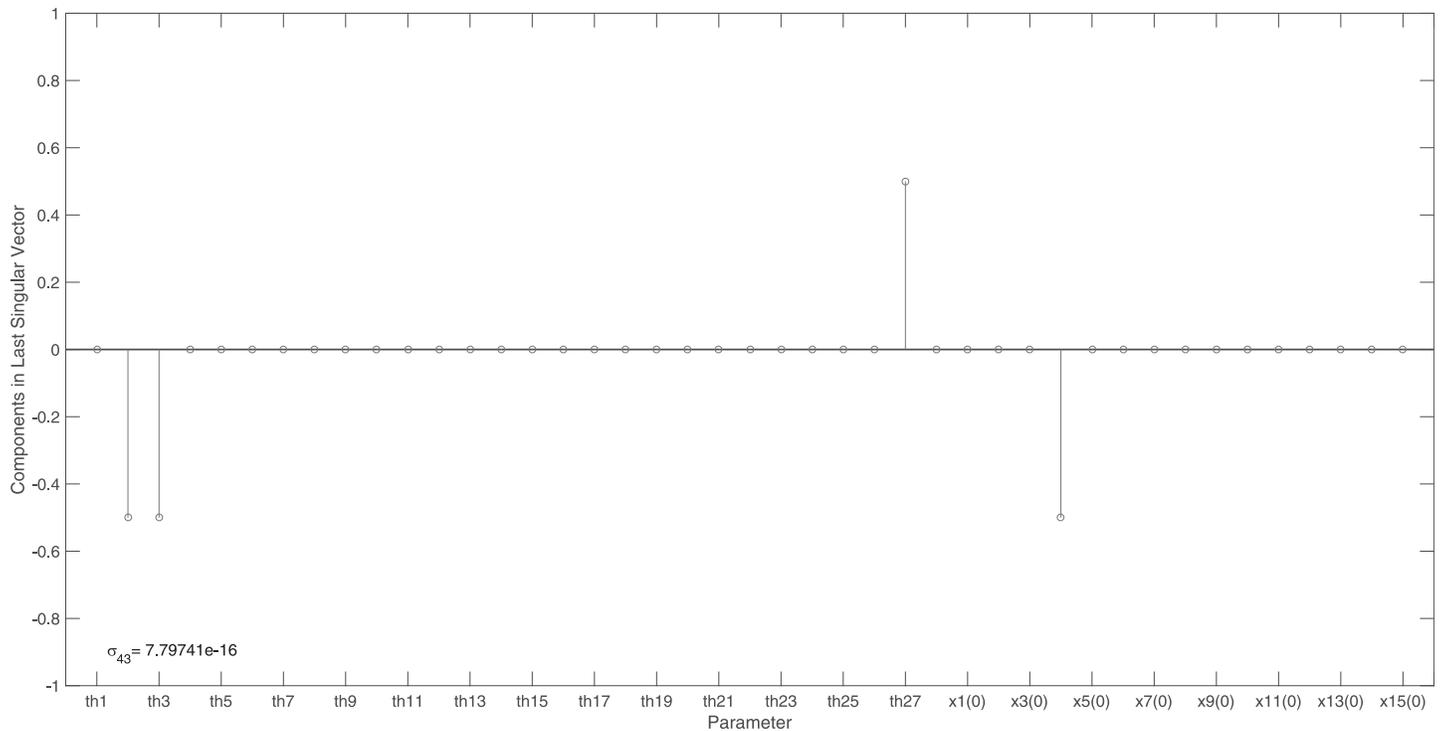
Fig 3. NF-κB model: Entries in the last right singular vector corresponding to the vanishing singular value, σ_{43}, in Fig 2. The corresponding non-trivial null-space indicates that parameters θ_2, θ_3, θ_{27} and initial condition $x_4(0)$ are totally correlated.

https://doi.org/10.1371/journal.pone.0207334.g003

corresponding to σ_{43}. The non-zero entries in Fig 3 reveal that parameters θ_2, θ_3, θ_{27} and the initial condition $x_4(0)$, are both totally correlated and unidentifiable. To ensure the model's structural identifiability, the omitted state, x_4, has to be measured and so is included into any minimal output set. In contrast, omitting state x_3 from the output set does not change this model's identifiability and therefore can be omitted from a minimal output set.

Minimal output set algorithm

Here, we present our algorithm to detect minimal output sets. We first outline the ideas underlying the algorithm and then discuss the subsequent steps. It is important to realise that the parameters to be identified may comprise both system parameters θ_j, $j = 1, .., p$, and initial values of the states, $x_j(0)$, $j = 1, . . ., n$. We assume that the numerical values assigned to the elements in both $\boldsymbol{\theta}$ and $\boldsymbol{x}(0)$ are regular points, where it is known that the rank of the sensitivity matrix does not change in the neighbourhood of a regular point. To ensure that this assumption holds, it may be useful to repeat the algorithm for a different values in the vicinity of a chosen regular point. System parameters are to be *inferred* from measurements of model states that may or may not be measured directly and so are usually not regarded as measurable outputs. For the time being, we assume that the pre-defined measurable outputs y_j, $j = 1, . . ., m$, also referred to as sensors, are identical to the states x_j, $j = 1, . . ., n$, and therefore $m = n$. Later on we show that this assumption can easily be relaxed. We may also take for granted that the system is identifiable when all sensors are measured. If this would not be the case, searching for minimal output sets would clearly not be possible.

The main idea of the algorithm is to systematically omit elements from the set of all available sensors, thereby searching for essential sensors that absolutely can not be omitted to keep

the system identifiable. As explained above, unidentifiability is detected by inspecting the calculated singular vales of the sensitivity matrix in (7). If these singular values show a gap of 3 decades or larger, we conclude unidentifiability and subsequently proceed to identify the essential sensors that need to be included into a model's minimal output sets.

Let $\boldsymbol{y}_{max}$ be the set of all available sensors with set-cardinality $|\boldsymbol{y}_{max}| = m$. The algorithm involves an iterative identifiability analyses in which sensors are omitted step-wise from the maximum starting set $\boldsymbol{y}_{max}$. Systematically more and more sensors are left out as to find all essential sensors that are needed for a minimal output set (MOS).

Let k be the number of sensors to be omitted from a set of available sensors, $\boldsymbol{y}^k$. Starting with $k = 1$, we leave out one-sensor-at-a-time from the initial set of *all available sensors*, $\boldsymbol{y}^1 := \boldsymbol{y}_{max}$. Each time measuring with a different set of sensors from $\boldsymbol{y}^1$, we conduct $\binom{m}{m-1} = m$ identifiability analyses for $k = 1$. If a lack of identifiability is detected, the unidentifiable parameters are stored in a set ϕ_1 and the corresponding omitted sensors that cause unidentifiability are stored in a set ψ_1. Continuing this way, we get unidentifiable parameter sets $\{\phi_i, i = 1, \ldots, l_1\}$, and the corresponding omitted sensor sets $\{\psi_i, i = 1, \ldots, l_1\}$ that cause a lack of identifiability. Here, l_1 is the total number of unidentifiable parameter sets identified for the case of omitting one-sensor-at-a-time ($k = 1$). The unidentifiable parameter sets ϕ_i can be found by inspecting the non-zero entries in the singular vectors of the matrix $\boldsymbol{V}$ corresponding with the zero-valued singular values (as can be seen from the identifiability signature).

To ensure structural identifiability, the essential sensors form each $\{\psi_i, i = 1, \ldots, l_1\}$ *must be* included into *any* minimal output set. Having checked all possibilities of leaving out one-sensor-at-a-time, we can now define a new set of available sensors, say $\boldsymbol{y}^2$, that is created by excluding the previously found sensors in the sets $\{\psi_i, i = 1, \ldots, l_1\}$ from the set $\boldsymbol{y}^1$. Since we know for sure that these excluded sensors are needed for a model's structural identifiability, they are permanently included into all sensor sets that are measured from now on. Hence, the case $k = 1$ *reduces* the number of candidate sensors to choose from in the next iteration from m to $m' = m - l_1$.

Next, we leave out 2-sensors-at-a-time (the case $k = 2$) from $\boldsymbol{y}^2$ and check for identifiability. Since set cardinality $|\boldsymbol{y}_2|$ now equals $m' \leq m$, we have $\binom{m'}{m'-2}$ choices for omitting 2 sensors from this set. If unidentifiability is detected, a new set of unidentifiable parameters is compiled from the identifiability signature and stored in $\phi_{l_1 + 1}$, and the 2 corresponding left-out sensors are stored in $\psi_{l_1 + 1}$. Proceeding this way, the total number of unidentifiable sets that can be found for $k = 2$ are collected in the sets $\{\phi_i, i = l_1 + 1, \ldots, l_1 + l_2\}$ and the corresponding omitted sensor sets, $\{\psi_i, i = l_1 + 1, \ldots, l_1 + l_2\}$.

Assume now that for the case of leaving out 2-sensors-at-a-time ($k = 2$), we have found an unidentifiable parameter set ϕ_i. Apparently, this new set, ϕ_i, *only occurs when these 2 particular sensors are missing* and therefore, either 1 of these 2 essential sensors *must be* included in a MOS. Hence, the available sensor sets for the case $k = 3$ branch out into two sets, namely $\boldsymbol{y}^{3,1}$ and $\boldsymbol{y}^{3,2}$. When leaving out three-sensors-at-a-time in the next iteration of our algorithm (case $k = 3$), we have to iterate both of these available sensor sets to find more unidentifiable parameter sets $\{\phi_i, i = l_1 + l_2 + 1, \ldots, l_1 + l_2 + l_3\}$. Continuing in this way for $k = 3, 4, \ldots$, we complete our search for essential sensors when leaving out k sensors at a time. At the same time guaranteeing that the sensors that are needed for the identifiability of our model, identified in earlier iterations $k - 1, k - 2, \ldots, 1$, are included in each new measured output.

Clearly, for larger models the output, $\boldsymbol{y}_{max}$, will contain a large number of sensors and in these cases an exhaustive search will be computationally demanding. The computational burden may however be substantially reduced by randomly selecting outputs from an intermediate set of available sensors $\boldsymbol{y}^k$ (for a certain iteration step k) using a series of Bernoulli trial

experiments. The number of sensors to include into each sensor set can then be chosen in such a way that the chance of successfully detecting an unidentifiable set of parameters is more than 99.5% (refer to supplementary S9 File).

We further note that in practice our experience shows that the values $k = 1, 2, 3$ already summarise the *majority* of possible unidentifiable parameter sets ϕ_i. More importantly, once we have established a few required sensors on basis of lower k values, one can perform an additional check for a lack of identifiability when *using only the required sensors that have already been determined for the lower k values*. Such a test will immediately reveal additional correlations that still need to be found for larger k values, but these correlations are not yet neatly separated in a systematic way. This check does, however, demonstrate decisively whether we need to continue our search for larger k values (e.g. $k = 4, 5, \ldots$), yes or no or whether one can already define minimal outputs sets from the already identified essential sensors.

Finally, in reality the output, y_{max}, is not always identical to the states x. For example, one could have $y_{max} = \{x_1 + x_3 + x_4, \theta_{16}(x_3 + x_4 + x_5 + x_{12}), \theta_{17}(x_4 + x_5)\}$. Our algorithm allows for the user to define these more complex outputs in a straightforward manner: Instead of omitting states $\{x_i, i = 1, \ldots, n\}$, we now systematically omit *outputs* $\{y_j, j = 1, \ldots, m\}$ to find the essential sensors needed in a MOS.

Results and discussion

Example 1: A chemical reaction system

This model was used by Liu and co-authors to illustrate their method ensuring observability based on the graphical analysis of a model's structure [5]. It contains 11 states and 6 model parameters and potentially has 17 unknown parameters. Examining the structure of the model by evaluating its adjacency/Jacobi matrix, the authors detected 3 root strongly connected components and identified 6 minimal output sets.

These observability results were confirmed using our algorithm. Additionally, we expanded the scope of the problem to define minimal output sets that guarantee this model's structural identifiability. We found that the minimal output sets that ensure observability also ensure identifiability and these are: $\{x_4, x_6, x_7\}$, $\{x_4, x_6, x_8\}$, $\{x_4, x_6, x_9\}$, $\{x_5, x_6, x_7\}$, $\{x_5, x_6, x_8\}$ and $\{x_5, x_6, x_9\}$. These results were obtained in 6 minutes and 35 seconds using a Intel Core i7 processor with 8GB RAM (see S1 File for details).

Using our algorithm, we detected 3 different sets of unidentifiable parameters, $\{\phi_1, \phi_2, \phi_3\}$. Each of these sets can be verified symbolically, which also allows for the identification of different totally correlated sets of parameters within each set, ϕ_i (see supplementary S8 File for the symbolic verification of all 3 unidentifiable sets). The results obtained for the different values of k are summarised in Table 1.

Table 1. Results obtained analysing the chemical reaction model.

k value	Number of sets	Unidentifiable parameters sets	Omitted sensors	Computational time (seconds)
1	1	$\phi_1 = \{x_6(0)\}$	$\psi_1 = \{x_6\}$	4.5
2	1	$\phi_2 = \{\theta_2, \theta_3, x_4(0), x_5(0)\}$	$\psi_2 = \{x_4, x_5\}$	14.6
3	1	$\phi_3 = \{\theta_4, \theta_5, x_7(0), x_8(0), x_9(0)\}$	$\psi_3 = \{x_7, x_8, x_9\}$	53.6
4	0			53.5
5	0			112.3
6	0			90.4
7	0			48.7
8	0			17.1
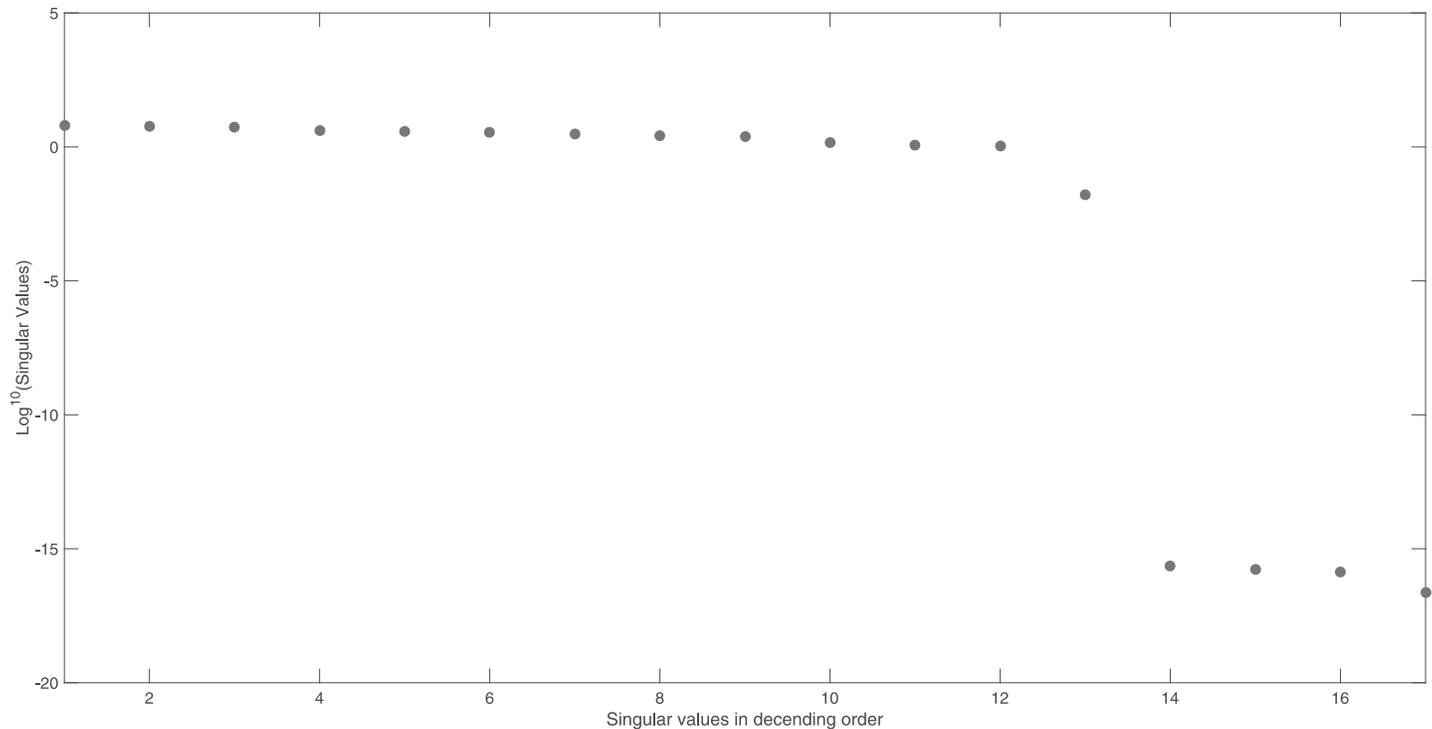
Fig 4. Example 1: Structural identifiability results of a chemical reaction system: Singular values of the output sensitivity matrix, S_{norm}, when measuring the output $\{x_1 \ldots, x_{11}\}$ omitting sensors x_4 and x_5. Singular values, arranged in descending order, reveal a clear gap. This gap, in conjunction with the smallest singular value, $\sigma_{17} = 2.4 \times 10^{-17}$, indicate that the model is structurally unidentifiable when measuring this output.

Figs 4 and 5 indicate the identifiability signature obtained when measuring the output, $\{x_1, x_2, x_3, x_6, x_7, x_8, x_9, x_{10}, x_{11}\}$, here $k = 2$. The 4 zero-valued singular values indicate that the model is unidentifiable when measuring this output. The unidentifiable parameters can be identified by looking at the non-zero entries in the last 4 columns of matrix V, each corresponding to a singular value beyond the gap. Fig 5 reveals the unidentifiable parameter set, $\phi_2 = \{\theta_2, \theta_3, x_4(0), x_5(0)\}$ and accordingly, the essential sensors are $\psi_2 = \{x_4, x_5\}$. The symbolic verification of this set yields a non-trivial null-space with 4 base vectors: $\mathcal{N}(\boldsymbol{J}(\tilde{\boldsymbol{x}}_0^{cor})) = \{1, 0, 0, 0\}, \{0, 1, 0, 0\}, \{0, 0, 1, 0\}, \{0, 0, 0, 1\}$, where $\tilde{\boldsymbol{x}}_0^{cor} = \{\theta_2, \theta_3, x_4(0), x_5(0)\}$.

Example 2: NF-κB model

This model describes the two-feedback-loop regulatory module of nuclear factor NF-κB signalling pathway. It involves two-compartment kinetics of the activators IκB (IKK) and NF-κB, the inhibitors, A20 and IκBα, and their complexes. In response to extra-cellular signals such as tumour necrosis factor, the activation of IKK ultimately stimulates the release of the main activator NF-κB, which enters the nucleus and triggers transcription of the inhibitors and numerous other genes [23] (See supplementary S2 File for a model description). The model contains 15 states and 28 model parameters and assuming the initial state conditions to be unknown, it has 43 unknown parameters in total.

Minimal output sets for this model were first identified by Anguelova and co-authors [3]. We found the model structural identifiable when measuring all states, $\boldsymbol{y_{max}} = \{x_1, \ldots, x_{15}\}$. Our algorithm identified 5 different sets of unidentifiable parameters: $\phi_1 = \{\theta_2, \theta_3, \theta_{27}, x_4(0)\}$,
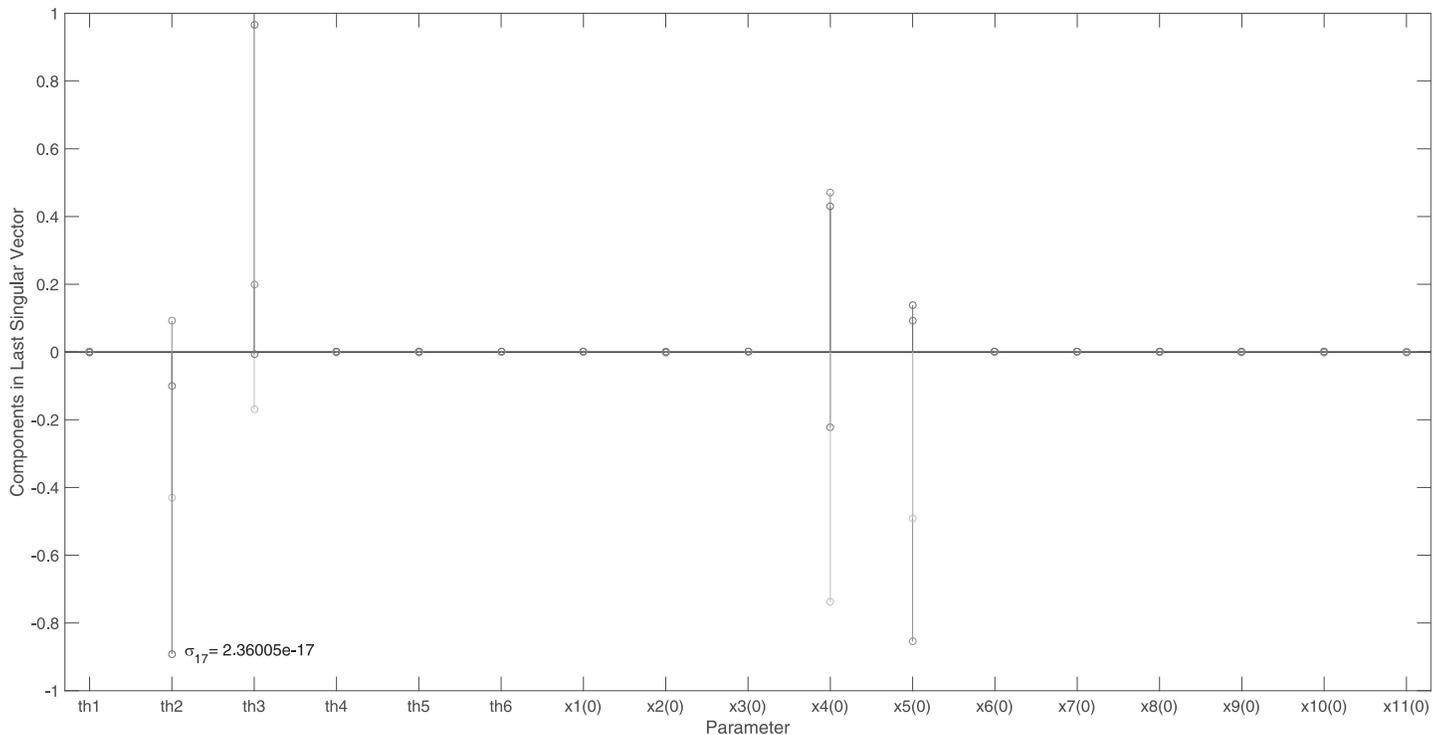
Fig 5. Example 1: Structural identifiability results of a chemical reaction system: Non-zero entries in the last 4 columns of matrix V. These indicate that initial conditions $x_4(0)$ and $x_5(0)$ and model parameters θ_2 and θ_3 are unidentifiable. Since x_4 and x_5 are defined in y_{max}, both of these sensors are essential.

https://doi.org/10.1371/journal.pone.0207334.g005

$\phi_2 = \{\theta_5, \theta_6, \theta_{18}, x_5(0)\}$, $\phi_3 = \{\theta_8, \theta_9, \theta_{10}, x_6(0)\}$, $\phi_4 = \{\theta_{19}, \theta_{27}, x_{10}(0)\}$ and $\phi_5 = \{x_{12}(0)\}$. The corresponding sets of essential sensors are: $\psi_1 = \{x_4\}$, $\psi_2 = \{x_5\}$, $\psi_3 = \{x_6\}$, $\psi_4 = \{x_{10}\}$ and $\psi_5 = \{x_{12}\}$ and these results were obtained in 29.5 seconds. Analysing the model for all the different values of k took 8 minutes and 20 seconds. The resulting minimal output set, $\{x_4, x_5, x_6, x_{10}, x_{12}\}$, is identical to the minimal output set defined by Anguelova and co-authors [3].

Figs 2 and 3 show the identifiability signature obtained when sensor x_4 is omitted from y_{max}. The symbolic verification of the unidentifiable set shown in Fig 3 yields the non-trivial null-space: $\mathcal{N}(J(\tilde{x}_0^{cor})) = \{\theta_2/x_4(0), -\theta_3/x_4(0), -\theta_{27}/x_4(0), 1\}$, were $\tilde{x}_0^{cor} = \{\theta_2, \theta_3, \theta_{27}, x_4(0)\}$. Refer to the supplementary S8 File for symbolic verification of the remaining 4 sets of unidentifiable parameters.

Example 3: JAK/STAT model

This model aims to describe the interaction of the suppressor cytokine signaling-1 (SOCS1), Janus kinase (JAK) and the transcription (STAT) signal transduction pathway [24] (S3 File). It contains 31 model states and 51 model parameters and therefore the total number of unknown parameters is 82. This model was structurally identifiable when measuring all states, $y_{max} = \{x_1, \ldots, x_{31}\}$. Applying our method, we identified 2 sets of unidentifiable parameters: $\phi_1 = \{x_{31}(0)\}$ with corresponding omitted sensor set $\psi_1 = \{x_{31}\}$, and $\phi_2 = \{\theta_{14}, \theta_{51}, x_{10}(0), x_{11}(0)\}$ with corresponding omitted sensor set $\psi_2 = \{x_{10}, x_{11}\}$. These results were obtained in 3 minutes and 2 seconds with sets ψ_1 and ψ_2 identified using iterative Bernoulli trails.

Due to the model size and the potential computational demand associated with larger values of k, we first measured outputs containing the already detected essential sensors to ascertain whether measuring these sensors resulted in the model's structural identifiability. The model
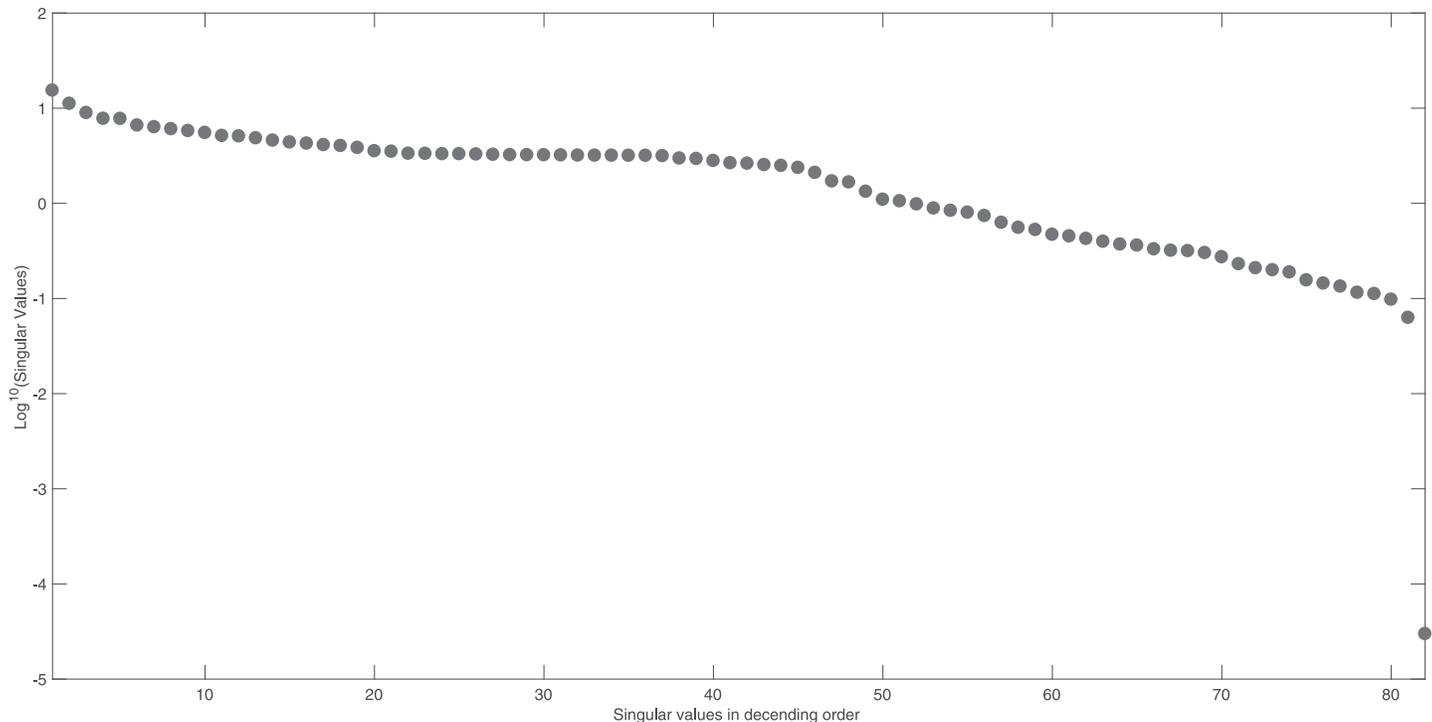
Fig 6. Example 3: JAK-STAT model: Singular values of the output sensitivity matrix, S_{norm}, when measuring the model output $\{x_1, \ldots, x_{31}\}$, omitting sensors x_{10} and x_{11}. Singular values reveal a clear gap and this, in conjunction with the smallest singular value of $\sigma_{82} = 7.3 \times 10^{-16}$, indicates that S_{norm} is not of full rank and therefore the model is structurally unidentifiable.

was found to be identifiable when measuring either of the outputs, $\{x_{10}, x_{31}\}$ or $\{x_{11}, x_{31}\}$. Accordingly, these are the minimal output sets of the JAK/STAT model and our results correspond to the findings of Anguelova and co-authors [3].

The identifiability signature obtained when states x_{10} and x_{11} are simultaneously omitted from the model's output is illustrated in Figs 6 and 7. The unidentifiable set illustrated in Fig 7, was confirmed by the symbolically computed non-trivial null-space: $\mathcal{N}(J(\tilde{x}_0 \, cor)) = \{\theta_{14}/x_{11}(0), -\theta_{51}/x_{11}(0), \theta_{10}/x_{11}(0), 1\}$. Here $\tilde{x}_0^{cor}$ is set ϕ_2.

Example 4: Ligand binding model

Next, we consider a Ligand binding model, previously analysed for structural identifiability [25]. This model describes the dynamic behaviour of the ligand (Epo) and its receptor (EpoR) in erythroid progenitor cells. In these cells, the dynamic characteristics of the Epo receptor (EpoR) determine how signals are encoded, in the presence of Epo, and processed at receptor level. These processed signals activate downstream signalling cascades such as the JAK2-STAT5 pathway which in turn leads to responses such as differentiation and proliferation of erythrocytes [25]. The model consists of 6 states and assuming their initial states are unknown, it contains 14 unknown parameters (see supplementary S4 File).

The minimal output set ensuring the observability of this model, $\{x_5, x_6\}$, was determined by Liu and co-authors using their graphical approach [5]. This set also ensures the structural identifiability of the model and this result was obtained in 12 seconds. Two sets of unidentifiable parameters were detected: $\phi_1 = \{x_5(0)\}$ and $\phi_2 = \{x_6(0)\}$. Set ϕ_2, shown in Fig 8, is indicated by the non-zero entry in the last right singular vector corresponding to the smallest singular value calculated to be precisely zero.
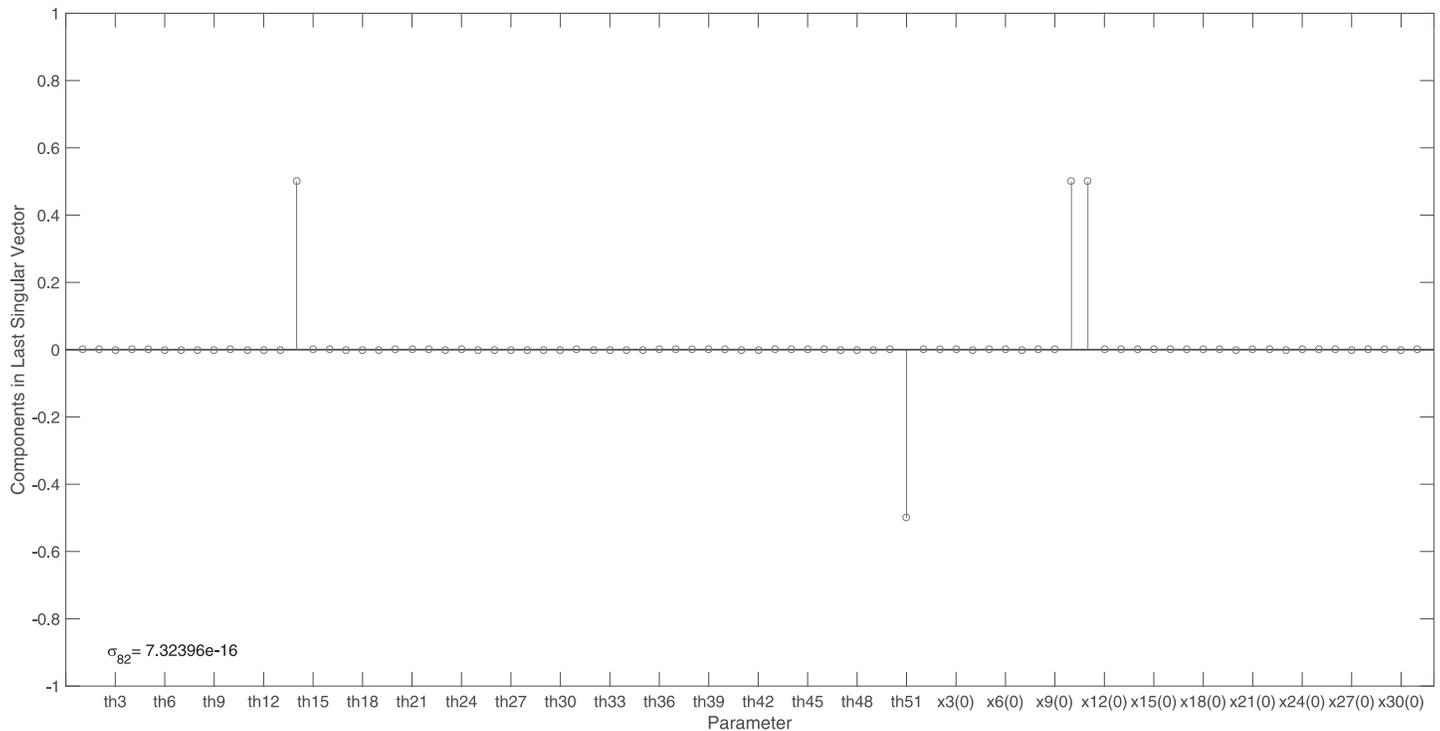
Fig 7. Example 3: JAK-STAT model: Entries in the last right singular vector corresponding to the vanishing singular value, σ_{82}, in Fig 6. The corresponding non-trivial null-space indicates that model parameters θ_{14}, θ_{51} and initial conditions $x_{10}(0)$ and $x_{11}(0)$ are totally correlated and so the model is not identifiable when model states x_{10} and x_{11} are simultaneously omitted from the model's output.

https://doi.org/10.1371/journal.pone.0207334.g007

Example 5: Simplified glycolytic reaction model

The simplified glycolytic reaction map consists of 10 chemical species: glucose, ADP, glucose 6-phosphate, ATP, glucose 1-phosphate, AMP, fructose 6-phosphate, fructose 2, 6-bipho-sphate, triose phosphate and pyruvate. The interaction between these chemicals are described by 9 reactions [26] (see supplementary S5 File). This model's minimal output set for observability was defined by Liu and co-authors as $\{x_{10}\}$ [5]. Our algorithm confirmed that this minimal set also ensures the model's structural identifiability. This result was obtained after 2 minutes and 43 seconds. The set of unidentifiable parameters, $\phi_1 = \{\theta_{13}, x_{10}(0)\}$, corresponding with the omitted sensor set, $\psi_1 = \{x_{10}\}$, is indicated in Fig 9.

Example 6: Goldbeter model

Consider a model describing the circadian oscillations in the Drosophila period protein (PER) [27]. It is based on both multiple phosphorylation of PER and on the negative feedback exerted by PER on the transcription of the period (per) gene. It provides a molecular basis for circadian oscillations of the limit cycle type in which the peak in per mRNA precedes the peak in total PER protein.

This model was analysed by Sedoglavic in 1995, in which he identified only 1 set of totally correlated parameters [28]. It contains 5 states and 17 model parameters and assuming that initial conditions are unknown, the total number of model parameters is 22. Measuring the output, $\{x_2, x_3, x_4, x_5\}$, our algorithm also found only the 1 totally correlated set, $\phi_1 = \{\theta_1, \theta_3, \theta_4, \theta_5, x_1(0)\}$, with its elements indicated by the non-zero values in Fig 10. The minimal output set of this model, $\{x_1\}$, was calculated in 12 seconds.
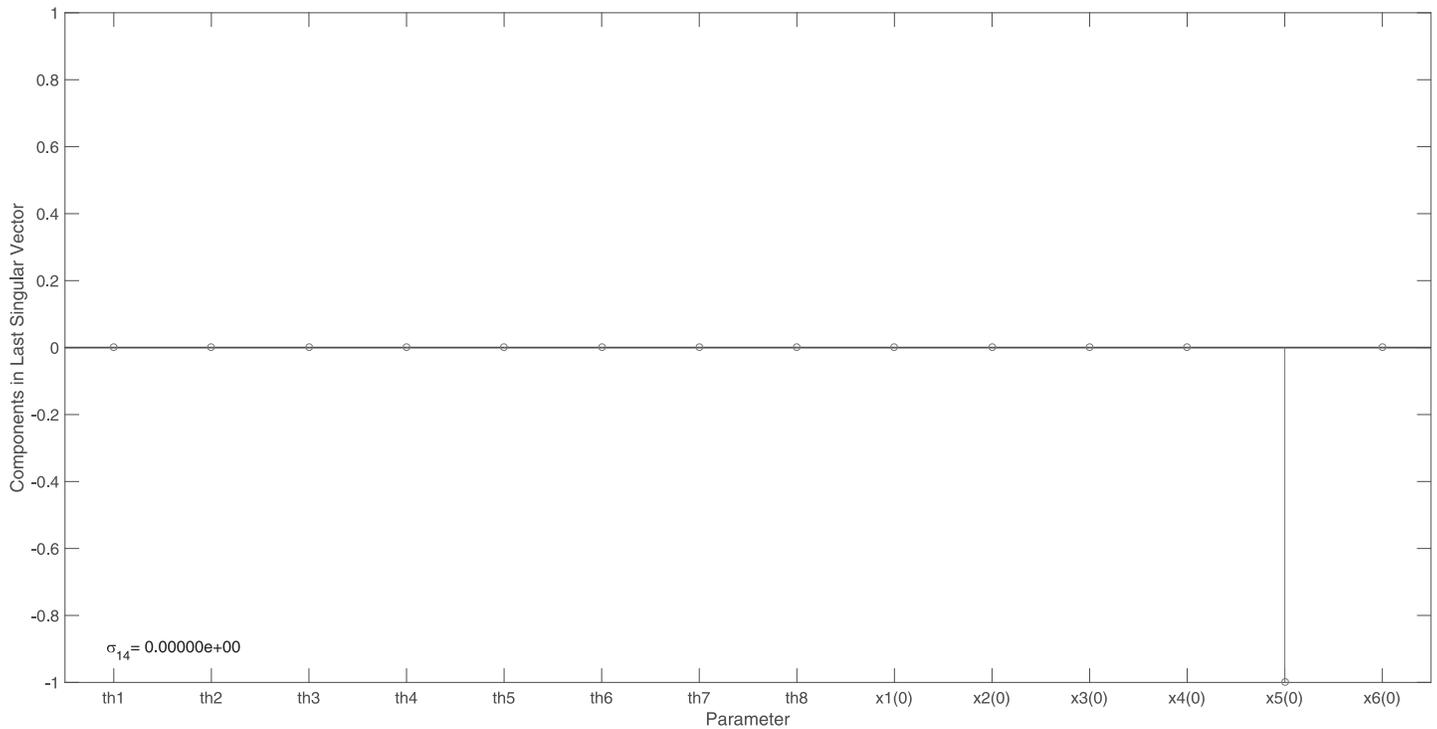
Fig 8. Example 4: Ligand binding model: Entries in the last right singular vector, corresponding to the smallest singular value of precisely zero, calculated for the measured output $\{x_1, x_2, x_3, x_4, x_6\}$. The non-trivial null-space indicates that the initial condition of state x_5 is unidentifiable when this state is not measured. Accordingly, x_5 should be included into the model's minimal output set.

https://doi.org/10.1371/journal.pone.0207334.g008
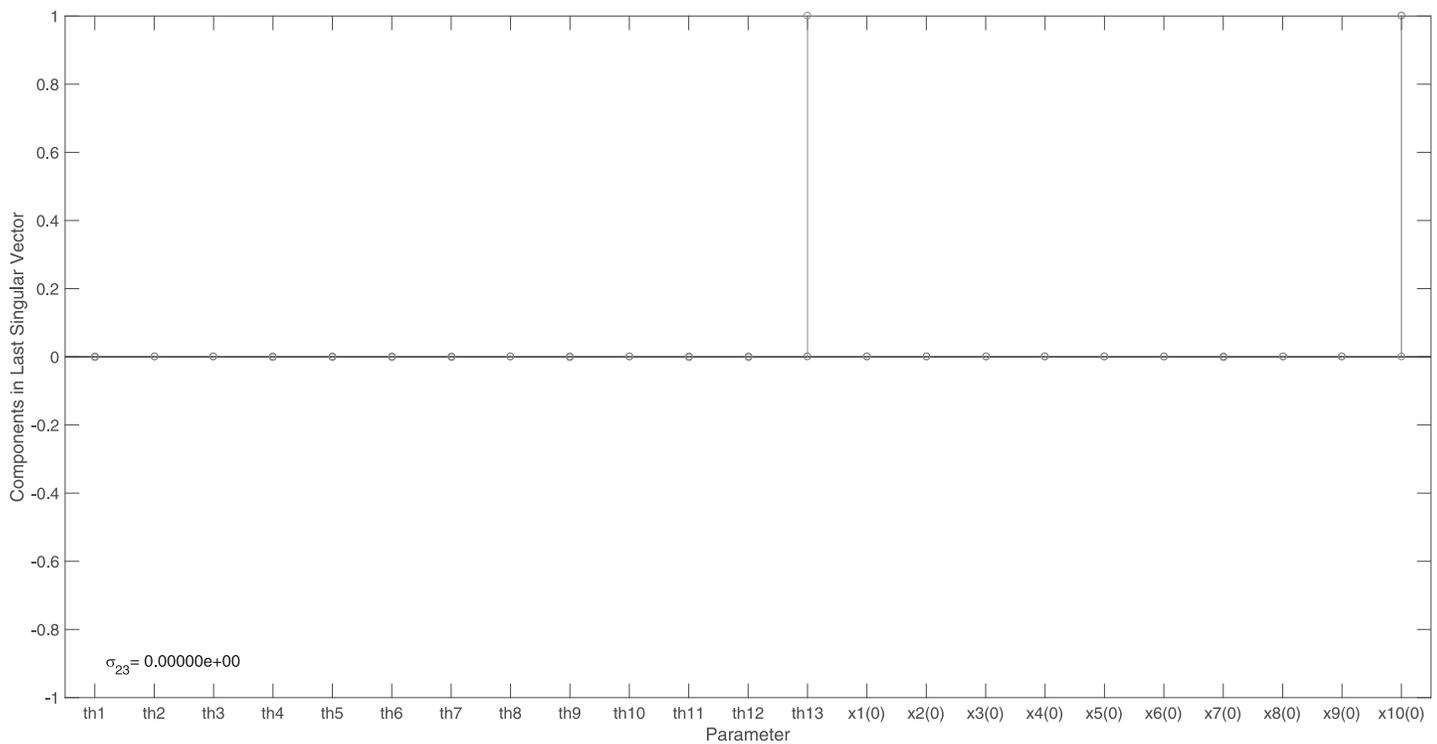
Fig 9. Example 5: Simplified glycolytic reaction model: Entries in the right singular vectors corresponding to 2 vanishing singular values. The non-zero values indicate that the initial condition $x_{10}(0)$ and parameter θ_{13} are unidentifiable when state x_{10} is not measured.

https://doi.org/10.1371/journal.pone.0207334.g009
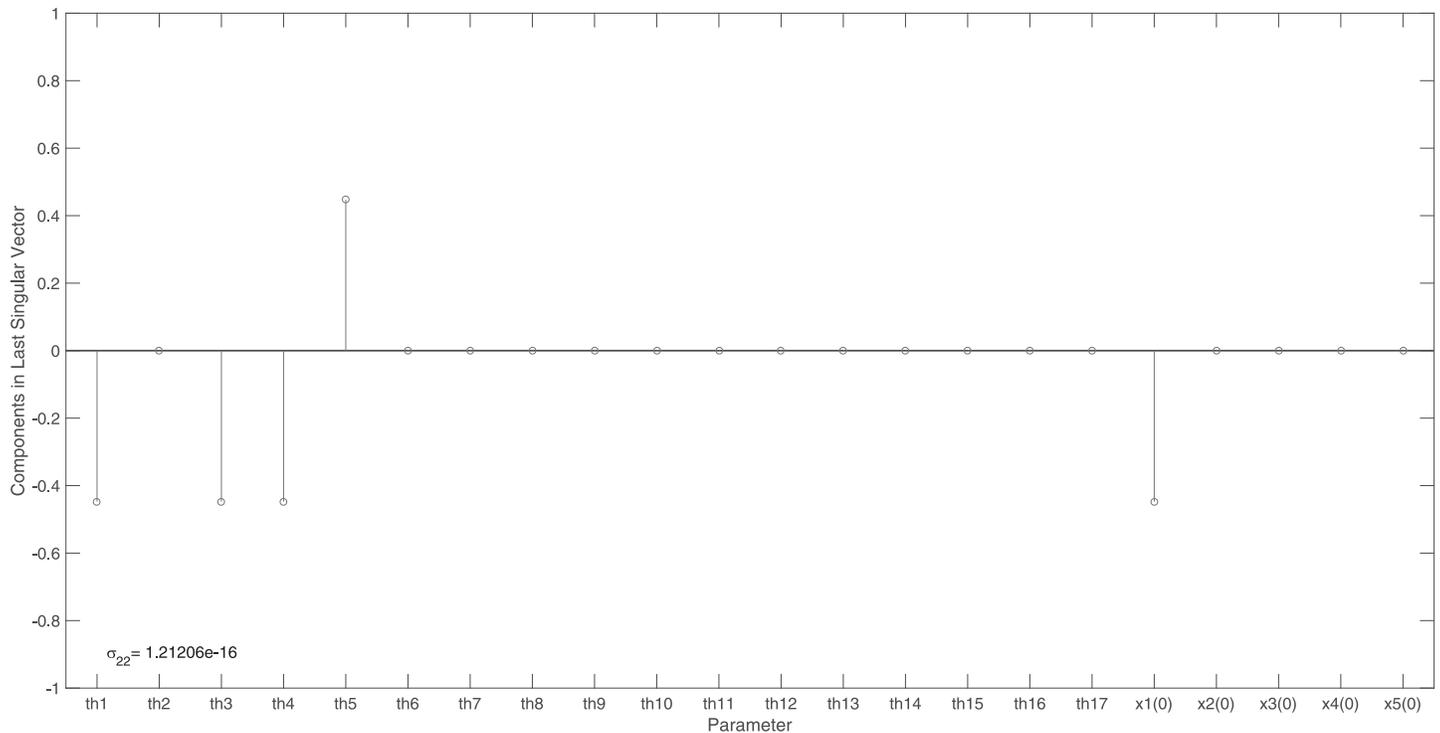
Fig 10. Example 6: Goldbeter model: Entries in the last right singular vector corresponding to single vanishing singular value calculated. The non-zero values indicate that parameters θ_1, θ_3, θ_4, θ_5 and initial condition $x_1(0)$ are unidentifiable when state x_1 is not measured.

Example 7: JAK/STAT model with specific model output

In this example, we illustrate how our method can be used to identify minimal output sets from a set of more complex model outputs. These outputs do not simply consist of single model states and in this example, also include additional model parameters. We consider a reparameterised JAK/STAT model, with the original unidentifiable model described by Raia and co-authors [29]. The constitutive activation of the JAK (Janus kinase)/STAT signalling pathway forms part of both the primary mediastinal B-cell lymphoma (PMBL) and the classical Hodgkin lymphoma (cHL). Raue and co-authors investigated the identifiability of this benchmark model using three different approaches [1].

The model definition also contains a specific set of initial conditions for model states, $x(0) = \{1.3, \theta_{21}, 0, 1, 0, 2.8, 0, 165, 0, 0, 0.34, 0, 0, 0\}$. These initial conditions, in conjunction with the predetermined set of model outputs, result in the model's structural unidentifiability. Structural identifiability can be reinstated by reparameterising the model (See supplementary S7 File for the structurally identifiable version of this JAK/STAT model). The reparameterised model contains 14 states and 21 parameters, with only the initial condition of state x_2 assumed to be unknown.

Considering the reparameterised model's output, $y_{max} = \{x_1 + x_3 + x_4, \theta_{16}(x_3 + x_4 + x_5 + x_{12}), \theta_{17}(x_4 + x_5), \theta_{18} x_7, \theta_{19} x_{10}, \theta_{20} x_{14}, x_{13}, x_9\}$, our algorithm can now be implemented to determine the model's minimal output sets. Setting $k = 1$, already revealed 6 essential sensors. The unidentifiable parameters obtained were: $\phi_1 = \{\theta_{12}, \theta_{16}\}$, when sensor $\psi_1 = \{\theta_{16}(x_3 + x_4 + x_5 + x_{12})\}$ was not measured, $\phi_2 = \{\theta_{17}\}$, when $\psi_2 = \{\theta_{17}(x_4 + x_5)\}$ was not measured, $\phi_3 = \{\theta_{18}\}$, when $\psi_3 = \{\theta_{18} x_7\}$ was not measured, $\phi_4 = \{\theta_{19}\}$, when $\psi_4 = \{\theta_{19} x_{10}\}$ was not measured, $\phi_5 = \{\theta_{20}\}$, when $\psi_5 = \{\theta_{20} x_{14}\}$ was not measured, and $\phi_6 = \{\theta_8, \theta_{13}\}$ when state $\psi_6 = \{x_{13}\}$ was
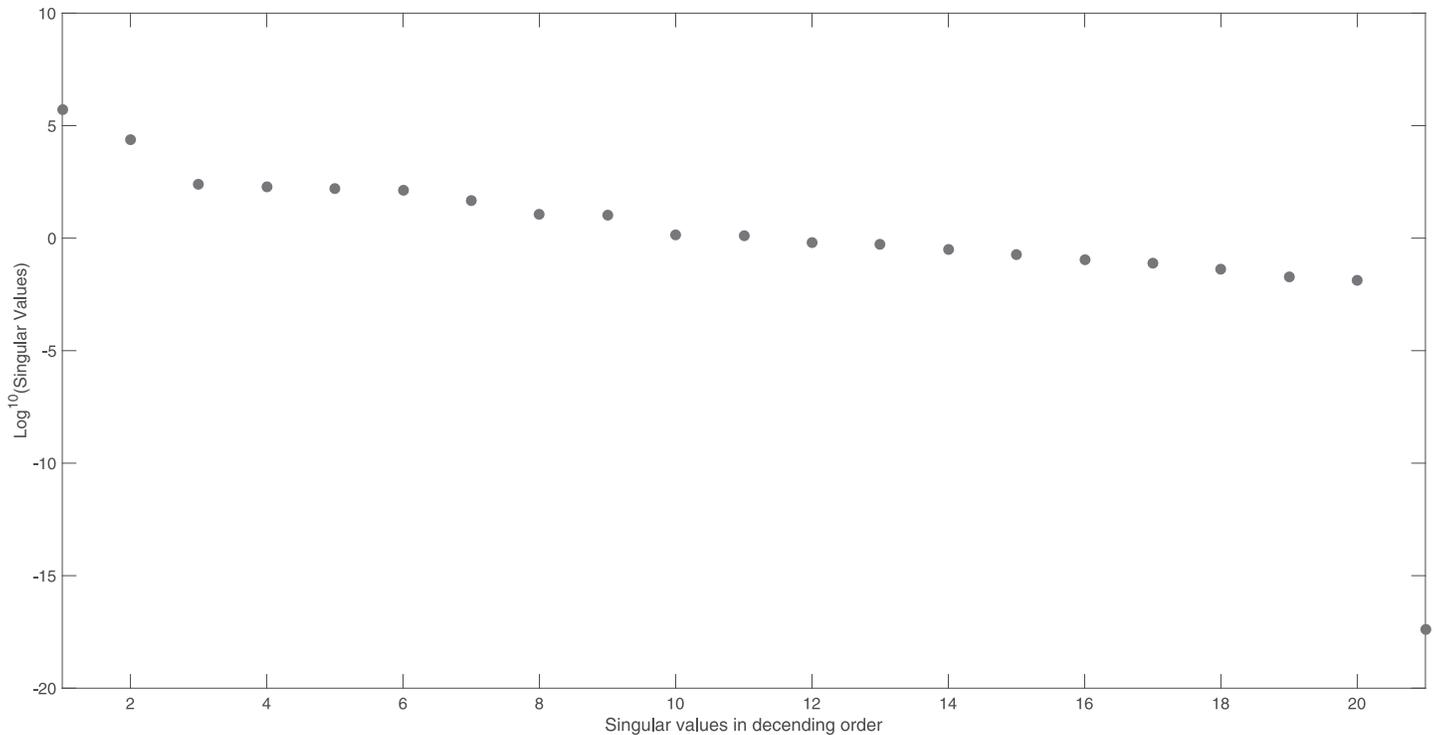
Fig 11. Example 7: JAK/STAT model with specific model output: Singular values of the output sensitivity matrix, S, when omitting sensor $\theta_{17}(x_4 + x_5)$ from y_{max}. Singular values, arranged in descending order, reveal a clear gap. This gap in conjunction with the smallest singular value of 4×10^{-18}, indicate that S is rank deficient.

https://doi.org/10.1371/journal.pone.0207334.g011

not measured. All these sensors are essential and the resulting minimal output set, obtained after 18 seconds, is: $\{\theta_{16}(x_3 + x_4 + x_5 + x_{12}), \theta_{17}(x_4 + x_5), \theta_{18} x_7, \theta_{19} x_{10}, \theta_{20} x_{14}, x_{13}\}$.

Figs 11 and 12 reveal the identifiability signature obtained when sensor $\theta_{17}(x_4 + x_5)$ was not measured. From this, one can see that parameter θ_{17} is unidentifiable.

Example 8: Non-rational JAK/STAT model with specific model output

In this final example, we show that our method can be used to analyse non-rational models. Consider a non-rational version of the JAK/STAT model in example 7:

$$\dot{x}_1 = \theta_1 c_1 u_1 x_1 - \theta_2 x_1^{\pi} + \theta_3 x_2, \tag{10}$$

$$\dot{x}_2 = \theta_2 x_1 - \theta_3 x_2, \tag{11}$$

$$\dot{x}_3 = \theta_1 c_1 u_1 x_1 - \theta_4 x_3 x_7, \tag{12}$$

$$\dot{x}_4 = \theta_4 x_3 x_7 - \theta_5 x_4, \tag{13}$$

$$\dot{x}_5 = \theta_5 x_4 - \theta_6 x_5, \tag{14}$$

$$\dot{x}_6 = -\frac{\theta_7 x_3 x_6}{1 + \theta_8 x_{13}} - \frac{\theta_7 x_4 x_6^{\sqrt{2}}}{1 + \theta_8 x_{13}} + c_2 \theta_9 x_7, \tag{15}$$
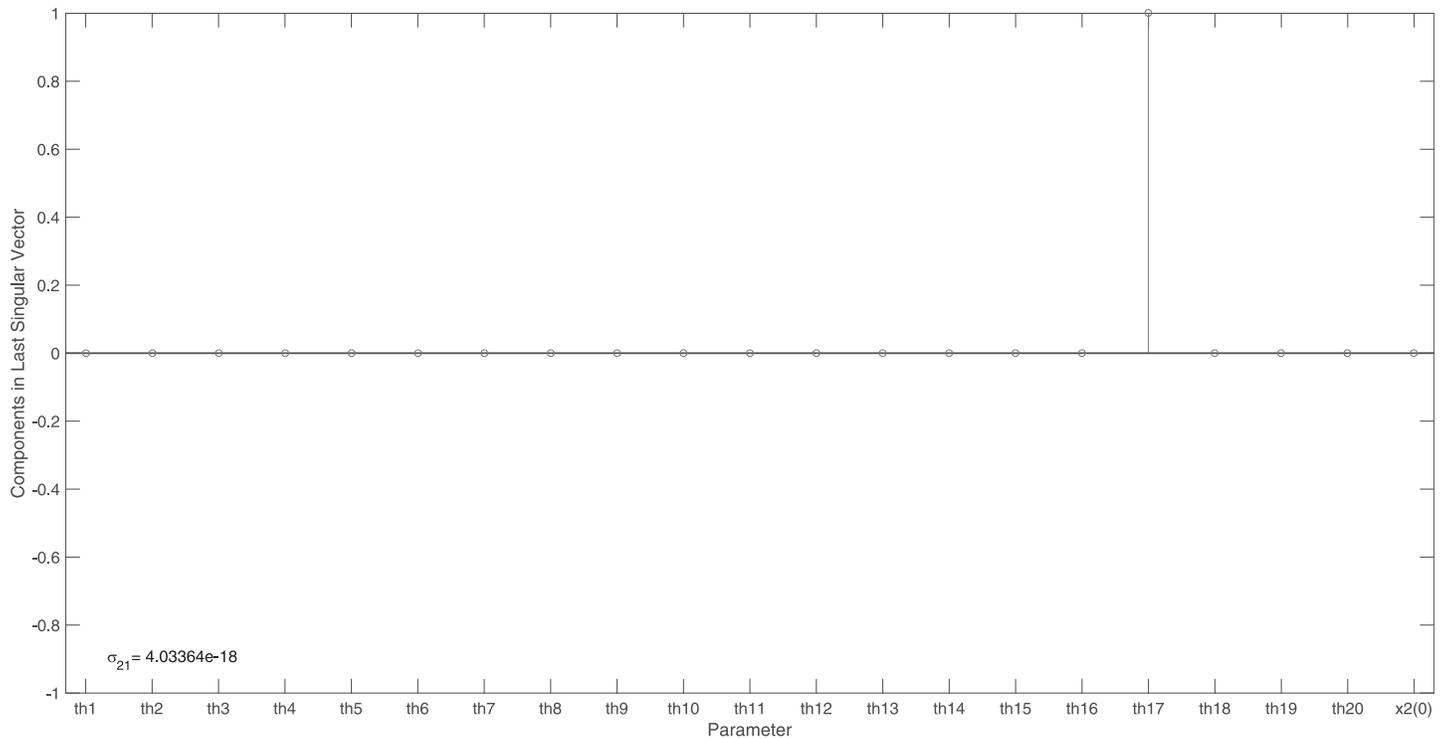
Fig 12. Example 7: JAK/STAT model with specific model output: Entries in the last right singular vector corresponding to the vanishing singular value in Fig 11. The non-trivial null-space indicates that model parameter θ_{17} is not identifiable when sensor $\theta_{17}(x_4 + x_5)$ is not measured.

$$\dot{x}_7 = \frac{\theta_7 x_3 x_6}{1 + \theta_8 x_{13}} + \frac{\theta_7 x_4 x_6^{\sqrt{2}}}{1 + \theta_8 x_{13}} - c_2 \theta_9 x_7, \tag{16}$$

$$\dot{x}_8 = -\theta_{10} x_8 x_7 + c_2 \theta_{11} x_9, \tag{17}$$

$$\dot{x}_9 = \theta_{10} x_8 x_7 - c_2 \theta_{11} x_9, \tag{18}$$

$$\dot{x}_{10} = x_9, \tag{19}$$

$$\dot{x}_{11} = -\theta_{12} c_1 u_1 x_{11}, \tag{20}$$

$$\dot{x}_{12} = \theta_{12} c_1 u_1 x_{11}, \tag{21}$$

$$\dot{x}_{13} = \frac{\theta_{13} \sin(x_{10})}{\theta_{14} + x_{10}} - \theta_{15} x_{13}, \tag{22}$$

$$\dot{x}_{14} = x_9. \tag{23}$$

Analysing this model, we find the results identical to those obtained in example 7 and therefore conclude that the predefined outputs, $x_1 + x_3 + x_4$ and x_9, do not have to be measured to ensure this model's identifiability.

The different identifiability signatures, calculated for each example, can be found in the supplementary material (see supplementary files S1 to S7 Files). The symbolic verification of the individual unidentifiable sets in ϕ can be found in the supplementary S8 File. The MATLAB code of our algorithm can be found at: https://sourceforge.net/u/djoubert-wur/profile.

Conclusions

In this paper we introduced an algorithm that can find minimal output sets for a wide range of models in a short time. It is not limited by any specific model structure. Proposing multiple plausible minimal output sets to experimental researchers, enables them to select model outputs based on factors such as measurement cost and complexity. Offering measurement flexibility whilst ensuring structural identifiability is a useful tool to scientists and our algorithm could propose these minimal sets within a couple of minutes. In the future we intent to increase the numerical accuracy of our method by making use of the increased integration accuracy obtained by using complex derivatives to compute the Jacobi matrices $\partial f/\partial x$ and $\partial f/\partial \theta$. This step will increase the tolerance of the elements of the output sensitivity matrix to 10^{-20} [30]. In addition, we are investigating the added advantages of concatenating the sensitivity matrix for different values of the model parameters. Preliminary results indicate that this can have a dramatic effect on the accuracy in our computations [22].

Supporting information

S1 File. A chemical reaction system description. A description of model kinetics and all model states and parameters.
(PDF)

S2 File. NF-κB model description. A description of model kinetics and all model states and parameters.
(PDF)

S3 File. JAK/STAT model description. A description of model kinetics and all model states and parameters.
(PDF)

S4 File. Ligand binding model description. A description of model kinetics and all model states and parameters.
(PDF)

S5 File. Simplified glycolytic reaction model description. A description of model kinetics and all model states and parameters.
(PDF)

S6 File. Goldbeter model with specific model output description. A description of model kinetics and all model states and parameters.
(PDF)

S7 File. JAK/STAT model with specific model output description. A description of model kinetics and all model states and parameters.
(PDF)

S8 File. Symbolically verified sets of correlated parameters.
(PDF)

S9 File. Bernoulli trials. How to ensure that a set of unidentifiable parameters is identified with 99.5% certainty.
(PDF)

Author Contributions

Conceptualization: D. Joubert, J. D. Stigter, J. Molenaar.

Software: D. Joubert, J. D. Stigter.

Writing – original draft: D. Joubert.

Writing – review & editing: D. Joubert, J. D. Stigter, J. Molenaar.

References

1. Raue A, Karlsson J, Saccomani MP, Jirstrand M, Timmer J. Comparison of approaches for parameter identifiability analysis of biological systems. Bioinformatics. 2014; 30(10):1440–1448. https://doi.org/10.1093/bioinformatics/btu006 PMID: 24463185

2. Cole DJ, Morgan BJT, Titterington DM. Determining the parametric structure of models. Math Biosci. 2010; 228(1):16–30 https://doi.org/10.1016/j.mbs.2010.08.004 PMID: 20800072

3. Anguelova M, Karlsson J, Jirstrand M. Minimal output sets for identifiability. Math Biosci. 2012; 239 (1):139–153. https://doi.org/10.1016/j.mbs.2012.04.005 PMID: 22609467

4. August, E. Parameter identifiability and optimal experimental design. Proceedings of the 12th IEEE International Conference on Computational Science and Engineering; 2009 Aug 29-31; Vancouver, Canada. IEEE Computational Society, 2009.

5. Liu YY, Slotine JJ, Barabási AL. Observability of complex systems. Proc Natl Acad Sci USA. 2013; 110 (7):2460–2465. https://doi.org/10.1073/pnas.1215508110 PMID: 23359701

6. Letellier C, Sendiña-Nadal I, Bianco-Martinez E, Baptista MS. A symbolic network-based nonlinear theory for dynamical systems observability. Sci Rep. 2018; 8(3785):1–15

7. Tunali ET, Tarn TJ. New results for identifiability of nonlinear systems. IEEE Trans Autom Control. 1987; 32(2):146–154. https://doi.org/10.1109/TAC.1987.1104544

8. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms. 3rd ed. Cambridge: The MIT Press; 2009.

9. Docherty PD, Chase J, Lotz TF, Desaive T. A graphical method for practical and informative identifiability analyses of physiological models: A case study of insulin kinetics and sensitivity. Biomed Eng Online. 2011; 10(1):39–59. https://doi.org/10.1186/1475-925X-10-39 PMID: 21615928

10. Stigter JD, Molenaar J. A fast algorithm to assess local structural identifiability. Automatica. 2015; 58:118–124 https://doi.org/10.1016/j.automatica.2015.05.004

11. Chis OT, Banga JR, Balsa-Canto E. Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods. PLoS ONE. 2011; 6(11):e27755. https://doi.org/10.1371/journal.pone.0027755 PMID: 22132135

12. Reid JG. Structural identifiability in linear time invariant systems. IEEE Trans Automat Contr. 1977; 22 (2):242–246 https://doi.org/10.1109/TAC.1977.1101474

13. Cobelli C, DiStefano JJ III. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. Am J Physiol. 1980; 239(1):R7–R24 https://doi.org/10.1152/ajpregu.1980.239.1.R7 PMID: 7396041

14. Cintrón-Arias A, Banks HT, Capaldi A, Lloyd AL. A sensitivity matrix based methodology for inverse problem formulation. Journal of Inverse and Ill-posed Problems. 2009; 17(6):545–564

15. Golub GH, Van Loan CF. Matrix computations. 3rd ed. Baltimore: The Johns Hopkins University Press; 1996.

16. Gantmacher FR. The Theory of Matrices. 1st ed. New York: Chelsea publishing company; 1960.

17. Bard Y. Nonlinear Parameter Estimation. 1st ed. New York: Academic Press; 1974.

18. Miao H, Xia X, Perelson AS, Wu H. On Identifiability of Nonlinear ODE Models and Applications in Viral Dynamics. SIAM Rev Soc Ind Appl Math. 2011; 53(1):3–39 https://doi.org/10.1137/090757009 PMID: 21785515

19. Bapat RB. Linear Algebra and Linear Models. 3rd ed. New York: Springer-Verlag; 2012.

20. Chu Y, Hahn J. Parameter set selection for estimation of nonlinear dynamic systems. AIChE J. 2007; 53(11):2858–2870 https://doi.org/10.1002/aic.11295

21. Quintana-Ortí G. and Quintana-Ortí E S. Parallel codes for computing the numerical rank. Linear Algebra Appl. 1998; 275-276(1998):451–470 https://doi.org/10.1016/S0024-3795(97)10032-5

22. Stigter JD, Joubert D, Molenaar J. Observability of Complex Systems: Finding the Gap. Sci Rep. 2017; 7(16566):1–9

23. Lipniacki T, Paszek P, Brasier AR, Luxon B, Kimmel M. Mathematical model of NF-KB regulatory module. J Theor Biol. 2004; 228(2004):195–215 https://doi.org/10.1016/j.jtbi.2004.01.001

24. Yamada S, Shiono S, Joo A, Yoshimura A. Control mechanism of JAK/STAT signal transduction pathway. FEBS Lett. 2003; 534(1-3):190–196 https://doi.org/10.1016/S0014-5793(02)03842-5 PMID: 12527385

25. Raue A, Becker V, Klingmüller U, Timmer J. Identifiability and observability analysis for experimental design in nonlinear dynamical models. Chaos. 2010; 20(4):045105 https://doi.org/10.1063/1.3528102 PMID: 21198117

26. Heinrich R, Schuster S. The regulation of Cellular Systems. Berlin: Springer; 1996.

27. Goldbeter A. A model for circadian oscillations in the Drosophila period protein (PER). Proc Biol Sci. 1995; 261(1362):319–324 https://doi.org/10.1098/rspb.1995.0153 PMID: 8587874

28. Sedoglavic A. A probabilistic algorithm to test local algebraic observability in polynomial time. J Symb Comput. 2002; 33:735–755 https://doi.org/10.1006/jsco.2002.0532

29. Raia V, Schilling M, Böhm M, Hahn B, Kowarsch A, Raue A, et al. Dynamic mathematical modeling of IL13-induced signaling in Hodgkin and primary mediastinal B-cell lymphoma allows prediction of therapeutic targets. Cancer Res. 2011; 71(3):693–704 https://doi.org/10.1158/0008-5472.CAN-10-2987 PMID: 21127196

30. Martins J, Sturdza P, Alonso J. The connection between the complex-step derivative approximation and algorithmic differentiation. Proceedings of the 39th Aerospace Sciences Meeting, Reno, USA, 2001.