Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Re-estimation-verified selection of small parameter subsets for systems biology models

Monica Maria Miranda Ramos

mmmirand@espol.edu.ec

Escuela Superior Politecnica del Litoral

Carlos Alberto Salazar López

Escuela Superior Politecnica del Litoral

Cesar Antonio Martin Moreno

Escuela Superior Politecnica del Litoral

Adriana Allison Aguirre Alonso

Escuela Superior Politecnica del Litoral

Article

Keywords: parameter reduction, practical identifiability, sensitivity analysis, ordinary differential equations, systems biology, formal verification

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11750537/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Re-estimation-verified selection of small parameter subsets for systems biology models

Mónica Miranda R.[1*], Carlos Salazar[1], César Martín[1], Adriana Aguirre[1]

[1*]Facultad de Ingeniería en Electricidad y Computación (FIEC), Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador.

*Corresponding author(s). E-mail(s): mmmirand@espol.edu.ec;
Contributing authors: asalazar@espol.edu.ec; cmartin@espol.edu.ec; aaaguirr@espol.edu.ec;

Abstract

Mathematical models of biological systems have far more unknown parameters than experiments can determine, and estimating all of them is slow, unstable and yields values that the data cannot pin down. A common remedy is to estimate only a few and keep the rest at published values, yet the methods that choose those few do not check that the simplified model still behaves like the original one. This work presents a selection method that accepts a small set of parameters only after verifying that re-estimating them reproduces the response of the full model to random changes of all its parameters, and then discards every parameter that is not needed. Applied to 34 published models with real data, from cell signalling to epidemics, it found that two parameters usually suffice, and that the data determine them. Against four established criteria, its sets were sufficient and determined by the data in at least as many models. The mathematical results behind the method are verified by a computer proof assistant. Recalibrating a published model with new data can thus start from a few checked parameters instead of a median of fifteen.

Keywords: parameter reduction, practical identifiability, sensitivity analysis, ordinary differential equations, systems biology, formal verification

Introduction

Ordinary differential equation (ODE) models are central to quantitative biology, pharmacology and epidemiology, where they describe how interacting molecular species or populations change over time as a function of kinetic parameters θ[1]. Using a model for prediction requires estimating θ from measurements, and this becomes difficult as models grow. The main obstacle is non-identifiability: different parameter values may produce the same measured output, because of model symmetries (structural non-identifiability) or because the data are insufficient (practical non-identifiability)[2–6]. The likelihood then has flat ridges on which optimisers stall and confidence intervals become unreliable[7–9]. Models in public benchmark collections routinely contain 20–300 parameters[10].

The sloppy-model phenomenon offers a structural explanation[11, 12]: the outputs are typically sensitive to a few "stiff" parameter directions and almost insensitive to many "sloppy" ones[13], which suggests that a small subset of parameters may suffice to describe what is actually measured. Global sensitivity indices[14, 15] and profile likelihoods[7, 16, 17] (which test whether the data bound each parameter) rank parameters by importance or identifiability, and subset-selection methods based on the sensitivity or Fisher information matrix additionally avoid redundant parameters, whose effects on the outputs are nearly proportional to those of others[18]: Fisher-information criteria[19, 20], collinearity indices such as Brun's[21, 22], orthogonalization[23–25], eigenvalue methods[26] and pivoted QR, a matrix factorization that orders the columns of the sensitivity matrix by how much new information each adds[28, 29], several of which have been compared on signalling and physiological models[26, 30]. All of them return a subset, but none appears to test whether re-estimating the selected subset reproduces the outputs of the full model when all parameters change, and the validity of statements such as "the discarded parameters can be fixed" is rarely made explicit.

This work was motivated by a 22-parameter model of Bandura's social cognitive theory[31–34], a psychological theory of how beliefs, goals and the social environment shape behaviour, in which many stable parameter sets fitted the same data with widely different quality[35].

This work presents a reduction method that works directly on the measured outputs defined in the PEtab format[36], a standard that bundles a model in SBML (Systems Biology Markup Language) with its measurements, experimental conditions, transformations and noise levels. It separates biological parameters from calibration parameters (scalings and offsets of the measurement), which are always re-estimated, and selects only among the biological ones: candidates are ranked by sensitivity, redundant ones are rejected by conditioning tests (the condition number and the variance inflation factor, which grow when the effects of parameters overlap), the reduced model is accepted only if re-estimating it reproduces the full model under random parameter perturbations, and backward elimination removes every parameter whose removal keeps the subset admissible under the rule $|S| \geq 2$. The method is evaluated on 34 problems with real data from a public benchmark collection of published models[10, 36], on held-out experiments and synthetic data, against four existing criteria and null baselines under a common protocol; its

robustness and computational cost are also reported. The contributions are threefold: (i) a selection criterion that accepts a subset only after verifying by re-estimation that it reproduces the full model on the measured outputs; (ii) a benchmark-wide comparison with four criteria and two null baselines under a common protocol, which shows both where the test separates subsets and where it does not; and (iii) general theorems behind each step verified in the Lean 4 proof assistant[37, 38], a program that accepts a proof only when every step passes a small trusted kernel, with its mathematical library Mathlib, with an explicit statement of what is proved in general and what is only checked numerically for each result.

Results

Overview of the method

Figure 1 summarises the procedure; all symbols are defined in Methods (*Measured outputs and parameter classes* to *Admissibility and re-estimation*) and listed in Supplementary Table S2. The input is a PEtab problem: an SBML model, the measured observables with their conditions, times, transformations and noise standard deviations σ_i, and the nominal parameter values θ_0 reported by the original authors. Each measurement defines a scaled output $y_i(\theta) = h(g_i(x(t_i;\theta),\theta))/\sigma_i$, in units of measurement noise. Only biological parameters θ_B (those of the SBML model) are candidates for selection; calibration parameters θ_C are always re-estimated with the selected ones.

The relative sensitivity matrix $J = \partial y/\partial \log\theta$ is computed once at θ_0. In stage 1, biological parameters are added in decreasing order of their sensitivity energy $E_j = \|J_{:,j}\|^2$, a candidate being accepted only if the condition number κ and the largest variance inflation factor (VIF) of the selected columns stay below a threshold, until a fraction τ_R of the energy is captured. A subset S is admissible if, in random scenarios in which all parameters are perturbed, re-estimating θ_S and θ_C with the other parameters at θ_0 gives a median relative fitting error $e_{\text{fit}} \le \tau_e$, and $|S| \ge 2$. The threshold $\tau_e = \sqrt{1-c^2} = 0.436$ ($c = 0.9$) is the smallest error attainable by rescaling an output change whose cosine similarity with the full change is c (Theorem 1). If stage 1 does not give an admissible subset, stage 2 orders the parameters by their finite-perturbation effect s_j and stops at the first admissible subset. Backward elimination finally removes any parameter whose removal keeps the subset admissible (Theorem 2).

Benchmark: a few biological parameters suffice in most models

The method was applied to 34 PEtab benchmark problems with real measurements[10, 36], covering signalling, metabolism, gene regulation, cell cycle, immunology and epidemiology, with 3 to 294 parameters and 16 to 27,132 measurements (Table 1; Methods, *Models, data and simulation*; Supplementary Note *System-by-system summary*). An admissible subset was found in 31 of the 34 problems (Fig. 2). In 25 of them admissibility was established by nonlinear re-estimation; in the six with the most expensive simulations (Alkan, Bachmann,

Chen, Isensee, Lucarelli, Raimundez) it was decided with a first-order approximation of e_{fit} and then checked by nonlinear re-estimation of the final subset, which confirmed it in all six (in Lucarelli with a margin of 4×10^{-5}; Methods, *Admissibility and re-estimation, First-order e_{fit} and nonlinear check*). In 28 problems the subset contains two biological parameters, which is also the minimum size enforced by the method ($k_{\min} = 2$): a single biological parameter already reaches the threshold in 15 of 27 tested problems (Methods, *Settings, Single-parameter subsets*). The other three subsets contain three, five and ten parameters (Zhao, Bruno, Lucarelli). Counting the calibration parameters, which are re-estimated in every case, the reduced problem has a median of 78% fewer parameters than those estimated in the PEtab problem (range 33–97%; Fig. 2a), or 73% (0–97%) if the noise parameters, which are kept fixed, are not counted, the 0 being Perelson, whose two biological parameters are both selected; 86% among biological parameters alone; Methods, *Admissibility and re-estimation*).

All 31 subsets satisfy the conditioning thresholds of the selection, which apply to the biological columns ($\tau_\kappa = \tau_{\mathrm{VIF}} = 10$; maximum $\kappa = 6.5$, VIF= 9.8; together with the calibration parameters the VIF exceeds 10 in ten problems; Methods, *Sensitivities, energy and conditioning*). In 17 of them the median cosine similarity is also at least c, so that a single rescaling of the reduced response already reaches the threshold, and in two (Alkan, Perelson) the captured energy also reaches τ_R, although the corresponding formal error bound requires a model verified in Lean, which these two are not; Table 1 gives this level of guarantee for every problem (Methods, *Admissibility and re-estimation*). Stage 1 sufficed in seven problems; in the other 24 the energy ordering selected parameters whose re-estimation could not reproduce the full model, and stage 2 was needed (Methods, *Selection*). The three problems without an admissible subset fail for different reasons: in Crauste, 1% changes of single parameters make the solution grow without bound within the measured horizon, so no local reduction is defined (Methods, *Models, data and simulation, Crauste*); in Beer and Lang the search stopped at the conditioning limit with 18 and 13 parameters ($e_{\mathrm{fit}} = 50\%$ and 56%), and in Lang other conditioning thresholds lead to admissible pairs (Methods, *Settings*), so the failure there is one of the greedy search, not of the model.

Re-estimation is essential. When the selected parameters are simply set to their perturbed values, the median error e_{rel} exceeds τ_e in 23 of the 31 admissible problems (Fig. 2b; Methods, *Admissibility and re-estimation*); after re-estimation the median e_{fit} is 15% (maximum 43.585%, in Lucarelli, against $\tau_e = 43.589\%$). The selected parameters do not replicate each perturbation; adjusting them compensates for the perturbations of all the other parameters, the property needed when the reduced model is fitted to data.

The selected parameters are also identifiable

A subset that reproduces the full model need not be determined by the data; a selected rate constant could, for example, compensate for a measurement scaling factor. The profile likelihood of every selected parameter was therefore computed in the reduced problem, with $S \cup \Theta_C$ free and the other parameters at θ_0 (Methods, *Identifiability of the selected parameters*). In 25 of the 31 problems all selected

parameters were practically identifiable, that is, their 95% profile-likelihood interval was bounded on both sides within the range of the study (65 of the 74 selected parameters; Table 1, column Ident.); with the threshold scaled by the misfit χ^2/d.f. of the reduced problem, 23 problems and 60 parameters. These intervals are conditional on the published values of the fixed parameters and therefore understate the uncertainty. The exceptions were Fujita (a rate constant confounded with a scaling factor), Giordano and Perelson (recorded noise far above the residuals), Okuonghae and Oliveira (one profile could not be simulated) and Smith (optimum at the bound). In the profile likelihoods of the full model, with all parameters free (computed for the comparison of criteria; Methods, *Comparison with other selection criteria*, *Profile likelihood*), only 20 of the 52 selected parameters that could be evaluated were identifiable: the data do not determine all parameters of these models but, once the others are fixed at their published values, they determine the few that the method selects. In every problem whose profiles could be evaluated, identifiability agreed with the profiled information $I_j = 1/[(A^\top A)^{-1}]_{jj}$, the quantity on which a sufficient condition verified in Lean is based (Methods, *Identifiability of the selected parameters*, *Formal condition*).

Prediction of held-out experiments

The measurements of each problem were split into training and test sets, by experimental condition (21 problems) or, with a single condition, by time (9 problems; Blasi could not be split; Methods, *External validation*); the reduced model (θ_S and θ_C) and the full model (all parameters) were fitted to the training data from θ_0, and both were evaluated on the test data. With the non-selected parameters at θ_0, the reduced model predicted the held-out experiments at least as accurately as the full model (test χ^2/n within 5% or lower) in 26 of the 29 comparable problems (23 of the 26 shown in Fig. 3a, which omits Giordano, Perelson and Rahman, fitted far below the noise level); the exceptions were Raia, Fiedler and Chen. Where the refitted full model had a much larger test error the difference was large (for example Brannmark, test χ^2/n 6.0 against 6,277; Bruno, 0.9 against 208).

This comparison favours the reduced model, because θ_0 was estimated from all data, test experiments included, and the reduced model inherits that information through the parameters it keeps fixed. Two checks quantify this (Fig. 3c; Methods, *External validation*, *Nested validation without the θ_0 anchor*): θ_0 alone predicted the test data better than the reduced model, by more than 5%, in 20 of 29 problems (median ratio 0.90), and a nested reduced model, with the non-selected parameters fixed at a full-model fit to the training data instead of θ_0, predicted worse than the θ_0-anchored one in 21 of 29 (median ratio 1.35). The advantage of the reduced model on real data therefore reflects the value of anchoring at published estimates, not a property of the selected subset alone.

To measure prediction without this confound, synthetic data were generated from a known truth θ^* within $\pm 20\%$ of θ_0 in every estimable parameter, with noise of the recorded size, and the experiment was repeated with 20 seeds (Methods, *External validation*, *Synthetic experiment*; Supplementary Note *Synthetic prediction experiment*); θ_0 has not seen these data, although it still indicates the region of the

truth, so the test is local. Among the cases in which the parameter change exceeded the noise, the reduced model reached a relative prediction error $e_{\text{pred}} \leq \tau_e$ in 56% (178 of 320) and predicted at least as well as the full model in 57% (160 of 282). The outcome was stable within each problem (Fig. 3b): the reduced model predicted better in almost every seed of the problems in which the full model overfits (Weber 20 of 20 seeds, Zheng 18 of 20) and worse in almost every seed of others (Oliveira 0 of 20). On data that θ_0 has not seen, the results show no overall predictive advantage of the reduced models over the full models, and equivalence was not tested; the reduced models predicted better where the refitted full model had a much larger test error. Success decreased with the distance of the truth from θ_0 (64%, 54% and 51% at ±5%, ±20% and ±50%, ten seeds) and grew with the size of the true change in noise units (Methods, *External validation, True signal*).

Comparison with existing selection criteria

The subsets of the proposed method were compared with those of four existing criteria: pivoted QR of the sensitivity matrix in the standard subset-selection form of Golub, Klema and Stewart[27, 28]; the collinearity index of Brun *et al.*[21]; Morris screening over $[\theta_0/10, 10\theta_0]$[15]; and profile likelihood[7, 16], retaining the parameters with the narrowest confidence intervals. Every criterion chose the same number $k = |S|$ of biological parameters from the training outputs (the proposed selection on the training outputs chose its own size), and every subset was evaluated with the same three tests: reproduction of the full model (the admissibility test, at ±5% and at ±20%), prediction of the held-out experiments and prediction of synthetic data (Methods, *Comparison with other selection criteria*; Table 2, Fig. 4). Two null baselines re-estimated only the calibration parameters ($k = 0$) or k biological parameters drawn at random (20 draws). The main proposed subsets were selected on all measured outputs at θ_0; to remove the difference in the outputs used for selection (the scenarios remain those of the proposed selection), the proposed selection was repeated on the training outputs alone and these subsets were passed through the same tests (Methods, *Comparison with other selection criteria*, *Selection on the training outputs and other families of truths*).

On real data the criteria did not differ significantly (Fig. 4c; Methods, *Statistics*): Wilcoxon tests over problems found no significant difference between the proposed subsets and those of Brun, Morris, profile likelihood, pivoted QR, random subsets or the proposed subsets selected on the training outputs ($P = 0.16$–0.46), and the calibration-only baseline was slightly better (median ratio of test χ^2/n 0.99, $P = 0.003$, $n = 28$). As expected for sloppy models[11], many small subsets fit the training data and predict similarly when the other parameters stay at θ_0, so this test does not discriminate between criteria.

The criteria differed in whether their subsets reproduced the full model (Fig. 4a). At ±5%, the proposed subsets were admissible in all 28 comparable problems, against 26 for Brun, 23 for Morris, 18 of 27 for profile likelihood and 17 for pivoted QR (exact McNemar test, Methods, *Statistics*; Holm-adjusted $P = 0.5$, 0.13, 0.012 and 0.004). The null baselines show that in many problems the test is not selective: re-estimating the calibration parameters alone was admissible in 9 of 28 problems,

and random subsets of the same size in a median of 16 of 28 (14–19 over the 20 draws); the test therefore separated subsets in about 12 of the 28 problems. Because the scenarios also perturb the calibration parameters, part of the change is undone by re-estimating them (a median of 30% of the output change in the 14 problems that have such parameters, up to 93% in Laske). With the calibration parameters kept at θ_0 and the same biological perturbations, the proposed subsets remained admissible in 13 of these 14 problems (Bachmann rose to 44.1%, just above τ_e), calibration-only re-estimation fell from 9 to 6 and the admissible random draws from 220 to 192 of 280; in Laske, where all random pairs had passed, only 2 of 20 did, whereas the proposed pair did (Methods, *Comparison with other selection criteria*, *Calibration parameters fixed*; Supplementary Note *Admissibility with the calibration parameters fixed*). This comparison favours the proposed method, which selects by this very test; selected on the training outputs alone, the proposed subsets were admissible in 24 of 27 problems, against 26 of 28 for Brun's (not tested). At ±20%, a level used by none of the criteria, the counts were 27, 25, 23, 18 of 27 and 16 of 28 ($P = 0.5$, 0.25, 0.023 and 0.004), and 23 of 27 for the proposed training-output subsets; the same directions were rescaled, so this tests the size of the change, not new directions.

In prediction on synthetic data (Fig. 4b), over 281 cases with signal in 24 problems, the proposed subsets reached $e_{\text{pred}} \leq \tau_e$ in 56% of the cases, Brun's in 55%, Morris's in 48%, those of profile likelihood in 39% and those of pivoted QR in 30%; resampling whole problems (Methods, *Statistics*), the advantage was 8 percentage points over Morris (95% confidence interval 2–16; Holm-adjusted $P = 0.010$), 17 over profile likelihood (7–28; $P = 0.002$) and 26 over pivoted QR (12–40; $P < 0.001$), against 1 over Brun (−1 to +3; $P = 0.43$). Because these truths belong to the family of perturbations that the proposed method uses to decide admissibility, the comparison was repeated with log-uniform truths within a factor of two of θ_0 and with sparse truths in which 15% of the parameters change by ±50% (Methods, *Comparison with other selection criteria*, *Selection on the training outputs and other families of truths*). The advantage over profile likelihood and pivoted QR persisted (log-uniform: 50% against 37% and 24%, $P = 0.018$ and < 0.001; sparse: 54% against 40% and 39%, $P = 0.008$ and 0.019), that over Morris held for sparse truths (46%, $P = 0.003$) but not for log-uniform ones (47%), and no family separated the proposed method from Brun's. Selected on the training outputs, the proposed subsets reached the threshold in 51%, 48% and 55% of the cases of the three families, again more often than those of profile likelihood and pivoted QR ($P \leq 0.046$), without a significant difference from Brun's (equivalence was not tested). The median prediction error differed significantly only from that of pivoted QR (Supplementary Note *Comparison of selection criteria on synthetic data*). What distinguishes the proposed method is thus not its ranking, which it shares with Brun's collinearity index and with orthogonalization methods[23, 25], but the verification that follows it.

Identifiable and sufficient

Every parameter selected by each criterion was profiled with the procedure used for the proposed subsets (Methods, *Identifiability of the selected parameters*, *Subsets of other criteria*). Over the 27 problems common to all criteria, 86% of the parameters

selected by the proposed method were practically identifiable, against 85% for Brun, 86% for profile likelihood, 80% for Morris and 76% for pivoted QR (Table 2), so identifiability alone separates only the criteria without a redundancy test. Requiring both properties separated them: every selected parameter was identifiable *and* the subset reproduced the full model in 21 of the 27 problems with the proposed subsets, 18 with Brun's, 14 with those of Morris and of profile likelihood, and 7 with pivoted QR's (Holm-adjusted McNemar $P = 0.38$, 0.047, 0.078 and 0.002; Supplementary Note *Identifiability of the subsets of other criteria*). Profile likelihood selects identifiable parameters that often do not suffice; selection by sensitivity verified by re-estimation gives both.

Robustness to the size of parameter changes

Admissibility is decided with perturbations of ±5%. Keeping each final subset fixed, the evaluation was repeated at ±1, 5, 10, 20, 30, 40 and 50% (15 scenarios per level) and the median e_{fit} was classified into four patterns (A, admissible at all levels; B, failing first at ±20% or more; C, irregular; D, failing at ±10% or less; Methods, *Robustness patterns*). When the same fifteen directions are rescaled at every level, 26 of the 31 problems follow pattern A (Fig. 5), four pattern B (Armistead, Bertozzi, Blasi, Elowitz, whose errors are already 33–40% at ±5%) and one pattern C (Lucarelli); none follows pattern D; for the problems marked in Table 1, robustness beyond the nonlinear ±5% check uses the first-order e_{fit}. With independent directions at each level, 24 problems follow pattern A, one pattern B and six pattern C, and 22 follow pattern A in both designs (five of them evaluated to first order beyond ±5%); the nine problems that change are those whose median error is close to the threshold at large perturbations (Supplementary Note *Robustness to the size of parameter changes*). The median hides individual failures: at ±5% the median pass rate is 93%, and 13 problems reproduce the full model in every valid scenario. Most reductions thus remain valid well beyond the perturbation used for selection; those with a small margin at ±5% fail for large changes.

Sensitivity to method settings and computational cost

The complete method was re-run on the 26 small and medium problems changing one setting at a time (Methods, *Settings*). The 25 admissible subsets were unchanged with finite-difference steps of 0.1% or 5%, with either alternative set of integrator tolerances and with τ_R between 0.80 and 0.95; the perturbation size (±1% or ±10%) changed one subset, and conditioning thresholds of 5 or 30 changed three or four subsets without changing the number of admissible problems. The error threshold has the largest influence: with $\tau_e = 0.10$, 0.20, 0.30 and 0.50, 10, 15, 18 and 23 subsets were unchanged, and $\tau_e = 0.10$ requires a mean of 3.7 biological parameters against 2.1–2.5 for τ_e between 0.30 and 0.50. The threshold is therefore an explicit accuracy–parsimony trade-off, not a tolerance validated for any biological application: e_{fit} is relative to the output change caused by the perturbation, and in units of the recorded noise its median remaining error per measurement was below 0.1σ in 17 of 31 problems and below σ in 29 (Supplementary Table S34). The random seed and the number of scenarios were not varied.

The number of sensitivity simulations grows linearly with the number of parameters ($3p$), they are independent and can run in parallel, and the admissibility test adds 15 re-estimations per candidate subset, whose cost depends on the simulation time and on the number of retained parameters rather than on p. Run from scratch on a shared server with the sensitivities in parallel, the selection took from a few seconds to 33 min per problem (median 19 s; about 3 min for Lang, with 294 parameters). Profile likelihood was as fast on small problems (its first round was faster than the proposed method in 16 of 28) but much slower on large ones: 70 min on 60 cores for Bachmann (90 free parameters; 40 h of processor time) against 10 s, 5.2 h for Isensee (110 h of processor time) against 10 min (the nonlinear check of its final subset added 3.9 h), and an estimated 10^3–10^5 h of processor time for Raimundez and Chen, which were not run. The method itself cost about 23 h of processor time for the 34 problems: 1.3 h for the selection and 21.3 h for the nonlinear verification of the final subset in the six problems selected to first order, almost all of it in Chen (12.7 h), Raimundez (4.1 h) and Isensee (3.9 h). The remaining ∼674 h were spent on the validations, robustness tests and comparisons carried out for this paper; all analyses reported here together took about 697 h of processor time, summed over cores; because they ran in parallel on servers with 88–256 cores, the elapsed time was much shorter (Methods, *Computational cost*; Supplementary Note *Computing time of the analyses*).

Formal verification of the mathematical foundations

The statements on which the method relies are verified in Lean 4 with Mathlib in 32 modules with 309 theorems and lemmas, together with 28 benchmark models; all proofs compile with a single command, without unproved assumptions (`sorry`) or additional axioms (Methods, *Formal verification*; Code availability; Supplementary Tables S4 and S7). Table 3 states, for each step of the method, what is proved, when the statement applies to a numerical result and to which results of this study it applies. Lean does not verify that the optimiser converges, the floating-point values (κ, VIF, $\cos\Delta$, e_{fit}, I_j), the identifiability of the full model, or the translation beyond the nominal parameters and base condition; these are checked numerically. Two statements are central:

Theorem 1. *For vectors $d \neq 0$ and $b \neq 0$ in a real inner-product space,*

$$\min_{\lambda \in \mathbb{R}} \frac{\|d - \lambda b\|}{\|d\|} = \sqrt{1 - \cos^2(d, b)}.$$

Theorem 2. *Let $\mathcal{A}$ be any admissibility predicate on finite sets and S an admissible set. Backward elimination (removing one element at a time whenever the remaining set is still admissible, until no element can be removed) returns an admissible subset of S from which no element can be removed without losing admissibility. If, in addition, $\mathcal{A}$ is monotone (every superset of an admissible set is admissible), no proper subset of the result is admissible.*

Theorem 1 links τ_e to the cosine: whenever $\cos\Delta \geq c$, rescaling the reduced response gives a relative error of at most $\sqrt{1-c^2}$; admissibility ($e_{\mathrm{fit}} \leq \tau_e$) does not require $\cos\Delta \geq c$ and is decided by the computed e_{fit}. The first part of Theorem 2 makes no assumption on the predicate and so holds for it as implemented; this is the sense in which the selected subsets are minimal (relative to the implemented test, including the rule $|S| \geq 2$). Minimality is relative to the criterion, which includes $|S| \geq 2$, and is not minimal cardinality. The second part requires monotonicity, which was checked empirically on all 660 nested pairs of subsets evaluated during the selections (no violation; Methods, *Selection*, *Monotonicity log*). Further verified statements cover the conditioning tests, re-estimation, the transfer of error bounds to transformed outputs and the first-order approximation (Methods, *Formal verification*, *Further verified statements*). A sufficient condition for practical identifiability (theorem `practically_identifiable_grad`) bounds the profile-likelihood interval of a parameter through the profiled information I_j, the residual gradient at the optimum and a bound ε on the nonlinear remainder over the admissible range. Its linear version (Methods, *Identifiability of the selected parameters*, *Formal condition*) agreed with the profile classification for all 72 selected parameters whose profile could be evaluated. With the nonlinear remainder estimated at the computed profile points, the hypotheses of the theorem were satisfied at every evaluated point for 24 parameters, and never for a parameter whose profile is not identifiable; this is a numerical check, not a certificate. Over the admissible set, as the theorem requires, the remainder bound holds only near the optimum: with two variants proved in Lean that need the bound only on the tail of the profile, sampling and adversarial search (Methods, *Identifiability of the selected parameters*, *Hypotheses over the admissible set*) found the hypotheses satisfied for 9 of the 74 parameters within a factor of 1.5 of the optimum, 7 within a factor of 2 and none over the whole range of the study (an apparent case found when sampling the widest range is contradicted by the sampling of a narrower one). Identifiability over the range of the study therefore rests on the computed profiles; with twice as many profile points, the same 65 parameters were identifiable. The verification is local and concerns implications: a conclusion of Table 3 applies to a result only when the model hypotheses are verified in Lean and the computed quantities satisfy the stated inequalities (Supplementary Table S5).

Discussion

In 31 of 34 benchmark problems with real data, re-estimating a subset of usually two biological parameters (the imposed minimum; one sufficed in 15 of 27) together with the calibration parameters reproduced, within the chosen tolerance, the local response of the measured outputs, and, with the other parameters fixed at their published values, the data also determined the selected parameters in 25 problems (23 with a misfit-scaled threshold). Over these 31 problems, the number of biological parameters to be estimated fell from 750 to 74 (median 15 to 2; Table 1). In the terminology of system identification this is a parameter subset selection[28] whose result is checked twice: the subset suffices to reproduce the outputs of the full model and its parameters are practically identifiable. With all parameters free, most

selected parameters are not identifiable, which is the sloppy-model picture[11, 12]; this statement is more concrete because it refers to the outputs actually measured, under the actual conditions and noise levels, and it is checked by re-estimation and profiles rather than inferred from the sensitivity spectrum. It is not a claim of predictive superiority: on data unseen by the published estimate, the reduced models predicted neither consistently better nor worse than the full models, and their advantage on real data reflects the published values.

The comparison with existing criteria clarifies what the method adds. When the other parameters stay at their published values, no criterion predicted held-out experiments significantly differently from the proposed method, except calibration-only re-estimation, which was slightly better (equivalence was not tested), and the criteria with a redundancy test (Brun's, profile likelihood and the proposed one) chose similarly identifiable parameters. They differ in whether their subsets reproduce the full model: random subsets and calibration-only re-estimation failed in 43% and 68% of the problems, profile likelihood and pivoted QR in 33% and 39%, Brun's and the proposed method rarely. Part of the success of the baselines reflects the perturbation of the calibration parameters themselves: without it, the proposed subsets remained admissible in 13 of the 14 problems that have such parameters and calibration-only re-estimation in 6, but in six problems (Brannmark, Fujita, Isensee, Raia, Raimundez, Smith) re-estimating the scalings still compensated for the biological change and almost any subset passed, so there the test does not single out the selected parameters. The proposed subsets were identifiable and sufficient in 21 of 27 problems, against 18 for Brun's index (Holm-adjusted $P = 0.38$) and 14 for profile likelihood ($P = 0.078$); this count partly reflects selection by the sufficiency test itself, and with selection on the training outputs it was 16 of 26. The ranking itself is not new—E_j is the diagonal of the Fisher information matrix of the scaled outputs and, up to a constant, the squared importance measure of Brun *et al.*[21], and orthogonalization methods[23–25] combine the same magnitude with a greedy redundancy test, much like stage 1—and energy alone sufficed in only seven problems, because it does not measure whether adjusting a parameter can compensate for changes in the others. What the method adds is the verification by re-estimation on the outputs of interest, the reduction by backward elimination until no parameter can be removed, and the fact that failures become visible (Beer, Lang, Crauste) instead of returning an untested subset.

The method is practical: its sensitivity simulations grow linearly with the number of parameters and run in parallel, and, in the 26 problems tested, its subsets were unchanged by the finite-difference step, the integrator tolerances and τ_R, whereas the error threshold τ_e sets an explicit accuracy–parsimony trade-off. The mathematical statements behind each step are proved in Lean 4; a statement applies to a particular result when the computed quantities satisfy the inequalities listed in Table 3. Other approaches answer different questions: likelihood-based observability analysis[9] assesses prediction uncertainty after calibration, Bayesian approaches[39] require many more simulations, structural identifiability analysis[2–4] addresses ideal noise-free data, manifold-based methods[40] rewrite the model in terms of emergent parameter combinations, balanced truncation[41, 42] reduces state variables, and

hierarchical optimisation[43] eliminates calibration parameters without selecting among biological ones.

The study has limitations. (i) The analysis is local: sensitivities are computed at θ_0, assumed to be a good fit, admissibility is tested around it, and the synthetic truths are generated around it. (ii) Because the non-selected parameters stay at θ_0, predictions on real held-out data benefit from information that θ_0 already contains. (iii) The minimum subset size of two is a conservative choice: one parameter already reaches the threshold in about half of the problems and is as identifiable and predictive there, but fails in more individual scenarios, especially for large changes (Supplementary Note *Minimum subset size: one parameter or two*). (iv) The error threshold $c = 0.9$ is a design choice relative to the perturbation, and in 25 of the 31 problems a $\pm 5\%$ perturbation changes the outputs by less than the measurement noise (by less than 10^{-4} of it in Giordano and Perelson), so admissibility at this level concerns the response of the model rather than differences detectable in the data. (v) The explicit finite-perturbation certificates of the formal development (Methods) are too conservative to certify any scenario at $\pm 5\%$ in these models (constants 10^2–10^{24}), so admissibility rests on the local theorems and on the numerical test. (vi) The method assumes smooth models, Gaussian noise with the recorded standard deviations and re-estimable calibration parameters; in five problems the residuals do not match the recorded noise levels, one problem with 9,169 conditions was too expensive to analyse on its measured outputs, and three problems had no admissible subset.

For a modeller, the practical message is simple: before calibrating a large model, select a small set of biological parameters, verify that re-estimating them together with the calibration parameters reproduces the full model (and whether the calibration parameters alone already do) and that their profiles are bounded, keep the remaining parameters at their published values, and check the reduced model on held-out experiments. The selection itself is cheap: with the sensitivities computed in parallel it took a median of 19 s per problem, and 10 s for Bachmann, whose profile likelihood took 70 min on 60 cores. The subset is often one of many admissible ones: selected on the training outputs, 20 of the 29 problems with an admissible subset chose a different one, mostly another pair. Natural extensions are forward sensitivity equations for genome-scale models[44], sensitivities at several operating points and the use of the selected subsets in experimental design[45]. In most of the benchmark problems analysed, recalibrating a published model with new data could thus start from two parameters whose sufficiency and identifiability have been checked, instead of a median of 15, with the others kept at their published values.

Methods

Models, data and simulation

This work uses the problems of the PEtab benchmark collection[10, 36] that provide real measurements, an SBML model and nominal parameter values (34 problems; Table 1). This is a secondary analysis of published, de-identified data, for which ethics approval was not required (Ethics declarations). Each problem specifies the observables g_i, their transformation h (linear, log or $\log_{10}$), the noise standard

deviation σ_i of each measurement, the experimental conditions (parameter or initial-value overrides) and, where applicable, pre-equilibration conditions. Models were simulated with libRoadRunner[46] (CVODE, relative and absolute tolerances 10^{-7} and 10^{-10}). Initial assignments that depend on parameters were re-evaluated for every parameter vector, condition-specific overrides were applied before simulation, pre-equilibration was run to steady state before the main simulation, and models with discontinuous inputs were integrated piecewise between discontinuities. Simulated observables were compared with the reference simulations distributed with the benchmark wherever available (Supplementary Note *Verification of the simulator*, Supplementary Table S10). *Crauste.* To establish why no local reduction exists for this problem, each of its 12 estimated parameters was changed by $\pm 1\%$ one at a time and the states were simulated over the measured horizon of 28 days (`crauste_blowup.py`); the solution grows without bound before the end of the horizon in 9 of these 24 changes and in 7 of the 15 admissibility scenarios at $\pm 5\%$, with the same counts for relative integrator tolerances from 10^{-10} to 10^{-5}. One further benchmark problem with real data (Froehlich, 4,231 parameters and 9,169 experimental conditions) was analysed only on its state trajectories, because each evaluation of its measured outputs requires about 9,000 simulations; it is not counted in the 34 problems (see Computational cost).

Measured outputs and parameter classes

For a parameter vector $\theta \in \mathbb{R}^p$ (in the units of the benchmark, excluding noise parameters), the vector of scaled outputs is $y(\theta) \in \mathbb{R}^m$ with entries

$$y_i(\theta) = \frac{h\big(g_i(x(t_i;\theta,c_i),\theta)\big)}{\sigma_i}, \qquad i = 1,\dots,m, \tag{1}$$

where $x(t;\theta,c)$ is the state trajectory under condition c and (t_i, c_i) are the time and condition of measurement i. Parameters that appear in the SBML model are biological (Θ_B); parameters that appear only in the observation functions and not in the model or the condition table (scalings and offsets) are calibration parameters (Θ_C). Only biological parameters are candidates for selection; calibration parameters are always re-estimated.

Sensitivities, energy and conditioning

The relative sensitivity matrix $J \in \mathbb{R}^{m\times p}$ is computed by central differences,

$$J_{:,j} = \frac{y(\theta_0 + \delta\theta_{0,j}e_j) - y(\theta_0 - \delta\theta_{0,j}e_j)}{2\delta}, \tag{2}$$

which approximates $\partial y/\partial \log\theta_j$. The energy of parameter j is $E_j = \|J_{:,j}\|^2 = (J^\top J)_{jj}$ and the fraction of energy captured by a subset $S \subseteq \Theta_B$ is $R_{\mathrm{var}}(S) = \sum_{j\in S} E_j / \sum_{j\in\Theta_B} E_j$. For a subset S, the columns of J_S are centred and scaled to unit variance, giving a matrix F whose Gram matrix is proportional to the correlation matrix R_S of the sensitivity columns; the condition number is

$\kappa(S) = \sigma_{\max}(F)/\sigma_{\min}(F)$ and the largest variance inflation factor is $\mathrm{VIF}(S) = \max_j [R_S^{-1}]_{jj}$[18, 21]. Scaling removes differences of scale between parameters, so that κ and VIF measure only the angles between sensitivity directions. Because R_S is a principal submatrix of the correlation matrix of all columns, both quantities can only increase when parameters are added. The statements of the formal development on the linearised least-squares problem are formulated for columns scaled to unit length without centring, $Z = J_S \,\mathrm{diag}(\|J_{:,j}\|)^{-1}$; computed in that form for the final subsets (`conditioning_check.py`), $\kappa(Z) \leq 6.6$ and $\mathrm{VIF}(Z) \leq 10$ in 30 of the 31 subsets, whereas in Rahman $\kappa(Z) = 12.4$ and $\mathrm{VIF}(Z) = 39$, so the formal statement on the linearised problem does not apply to that subset.

Selection

Greedy selection. Candidates are visited in a given order. A candidate j is accepted if $\kappa(S \cup \{j\}) \leq \tau_\kappa$ and $\mathrm{VIF}(S \cup \{j\}) \leq \tau_{\mathrm{VIF}}$; otherwise it is discarded permanently. Because κ and VIF can only increase when columns are added, a discarded candidate would also be rejected later. *Stage 1* visits biological parameters in decreasing order of E_j and stops when $R_{\mathrm{var}}(S) \geq \tau_R$. *Stage 2*, used if the stage-1 subset is not admissible, visits biological parameters in decreasing order of the finite-perturbation effect $s_j = \|y(\theta_0 + \eta\theta_{0,j}e_j) - y(\theta_0)\|$ and stops at the first admissible subset. *Backward elimination* visits the parameters of the selected subset and removes each one whose removal leaves the subset admissible (between one and five parameters were removed in five problems; Supplementary Table S1). *Monotonicity log.* All subsets evaluated during a selection and their admissibility are stored; for every nested pair $S \subset T$ of evaluated subsets, the violations of monotonicity are counted (S admissible, T not), which bear on the second part of Theorem 2.

Admissibility and re-estimation

Admissibility is evaluated on N random scenarios $\theta^{(k)} = \theta_0 \odot (1 + \varepsilon u^{(k)})$, with $u^{(k)}$ uniform in $[-1, 1]^p$. For each scenario let $\Delta y_{\mathrm{full}} = y(\theta^{(k)}) - y(\theta_0)$. The relative error without re-estimation is $e_{\mathrm{rel}} = \|\Delta y_{\mathrm{full}} - \Delta y_S\|/\|\Delta y_{\mathrm{full}}\|$, where $\Delta y_S = y(\theta^{(k)}_{S\cup\Theta_C}, \theta_{0,\overline{S\cup\Theta_C}}) - y(\theta_0)$ is the output change when only the parameters in $S \cup \Theta_C$ take their scenario values and all others stay at θ_0, and the cosine similarity is $\cos\Delta = \langle \Delta y_{\mathrm{full}}, \Delta y_S\rangle/(\|\Delta y_{\mathrm{full}}\|\|\Delta y_S\|)$. The relative error after re-estimation is

$$e_{\mathrm{fit}} = \min_{\theta_S, \theta_C} \frac{\|y(\theta_S, \theta_C, \theta_{0,\overline{S}}) - y(\theta^{(k)})\|}{\|\Delta y_{\mathrm{full}}\|}, \tag{3}$$

where $\theta_{0,\overline{S}}$ denotes the nominal values of all other biological parameters. The minimisation is carried out in $\log|\theta|$ within $[|\theta_0|/10,\ 10|\theta_0|]$ by a trust-region reflective least-squares method (SciPy[47], at most $100(|S| + |\Theta_C| + 1)$ function evaluations), from two starting points (the scenario value and θ_0); the error without re-estimation is also included, so that $e_{\mathrm{fit}} \leq e_{\mathrm{rel}}$. Scenarios whose full-model output change is below $\max(10^{-10}\|y(\theta_0)\|, 10^{-6})$, or in which the simulation of the full model fails, are discarded, and the number of valid scenarios is reported; this set

does not depend on S. A scenario in which the simulation with only $S \cup \Theta_C$ at their scenario values fails is kept and counts as not reproduced ($e_{\mathrm{rel}} = \infty$). Parameters with a negative nominal value keep their sign in the scenarios and in the re-estimation, which acts on $\log|\theta|$; a parameter with nominal value zero (four in Lang and two in Raimundez) is represented by 10^{-12}, so its relative perturbations and its column of J are negligible and, like any parameter with a zero or negligible column, it is in practice never selected (Supplementary Note *Random scenarios and large problems*). Supplementary Table S3 reconciles the parameter counts used throughout (estimated, biological, calibration and noise parameters, and the free parameters of each analysis). S is admissible if

$$|S| \geq k_{\min} \quad \text{and} \quad \bar{e}_{\mathrm{fit}}(S) = \underset{k}{\mathrm{median}}\; e_{\mathrm{fit}}^{(k)}(S) \leq \tau_e = \sqrt{1-c^2}, \tag{4}$$

the median being taken over the valid scenarios. The reduction in the number of free parameters reported in Results is $1 - (|S| + |\Theta_C|)/p$, and $1 - |S|/|\Theta_B|$ among biological parameters. Because e_{rel} bounds the exact e_{fit} from above, the implementation also accepts a subset whose median e_{rel} is at most τ_e, which makes re-estimation unnecessary; with nonlinear re-estimation $e_{\mathrm{fit}} \leq e_{\mathrm{rel}}$ in every scenario, so the rule $\min(\mathrm{median}\, e_{\mathrm{fit}}, \mathrm{median}\, e_{\mathrm{rel}}) \leq \tau_e$ reduces to the median of e_{fit}, and it differs only for the first-order e_{fit} described below. For each admissible subset a level of guarantee is also reported: level 1 when only this condition holds, level 2 when in addition the median $\cos\Delta$ is at least c, and level 3 when in addition $R_{\mathrm{var}}(S) \geq \tau_R$; these levels are reported, not required. *First-order e_{fit} and nonlinear check.* For the eight problems in which the cost of simulation made repeated re-estimation impractical (marked in Table 1), the first-order version $e_{\mathrm{fit,lin}} = \min\{e_{\mathrm{rel}}, \min_c \|\Delta y_{\mathrm{full}} - J_{S\cup\Theta_C} c\|/\|\Delta y_{\mathrm{full}}\|\}$, evaluated scenario by scenario, was used during the selection. This equals the minimum relative error attainable if the model response were exactly linear in parameter changes; for a truly linear model it equals the exact e_{fit}, while for nonlinear models it is an approximation that replaces per-scenario optimisation with a single projection; its absolute error is bounded by a second-order term (Supplementary Table S7), but no bound on the relative error was evaluated for the individual scenarios. The admissibility decision during the selection is therefore made by linear algebra alone. For the six of these problems with an admissible subset, the final subset was afterwards re-estimated nonlinearly in the same 15 scenarios, with the same starting points, bounds and evaluation budget as for the other problems, the scenarios being distributed over processes (`efit_nonlinear.py`); it confirmed admissibility in all six (median e_{fit}, the values reported in Table 1 and Fig. 2: Alkan 0.120, Bachmann 0.338, Chen 0.081, Isensee 0.113, Lucarelli 0.43585 against $\tau_e = 0.43589$, Raimundez 0.186; Supplementary Note *Nonlinear check of the first-order problems*). The columns of J for the calibration parameters are computed together with those of the biological parameters, because the first-order re-estimation uses $J_{S\cup\Theta_C}$.

Settings

Unless stated otherwise: $\delta = 0.01$; $\eta = 0.1$; $\tau_R = 0.89$; $\tau_\kappa = \tau_{\mathrm{VIF}} = 10$; $N = 15$ scenarios; $\varepsilon = 0.05$; random seed 42; $c = 0.9$, so $\tau_e = 0.436$; $k_{\min} = 2$. The sensitivity analysis changed one setting at a time: $\delta \in \{0.1\%, 5\%\}$; relative/absolute integrator tolerances $10^{-9}/10^{-12}$ and $10^{-6}/10^{-8}$; $\varepsilon \in \{0.01, 0.10\}$; $\tau_R \in \{0.80, 0.85, 0.95\}$; $\tau_\kappa = \tau_{\mathrm{VIF}} \in \{5, 30\}$; and $\tau_e \in \{0.10, 0.20, 0.30, 0.50\}$ (Supplementary Note *Sensitivity to the method settings*). It was run on the 26 problems that could be re-run completely fifteen times in reasonable time; the other eight (Alkan, Bachmann, Beer, Chen, Isensee, Lang, Lucarelli and Raimundez) are those with the most expensive simulations. To relate e_{fit} to the precision of the data, the remaining error of each scenario was also expressed per measurement in units of the recorded noise, $e_{\mathrm{fit}}\|\Delta y_{\mathrm{full}}\|/\sqrt{m}$ (the outputs are already divided by σ_i), and its median over the scenarios was reported. *Single-parameter subsets.* Single-parameter subsets were tested on 27 admissible problems (all except the four with the most expensive simulations, Bachmann, Chen, Isensee and Raimundez): the ten biological parameters with the largest E_j and the ten with the largest s_j were each tested alone, with Θ_C re-estimated, on the same 15 scenarios at $\pm 5\%$ and without the requirement $|S| \geq 2$; the admissible candidate with the smallest median e_{fit} was then validated like the selected subset, and, to test the rule $k_{\min} = 2$, put through the robustness, $\pm 20\%$ admissibility, synthetic and identifiability tests of the selected pair (5.7 h of processor time, 27 min of elapsed time on an 88-core server); in the synthetic experiment the selected subset predicted better than the best single parameter in 9 of these 15 problems, whereas on real data the single parameter had the lower test χ^2/n in 10 of 15 (Supplementary Note *Single-parameter subsets*). In the scan of settings, the perturbation size changed the subset only in Zhao, and admissibility only for Perelson at $\pm 1\%$, where the output change falls below the numerical noise floor. The admissibility test uses the median over $N = 15$ scenarios rather than requiring all of them to pass, because individual scenarios near the boundary of the parameter domain can fail through ODE nonlinearity; seven problems pass with 8–9 of the 15 scenarios. For Lucarelli the median e_{fit} lies within 10^{-3} of τ_e to first order and within 4×10^{-5} after nonlinear re-estimation, so its admissibility is borderline, and its median crosses τ_e at several perturbation sizes (43–47% from $\pm 1\%$ to $\pm 50\%$; pattern C). To relate the size of a perturbation to the noise, the root mean square per measurement of Δy_{full} at $\pm 5\%$ (in units of σ) was computed for each admissible problem; it is below 1 in 25 of the 31 problems and below 10^{-4} in Giordano and Perelson (Supplementary Table S34).

External validation

Measurements were split into training and test sets by experimental condition (every third condition after sorting, or the second of two conditions, used for testing) or, for problems with a single condition, by time (the last third of the distinct measurement times). Parameters that only affect test conditions were excluded from re-estimation. In two problems a split was not possible: in Bertozzi the selected parameters act only in the test conditions, and Blasi has a single condition and a single time point.

All real-data numbers come from a single run of each fit; the re-estimation of over-parameterised full models can end in different local optima[48]. In five problems even θ_0 has a test $\chi^2/n > 10$ (Chen, Okuonghae, Oliveira, Salazar-Cavazos, Smith); the simulation of this work agrees with the reference simulation of the benchmark for Chen, Salazar-Cavazos and Smith, whereas no reference is available for Okuonghae and Oliveira. A large χ^2/n can reflect recorded noise levels that understate the scatter of the data, a misspecified model or inadequate parameter values, and the analysis does not distinguish between these causes. The reduced model (re-estimating $S \cup \Theta_C$) and the full model (re-estimating all parameters) were fitted to the training data from θ_0 with the same least-squares method and bounds, $\hat{\theta} = \arg\min \sum_{i \in \text{train}} (y_i(\theta) - \tilde{y}_i)^2$ over the free parameters of each model within $[|\theta_0|/10, 10|\theta_0|]$, and evaluated by the mean squared noise-scaled residual $\chi^2/n = \frac{1}{n} \sum_{i \in \text{test}} (y_i(\hat{\theta}) - \tilde{y}_i)^2$, where $\tilde{y}_i$ are the scaled measurements. *Synthetic experiment.* In the synthetic experiment, the truth θ^* had each estimable parameter drawn uniformly within ±20% of θ_0 (parameters acting only in test conditions stayed at θ_0), data were simulated with Gaussian noise of standard deviation σ_i, and $e_{\text{pred}} = \|y_{\text{test}}(\hat{\theta}) - y_{\text{test}}(\theta^*)\| / \|y_{\text{test}}(\theta^*) - y_{\text{test}}(\theta_0)\|$; the effect of the parameter change was considered larger than the noise when χ^2/n of θ_0 on the synthetic test data was at least 2. The experiment was repeated with 20 seeds (42–61), each defining a different truth and noise realisation; the full model was not re-estimated for the two problems with more than 100 free parameters after removing those that act only in the test conditions (Chen and Raimundez; Bachmann has 101 parameters that affect the outputs but 84 free ones, and its full model was re-estimated). The experiment was repeated with truths within ±5% and ±50% of θ_0 (seeds 42–51, ten seeds), and compared with the same ten seeds at ±20%. With these amplitudes, the reduced model reached $e_{\text{pred}} \le \tau_e$ in 64% and 51% of the cases with signal at ±5% and ±50% (54% at ±20% on the same seeds) and predicted at least as well as the full model in 67% and 47%. *True signal.* The true signal of a case is the root mean square, over the test measurements, of $y_{\text{test}}(\theta^*) - y_{\text{test}}(\theta_0)$, which is in units of σ because the outputs are scaled by the noise; it is computed from the truth, without noise, whereas the criterion "larger than the noise" above uses the noisy synthetic data. At ±20%, the reduced model reached $e_{\text{pred}} \le \tau_e$ in 1%, 24%, 44% and 75% of the cases whose true signal was below 0.3σ, between 0.3σ and σ, between σ and 3σ, and above 3σ; across seeds, the success rate ranged from 31% to 76% and the fraction of cases predicted at least as well as by the full model from 40% to 73%. Problems whose test χ^2/n at θ_0 was below 0.01 (Giordano, Perelson, Rahman) reproduce the data to far below the recorded noise level and are not shown in Fig. 3a. Test errors were compared through the ratio of the reduced to the reference χ^2/n; ratios between 0.95 and 1.05 were described as similar (a descriptive band, not a test). In the synthetic experiment, a case is a pair (problem, seed); the reduced model predicted at least as well as the full model when $e^{\text{red}}_{\text{pred}} \le e^{\text{full}}_{\text{pred}}$, and the success rate is the number of cases with signal and $e_{\text{pred}} \le \tau_e$ divided by the number of cases with signal. *Nested validation without the θ_0 anchor.* For the nested validation without the θ_0 anchor (a partial control of leakage) (Supplementary Note *Nested validation without the θ_0 anchor*), the full model was first re-estimated on the training data (yielding $\hat{\theta}_{\text{train}}$), and the reduced model was then re-estimated with the

non-selected parameters fixed at $\hat{\theta}_{\mathrm{train}}$ and search bounds derived from $\hat{\theta}_{\mathrm{train}}$, for 29 problems (Bertozzi and Blasi excluded; for Chen the columns of the finite-difference Jacobian were evaluated in parallel and the integrator tolerances were $10^{-6}/10^{-8}$).

Comparison with other selection criteria

Each criterion selected the same number $k = |S|$ of biological parameters as the proposed method, from the training data, and the calibration parameters of the problem were always added. *Pivoted QR*: the standard subset selection of Golub, Klema and Stewart[27, 29], as used for parameter selection by Vélez-Reyes and Verghese[28]: the sensitivity matrix of the training outputs at θ_0, with columns normalised to unit length, is decomposed as $U\Sigma V^{\top}$, and the first k pivots of the column-pivoted QR decomposition of $V_k^{\top}$ (the k dominant right singular vectors) are selected. A simpler variant, the first k pivots of the column-pivoted QR decomposition of the normalised matrix itself, in which the first pivot is decided by rounding and column order when several columns have nearly equal norm, is reported in Supplementary Note *Pivoted QR selection*. *Brun*[21]: dimensionless sensitivities $s_{ij} = \theta_{0,j}\,\partial y_i/\partial\theta_j$ on the n training outputs (forward differences, relative step 10^{-4}), importance $\delta_j^{\mathrm{msqr}} = \|s_{:,j}\|/\sqrt{n}$, collinearity index $\gamma_K = 1/\sqrt{\lambda_{\min}(\tilde{s}_K^{\top}\tilde{s}_K)}$, where $\tilde{s}_K$ is s_K with its columns scaled to unit length, and determinant measure $\rho_K = \det(s_K^{\top}s_K)^{1/(2k)}$; among all k-subsets K of the 15 parameters with the largest δ_j^{msqr}, the one with $\gamma_K \le 10$ and the largest ρ_K is selected (the smallest γ_K if none satisfies the bound). *Morris*[15]: each biological parameter is represented by $x_j \in [0,1]$, mapped linearly onto $\log|\theta_j| \in [\log|\theta_{0,j}| - \log 10,\ \log|\theta_{0,j}| + \log 10]$; ten trajectories (seed 42) on the grid $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$ change one parameter at a time, in random order, by $\Delta = \frac{2}{3}$, giving elementary effects $\mathrm{EE}_j = \|y(x + \Delta e_j) - y(x)\|/\Delta$ on the training outputs, and the k parameters with the largest mean effect μ_j^* over the trajectories are selected. *Profile likelihood*[7, 16]: the full model was first re-estimated on the training data; each biological parameter was then fixed at five values on each side of the optimum (up to a factor of 10, within the same bounds as all other fits) and all other parameters were re-estimated; whenever a point of a profile improved the optimum by more than one unit of χ^2, the full model was re-estimated from that point and the profiles were repeated (at most three times; ten for Raia). A parameter was classified as identifiable when its profile exceeded $\Delta\chi^2 = 3.84$ (95%, one degree of freedom) on both sides, and the k parameters with the narrowest confidence intervals were selected, the width being $w_j = t_j^- + t_j^+$, where $t_j^{\pm}$ is the displacement in $\log|\theta_j|$ at which $\Delta\chi^2$ first exceeds 3.84 on each side, interpolated linearly between profile points (∞ if it is not exceeded); ties are broken in favour of the larger of the smaller maxima of $\Delta\chi^2$ on the two sides. It was run for the 26 admissible problems with at most 60 free parameters and, in a separate run with the profile points distributed over 60 cores, for Bachmann and Lucarelli (90 and 72 free parameters); it was not run for Chen and Raimundez because of its cost. Extrapolating from Bachmann with the simulation time and the numbers of free parameters and profile points, Raimundez (126 free parameters, about 570 profile points, 0.3 s per simulation) would need of the order of 10^3 h of computing time and Chen (155 free

parameters, about 1,520 points, 35 s per simulation) of the order of 10^5 h (Supplementary Note *Profile likelihood*). The criteria were implemented as they are usually applied and therefore differ in some numerical settings (data used to select, derivative scheme and step, candidate set, normalisation and computational budget); these differences are listed in Supplementary Note *Brun and Morris selections*. The *calibration-only* baseline re-estimated Θ_C alone ($k = 0$), and the *random* baseline used 20 random k-subsets of the biological parameters (seeds 0–19); for the latter, the number of admissible problems was counted for each draw and its median over the draws is reported. Every subset was then evaluated with the external validation described above (for random subsets, the median test χ^2/n over the draws), with the admissibility test of the proposed method (the same 15 scenarios and the same rules, with the fixed scenario set described in Admissibility and re-estimation, at $\varepsilon = 0.05$ and, with the same directions $u^{(k)}$, at $\varepsilon = 0.20$; for $k = 0$ the rule $|S| \geq 2$ was not applied) and with the synthetic experiment (20 seeds). *Selection on the training outputs and other families of truths.* The main proposed subsets were selected on all measured outputs at θ_0, whereas the other criteria selected on the training outputs; to remove this difference, the proposed selection was repeated on the rows of the training set only (`REPRO_OUTPUTS=train`), and the resulting subsets were evaluated with the same three tests. The synthetic truths of the main experiment, $\theta^* = \theta_0 \odot (1 + 0.2u)$, belong to the same family of perturbations that the proposed method uses to decide admissibility; the comparison of criteria was therefore repeated, with ten seeds (42–51), for two other families: log-uniform truths, $\theta_j^* = \theta_{0,j}\, 10^{v_j}$ with v_j uniform in $[-0.3, 0.3]$, and sparse truths, in which a random 15% of the estimable parameters change by $\pm 50\%$ and the others stay at θ_0. *Calibration parameters fixed.* Because the scenarios perturb every parameter and Θ_C is always re-estimated, part of the change to be reproduced is produced by Θ_C itself. For the 14 problems of the comparison that have calibration parameters, the admissibility test at $\pm 5\%$ was therefore repeated with Θ_C kept at θ_0 (the same 15 scenarios with the components of Θ_C reset, so the biological perturbations are identical and the comparison is paired) for all subsets: the proposed ones, the training-output ones, the other criteria, calibration only and the 20 random draws (378 tasks, 16.7 h of processor time, 34 min on 90 cores). For the proposed subsets the median fraction $\|\Delta y_C\|^2/\|\Delta y_{\text{full}}\|^2$ of the standard scenarios, where Δy_C is the output change produced by the calibration parameters alone, was also computed. In the other 14 problems the test is unchanged. The problems entering each comparison follow from these requirements. Admissibility was compared in 28 of the 31 admissible problems: Bertozzi was excluded (see Statistics), Blasi has no training/test split on which the other criteria could select, and Chen was not included because the selections of pivoted QR, Brun and Morris on its training outputs require 1,734 simulations of 35 s each (about 17 h of computing time) before any re-estimation. Profile likelihood was compared in 27 of them (all except Raimundez). The synthetic comparison used the 24 problems with at least one seed in which all five criteria had a result and the parameter change exceeded the noise: of the 28, Raimundez lacks the profile-likelihood subset, and Giordano, Laske and Lucarelli have no seed with signal. The computation time of the complete method was measured by running it

from scratch, without cached results, with the sensitivity computations distributed over 16–64 cores; all timings were obtained on a 128-core server shared with other computations and may therefore be inflated by resource contention.

Statistics

Comparisons between criteria were made over problems. For real data and for the synthetic prediction error, the statistic of each problem was the logarithm of the ratio of the errors of two criteria (for synthetic data, its median over seeds) and the two-sided Wilcoxon signed-rank test was applied. For the fraction of synthetic cases reaching $e_{\mathrm{pred}} \le \tau_e$, the success rate of a criterion is $\hat{\pi} = \sum_s a_s / \sum_s n_s$, where n_s is the number of cases with signal of problem s that are common to all criteria and a_s the number of them with $e_{\mathrm{pred}} \le \tau_e$; 95% confidence intervals of the difference $\hat{\pi}_{\mathrm{prop}} - \hat{\pi}_m$ are the 2.5 and 97.5 percentiles of the difference recomputed in 10,000 resamples of whole problems with replacement (seed 0), so that seeds of the same problem are not treated as independent. The two-sided p value of the bootstrap is twice the smaller of the fractions of resampled differences at or below and at or above zero; in the Wilcoxon test, problems with equal errors were discarded (zero method of Wilcoxon). Admissibility, one outcome per problem, was compared with the exact McNemar test: if b problems are admissible only with the proposed subset and c only with the other, $P = \min\{1,\, 2\sum_{i=0}^{\min(b,c)} \binom{b+c}{i} 2^{-(b+c)}\}$. The comparison of admissibility and of the synthetic success rate with the four existing criteria is the primary analysis; within each family of four comparisons (admissibility at one perturbation level, or synthetic success rate) p values were adjusted with the Holm procedure, $\tilde{p}_{(r)} = \max_{q \le r} \min\{1, (5-q)\, p_{(q)}\}$ for the ordered values $p_{(1)} \le \cdots \le p_{(4)}$, and the comparisons on real data are exploratory. The absence of a significant difference is not evidence of equivalence; no equivalence margin was tested. Bertozzi was excluded from the comparison of admissibility because its selected parameters act only in the test conditions; selected on the training outputs, the proposed method finds no admissible subset of two parameters there (a single parameter, `RO_CA`, has $e_{\mathrm{fit}} = 0.08$). Exact P values are reported; for the bootstrap the resolution is 2×10^{-4}.

Identifiability of the selected parameters

For each problem with an admissible subset, the reduced problem ($S \cup \Theta_C$ free, other parameters at θ_0) was fitted to all measurements from θ_0, with the method and bounds of the external validation. For each selected parameter θ_j, the profile

$$\Delta\chi^2_j(t) = \min_{\log|\theta_j| = \log|\hat{\theta}_j| + t} \chi^2(\theta) - \chi^2(\hat{\theta}), \qquad \chi^2(\theta) = \sum_i \big(y_i(\theta) - \tilde{y}_i\big)^2, \tag{5}$$

where $\hat{\theta}$ is the optimum and the minimum is taken over the other free parameters of the reduced problem, was computed at ten displacements t per side, with $|t|$ from 0.05 to 4.6; points outside the bounds $[|\theta_0|/10, 10|\theta_0|]$ of θ_j were not used, and if a point improved the optimum the reduced problem was re-fitted from it and the profile repeated (at most three rounds). θ_j was classified as identifiable if $\Delta\chi^2$

exceeded 3.84 on both sides, practically non-identifiable if on one side, and flat otherwise[7]; a robust variant scaled the threshold by $\max(1, \chi^2_{\min}/\text{d.f.})$. *Formal condition.* Let $r(\delta)$ be the vector of noise-scaled residuals of all measurements when the logarithms of the free parameters move by δ from the optimum, $r_0 = r(0)$, so that $\chi^2 = \|r\|^2$, and let A be its sensitivity matrix at the optimum (central differences, step 0.01 in $\log|\theta|$). The profiled information I_j was computed as the squared norm of the component of the column of A for θ_j orthogonal to the other columns, which equals $1/[(A^\top A)^{-1}]_{jj}$ when $A^\top A$ is invertible; it is the output change, in noise units, that a unit change of $\log|\theta_j|$ produces when the other free parameters compensate as well as possible. The residual gradient γ is the norm of the projection of r_0 onto the range of A (zero at an exact optimum), and ε is the largest nonlinear remainder $\|r(\delta) - r_0 - A\delta\|$ over the computed profile points. The Lean theorem `practically_identifiable_grad` states that if $|\langle r_0, A\delta\rangle| \le \gamma\|A\delta\|$, the nonlinear remainder is at most ε, and $t^* = (\varepsilon + \gamma + \sqrt{\gamma^2 + 3.84 + 2\|r_0\|\varepsilon + 2\gamma\varepsilon})/\sqrt{I_j}$ is smaller than the distances d_j^- and d_j^+ from the optimum to the bounds of $\log|\theta_j|$, then θ_j is practically identifiable. With $\gamma = \varepsilon = 0$, t^* reduces to the Wald half-width $\sqrt{3.84/I_j}$, and the linear version of the condition, used in Results, is $\sqrt{3.84/I_j} < \min(d_j^-, d_j^+)$; the remainder can be evaluated only at points, so this condition was checked at the computed profile points. *Hypotheses over the admissible set.* Because the theorem requires the remainder bound over the whole admissible range, not only at profile points, two variants were proved (`practically_identifiable_grad_tail` and `practically_identifiable_rel_tail`) that need the bound only on the tail of the profile beyond the guaranteed half-width, either in absolute form ($\le \varepsilon$) or relative to the linear change ($\le \eta\|A\delta\|$ with $\eta < 1$). For every selected parameter and side, the tail bounds were estimated over the region of the free parameters within a factor $r \in \{10, 5, 3, 2, 1.5\}$ of the optimum (and within the bounds of the study) from 512 Sobol points, the profile points in the region and an adversarial search (L-BFGS-B maximisation of the remainder, or of the ratio, on the tail, started from the worst sampled points and repeated until the half-width was stable); a parameter satisfies the hypotheses in the region when one of the two conditions holds on both sides. As a check of the profile resolution, all profiles were recomputed with 20 instead of 10 displacements per side, and the profiles of the subsets selected on the training outputs were computed with the same procedure. Because only the selected and calibration parameters are free, the profiles of a problem took a median of 34 s. *Subsets of other criteria.* The subsets chosen by Brun's collinearity index, Morris screening, pivoted QR and profile likelihood (Comparison with other selection criteria) were profiled with exactly the same procedure: same reduced problem (selected and calibration parameters free, the others at θ_0), all measurements, the same ten displacements per side, refits and bounds, and the same classification (`scripts/run_identifiability_other_criteria.sh`). Where a criterion chose the proposed subset, its profiles were reused (10, 6, 2 and 1 problems for Brun, Morris, profile likelihood and pivoted QR). The comparison is made on the 27 problems in which every criterion has a subset (Bertozzi excluded, as in the comparison of admissibility); for each criterion, the identifiable parameters and the problems in which every selected parameter is identifiable and the subset is admissible at $\pm 5\%$

are reported, and the latter are compared with the proposed subsets by the exact McNemar test with Holm adjustment over the four criteria (Statistics). These profiles took 41 h of processor time (Brun 4.7 h, Morris 13.1 h, profile likelihood 8.6 h, pivoted QR 14.4 h). The complete mathematical derivation, the profiles of every problem and the values of all quantities for every parameter are given in Supplementary Note *Identifiability of the selected parameters*. The profiles were computed for all 31 problems with an admissible subset: those of Bachmann, Chen, Isensee and Smith on a 256-core server, that of Raimundez on a 128-core server, and the other 26 on an 8-core workstation (libRoadRunner 2.10.0 on the servers and 2.9.2 on the workstation).

Robustness patterns

For each admissible problem, the final subset was kept fixed and e_{fit} was evaluated at perturbation sizes $\varepsilon \in \{0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}$ with 15 scenarios per level. The fifteen directions $u^{(k)} \in [-1, 1]^p$, with independent uniform components, are drawn once per problem with a fixed seed (42) and reused at every level, $\theta^{(k)} = \theta_0 \odot (1 + \varepsilon u^{(k)})$; the directions are stored with the results. To check that the patterns do not depend on these particular directions, the evaluation was repeated with new directions at each size (seed 42 at ±5%, which reproduces the selection, and $42 + 1000i$ at the i-th other level). Let $\bar{e}(\varepsilon)$ be the median e_{fit} at level ε and ε_f the smallest level with $\bar{e}(\varepsilon_f) > \tau_e$. The pattern is A if no such level exists; C (irregular) if $\bar{e}(\varepsilon) \leq \tau_e$ again at a level above ε_f; otherwise B if $\varepsilon_f \geq 0.20$ and D if $\varepsilon_f \leq 0.10$; the pass rate is the fraction of valid scenarios with $e_{\mathrm{fit}} \leq \tau_e$. As in the selection, e_{fit} was computed to first order for the problems marked in Table 1. For Chen, between 8 and 13 of the 15 scenarios per level could be simulated in the run shown in Fig. 5 (7 to 13 over all runs, because some simulations fail in a run-dependent way); the remaining scenarios failed to integrate and were discarded. A scenario is valid when it could be simulated and its output change exceeds the numerical noise floor; all fifteen scenarios were valid in 28 of the 31 problems, and only 8, 4 and 2 in Chen, Isensee and Perelson. Among the 26 problems of pattern A, 9 pass every scenario at every level, whereas 7 (Borghans, Bruno, Fujita, Liu, Sneyd, Zhao, Zheng) fall to a pass rate of 60–73% at some level. The full distribution of the error for every problem and level is given in Supplementary Tables S11 and S12 and Supplementary Fig. S4.

Computational cost

The cost of the selection is dominated by the computation of J ($2p$ simulations) and of the stage-2 ordering (p simulations); the selection, the conditioning tests and the first-order e_{fit} only use linear algebra on J (Supplementary Note *Computational cost*). The computation time of the complete method was measured by running it from scratch, without cached results, with the sensitivity computations distributed over 16–64 cores of a 128-core server shared with other computations, so the times may be inflated by resource contention. Times are given as elapsed (wall-clock) time when the computation of a problem ran in parallel, and as processor time, the sum over tasks, when it consisted of independent single-core tasks (scenarios, seeds,

profile points). For Armistead, Lang and Raimundez the times were measured with a slightly different scenario set; the sensitivity simulations, which dominate the cost, are the same. Chen, with 35 s per simulation, was the slowest problem (33 min). In the six problems selected with the first-order e_{fit}, the nonlinear re-estimation of the final subset in the 15 scenarios took 4 min (Lucarelli) to 12.7 h (Chen) of processor time (Alkan 8 min, Bachmann 23 min, Isensee 3.9 h, Raimundez 4.1 h). Profile likelihood was much slower on problems with expensive simulations (Alkan 27 min against 25 s for the proposed method; Zheng 5 min against 14 s; Isensee 5.2 h, or 110 h of processor time, against 10 min). For Bachmann (90 free parameters in the profile, 84 in the synthetic experiment after removing those that act only in test conditions; Supplementary Table S3) and Lucarelli (72), profile likelihood took 70 and 11 min with the profile points distributed over 60 cores (40 and 9 h of computing time), against 10 and 35 s for the proposed method; Chen and Raimundez were processed by the proposed method in 33 and 2 min. Selecting a subset took less than ten minutes in 32 of the 34 problems. All analyses reported in this paper together took about 697 h of processor time, the sum over all cores of the times recorded by each task in the final runs (only the final runs are counted, and the compilation of the Lean proofs is not included); because the analyses ran in parallel on servers with 88–256 cores and on an 8-core workstation, the elapsed time was much shorter. 87% of the processor time was spent in Isensee, Chen, Raimundez, Smith and Bachmann; the largest items were profile likelihood of the full model (193 h), the identifiability profiles (172 h: 26 h for the proposed subsets, 102 h for the check with 20 points per side, 3 h for the subsets selected on the training outputs and 41 h for the subsets of the other criteria), the tests of the subsets of all criteria (86 h, and 17 h more with the calibration parameters fixed) and the synthetic experiments (130 h over all amplitudes, families and criteria). The check of the identifiability hypotheses over the admissible set took a further 3 h of elapsed time on an 88-core server (Supplementary Note *Computing time of the analyses*). On state trajectories, the 4,231-parameter Froehlich model required 12,693 simulations and 2.6 h for the sensitivities; for models of this size on measured outputs, forward sensitivity equations[49] compute J in a single integration of an augmented system, whereas adjoint methods are efficient for gradients of a scalar objective but not for the full output Jacobian needed here.

Formal verification

The formal development uses Lean 4 (version 4.23.0) and Mathlib[37, 38] and has two parts (Table 3 and Supplementary Fig. S3). The first proves, for any model that satisfies stated hypotheses, the mathematical results on which each step of the method relies (Supplementary Table S4, steps 1–22). The second establishes those hypotheses for concrete benchmark models. For this, each PEtab problem was translated automatically into a Lean file (`lean_translation/sbml_to_lean.py` of the reproducibility package): the rate laws and stoichiometry of the SBML model and the nominal parameter values and initial conditions of the PEtab tables are written as Lean expressions in which every number is an exact rational. Verified checkers, Boolean functions proved sound in Lean, then test that every rate law is defined and smooth on a domain containing the trajectory, that concentrations

cannot become negative (quasi-positivity) and that the growth of the vector field is bounded so that the solution exists on the simulated horizon (linear growth, or a Riccati bound on a finite horizon for Crauste). The checkers are evaluated by the Lean kernel itself (`decide +kernel`), so their results are part of the proof, and each model ends with a theorem stating that it is smooth, that its solution exists and remains non-negative on the horizon, and that the solution is differentiable with respect to the parameters, which are the hypotheses of steps 14–22. This was achieved for 28 benchmark models (27 of the 34 problems analysed, and Froehlich); the reasons why the other seven problems were not verified are given in Supplementary Note *Verified benchmark models*. Because the theorems are only as relevant as the translation is faithful, the translated vector fields, parameter values and initial values were compared independently with the SBML models evaluated by libRoadRunner at the nominal state and at five random states; they agree to rounding error (largest relative difference 10^{-12}), and changing a single number of a Lean file by 1% is detected. All files compile from scratch without `sorry`; `#print axioms` reports only the standard axioms of Lean's logic (`propext`, `Classical.choice`, `Quot.sound`) for all 309 theorems and lemmas and the 28 model theorems, and every module is accepted by the independent checker `lean4checker`. The file `NonVacuity.lean` gives, for each central theorem, a concrete instance in which every hypothesis is proved (including a nonlinear residual for the identifiability theorems), so that no central statement is vacuous, and `DEFINITIONS_AUDIT.md` compares each Lean definition with the paper and the code. The numerical analyses of this paper (J, R_{var}, κ, VIF, $\cos\Delta$, e_{fit} and the validation) were carried out in Python in double precision; a formal conclusion applies to a numerical result when the model's hypotheses are verified in Lean and the computed quantities satisfy the inequalities that the corresponding theorem requires (Supplementary Table S5).

Further verified statements. Besides Theorems 1 and 2, the first part proves that, to first order, fixing the non-selected parameters changes the outputs by at most $\sqrt{(1 - R_{\mathrm{var}}(S))\, E}\, \|\Delta\theta\|$, where $E = \sum_j E_j$ (by the Cauchy–Schwarz inequality applied to the discarded columns of J), which is the error bound behind level 3; that re-estimating more parameters never increases the fitting error and that $e_{\mathrm{fit}} \le e_{\mathrm{rel}}$, so that the admissibility criterion is monotone when the re-estimation reaches its optimum (with a local optimiser this holds only approximately, which is why monotonicity was also checked empirically); that the greedy selection keeps κ and VIF below their thresholds and that discarding a rejected candidate permanently is correct; that bounds on κ and VIF of the unit-length columns imply a unique and stable solution of the linearised least-squares problem for the retained parameters (which is not a proof of identifiability of the nonlinear model); that error bounds on the states transfer to the measured outputs, including logarithmic transformations; and that the first-order e_{fit} differs from the exact one by a term of second order in the absolute error (which does not by itself bound the relative error used in the decision). Explicit versions of the error bound and of the cosine for finite perturbations are also proved; they bound the second-order remainder through a Lipschitz constant L_c computed from bounds on the first and second derivatives of the vector field. Evaluated for the verified models, L_c ranged from about 10^2 to 10^{24}, too large to

certify any scenario at ±5%, so these finite certificates are of theoretical value only. Theorem `practically_identifiable_grad` (step 32) is stated in the Methods section *Identifiability of the selected parameters*; its hypothesis on the nonlinear remainder is required for all parameter values in the admissible range, whereas it was evaluated at the computed profile points. In seven problems (Armistead, Blasi, Boehm, Bruno, Rahman, Sneyd, Zheng), every selected parameter satisfied these hypotheses at all evaluated points and the model itself is verified in Lean; in the other 24 admissible problems identifiability rests on the computed profiles alone. Supplementary Table S43 lists, for every claim of the Results that refers to the formal development, the Lean theorem, its hypotheses and the numerical quantities that instantiate them.

Use of large language models

An AI assistant (Claude, Anthropic) was used to help write and debug analysis code and Lean proofs and to edit the manuscript text. All code, proofs and results were checked by the authors, who take full responsibility for the content.

Ethics declarations

Ethics approval and consent to participate: Ethics approval was not required for this secondary analysis of publicly available, de-identified data. No new data from human participants were collected for this study. All analyses used publicly available, previously published benchmark models and data (PEtab benchmark collection); some of these data are human-derived, such as the published viral-load measurements of a patient (Perelson problem) and aggregated epidemiological case counts, and were analysed secondarily, in the de-identified form in which they were published. No identifiable participant information was accessed. Ethical approval and informed consent for the original studies were the responsibility of the corresponding original investigations.
Consent for publication: Not applicable.

Data availability. The PEtab benchmark problems analysed in this study are publicly available in the Benchmark-Models-PEtab collection (https://github.com/Benchmarking-Initiative/Benchmark-Models-PEtab, commit `fcbddf1`). All results generated in this study (selected subsets, external validation, synthetic prediction experiments, comparison with other selection criteria including the profile likelihoods, identifiability of the selected parameters, robustness, single-parameter subsets, sensitivity to the method settings, simulator check, computation times and the checks of the Lean models), one file per benchmark problem with a data dictionary, are included in the reproducibility package archived on Zenodo (doi:10.5281/zenodo.23338174) and in the folder `reproducibility/` of https://github.com/monicammr/archivos.

Code availability. The reproducibility package (Zenodo, doi:10.5281/zenodo.23338174; folder `reproducibility/` of https://github.com/monicammr/archivos) contains the Python code of the method and of every analysis, the programs that produce every figure and table of the paper

from the stored results, the translator of the SBML models into Lean, and all Lean 4 proofs (folder `lean_proofs/`, a self-contained project that compiles with a single command, `./verify.sh`, and checks that no proof uses `sorry` or additional axioms). Its manual (`README.md`) describes three independent levels of checking: (A) regenerating every figure, table and number of the paper from the stored results, without simulation, in a few minutes, which shows that the paper follows from those results; (B) compiling all proofs from scratch with `./verify.sh`, in 20 min to 2 h, which shows that every theorem is proved with no unproved assumption; and (C) re-running the analyses from the public PEtab problems and comparing the new results with the stored ones, which shows that the stored results are those the code produces (minutes per problem for the method itself, about 697 h of processor time for all analyses of the paper). Supplementary Note *Reproducibility package* describes each level in detail.

References

[1] Kholodenko, B. N. Cell-signalling dynamics in time and space. *Nature Reviews Molecular Cell Biology* **7**, 165–176 (2006).

[2] Bellman, R. & Åström, K. J. On structural identifiability. *Mathematical Biosciences* **7**, 329–339 (1970).

[3] Chiş, O.-T., Banga, J. R. & Balsa-Canto, E. Structural identifiability of systems biology models: A critical comparison of methods. *PLOS ONE* **6**, e27755 (2011).

[4] Villaverde, A. F., Barreiro, A. & Papachristodoulou, A. Structural identifiability of dynamic systems biology models. *PLOS Computational Biology* **12**, e1005153 (2016).

[5] Miao, H., Xia, X., Perelson, A. S. & Wu, H. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Rev.* **53**, 3–39 (2011).

[6] Wieland, F.-G., Hauber, A. L., Rosenblatt, M., Tönsing, C. & Timmer, J. On structural and practical identifiability. *Curr. Opin. Syst. Biol.* **25**, 60–69 (2021).

[7] Raue, A. *et al.* Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929 (2009).

[8] Raue, A., Kreutz, C., Theis, F. J. & Timmer, J. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A* **371**, 20110544 (2013).

[9] Kreutz, C., Raue, A. & Timmer, J. Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Systems Biology* **6**, 120 (2012).

[10] Hass, H. *et al.* Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics* **35**, 3073–3082 (2019).

[11] Gutenkunst, R. N. *et al.* Universally sloppy parameter sensitivities in systems biology models. *PLOS Computational Biology* **3**, e189 (2007).

[12] Transtrum, M. K. *et al.* Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J. Chem. Phys.* **143**, 010901 (2015).

[13] Monsalve-Bravo, G. M. *et al.* Analysis of sloppiness in model simulations: Unveiling parameter uncertainty when mathematical models are fitted to data. *Sci. Adv.* **8**, eabm5952 (2022).

[14] Saltelli, A., Tarantola, S., Campolongo, F. & Ratto, M. *Sensitivity analysis in practice: A guide to assessing scientific models* (Wiley, 2004).

[15] Morris, M. D. Factorial sampling plans for preliminary computational experiments. *Technometrics* **33**, 161–174 (1991).

[16] Maiwald, T. *et al.* Driving the model to its limit: Profile likelihood based model reduction. *PLOS ONE* **11**, e0162366 (2016).

[17] Eisenberg, M. C. & Hayashi, M. A. L. Determining identifiable parameter combinations using subset profiling. *Math. Biosci.* **256**, 116–126 (2014).

[18] Belsley, D. A., Kuh, E. & Welsch, R. E. *Regression diagnostics: Identifying influential data and sources of collinearity* (Wiley, 1980).

[19] Weijers, S. R. & Vanrolleghem, P. A. A procedure for selecting best identifiable parameters in calibrating activated sludge model no. 1 to full-scale plant data. *Water Sci. Technol.* **36**, 69–79 (1997).

[20] Cintrón-Arias, A., Banks, H. T., Capaldi, A. & Lloyd, A. L. A sensitivity matrix based methodology for inverse problem formulation. *J. Inverse Ill-Posed Probl.* **17**, 545–564 (2009).

[21] Brun, R., Reichert, P. & Künsch, H. R. Practical identifiability analysis of large environmental simulation models. *Water Resources Research* **37**, 1015–1030 (2001).

[22] Gábor, A., Villaverde, A. F. & Banga, J. R. Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC Systems Biology* **11**, 54 (2017).

[23] Yao, K. Z., Shaw, B. M., Kou, B., McAuley, K. B. & Bacon, D. W. Modeling ethylene/butene copolymerization with multi-site catalysts: Parameter estimability and experimental design. *Polym. React. Eng.* **11**, 563–588 (2003).

[24] Li, R., Henson, M. A. & Kurtz, M. J. Selection of model parameters for off-line parameter estimation. *IEEE Trans. Control Syst. Technol.* **12**, 402–412 (2004).

[25] Lund, B. F. & Foss, B. A. Parameter ranking by orthogonalization—Applied to nonlinear mechanistic models. *Automatica* **44**, 278–281 (2008).

[26] Quaiser, T. & Mönnigmann, M. Systematic identifiability testing for unambiguous mechanistic modeling – application to JAK-STAT, MAP kinase, and NF-κB signaling pathway models. *BMC Syst. Biol.* **3**, 50 (2009).

[27] Golub, G. H., Klema, V. & Stewart, G. W. Rank degeneracy and least squares problems. *NBER Working Paper* **165**; https://doi.org/10.3386/w0165 (1977).

[28] Vélez-Reyes, M. & Verghese, G. C. Subset selection in identification, and application to speed and parameter estimation for induction machines. In *Proceedings of the 4th IEEE Conference on Control Applications*, 991–997 (1995).

[29] Golub, G. H. & Van Loan, C. F. *Matrix Computations* (4th edn, Johns Hopkins University Press, 2013).

[30] Olufsen, M. S. & Ottesen, J. T. A practical approach to parameter estimation applied to model predicting heart rate regulation. *J. Math. Biol.* **67**, 39–68 (2013).

[31] Bandura, A. *Social Foundations of Thought and Action: A Social Cognitive Theory* (Prentice-Hall, 1986).

[32] Martín, C. A. *et al.* A dynamical systems model of social cognitive theory. In *2014 American Control Conference (ACC)*, 2407–2412 (IEEE, 2014).

[33] Riley, W. T. *et al.* Development of a dynamic computational model of social cognitive theory. *Transl. Behav. Med.* **6**, 483–495 (2016).

[34] Martín, C. A. *et al.* Development of a control-oriented model of social cognitive theory for optimized mHealth behavioral interventions. *IEEE Trans. Control Syst. Technol.* **28**, 331–346; https://doi.org/10.1109/TCST.2018.2873538 (2020).

[35] Miranda R., M., Salazar L., C., Martín M., C. & Aguirre A., A. Hybrid estimation of dynamic behavioral models using differential evolution and grey-box refinement. In *2025 IEEE Ninth Ecuador Technical Chapters Meeting (ETCM)*, 1–7; https://doi.org/10.1109/ETCM67548.2025.11304524 (IEEE, 2025).

[36] Schmiester, L. *et al.* PEtab—interoperable specification of parameter estimation problems in systems biology. *PLOS Computational Biology* **17**, e1008646 (2021).

[37] de Moura, L. & Ullrich, S. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, Lecture Notes in Computer Science **12699**, 625–635 (Springer, 2021).

[38] Mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020)*, 367–381; https://doi.org/10.1145/3372885.3373824 (ACM, 2020).

[39] Liepe, J. *et al.* A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nature Protocols* **9**, 439–456 (2014).

[40] Transtrum, M. K. & Qiu, P. Model reduction by manifold boundaries. *Physical Review Letters* **113**, 098701 (2014).

[41] Moore, B. C. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control* **26**, 17–32 (1981).

[42] Antoulas, A. C. *Approximation of large-scale dynamical systems* (SIAM, 2005).

[43] Loos, C., Krause, S. & Hasenauer, J. Hierarchical optimization for the efficient parametrization of ODE models. *Bioinformatics* **34**, 4266–4273 (2018).

[44] Fröhlich, F., Kaltenbacher, B., Theis, F. J. & Hasenauer, J. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLOS Computational Biology* **13**, e1005331 (2017).

[45] Transtrum, M. K. & Qiu, P. Optimal experiment selection for parameter estimation in biological differential equation models. *BMC Bioinformatics* **13**, 181 (2012).

[46] Welsh, C. *et al.* libRoadRunner 2.0: a high performance SBML simulation and analysis library. *Bioinformatics* **39**, btac770 (2023).

[47] Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

[48] Villaverde, A. F., Fröhlich, F., Weindl, D., Hasenauer, J. & Banga, J. R. Benchmarking optimization methods for parameter estimation in large kinetic models. *Bioinformatics* **35**, 830–838 (2019).

[49] Fröhlich, F. *et al.* AMICI: High-performance sensitivity analysis for large ordinary differential equation models. *Bioinformatics* **37**, 3676–3677 (2021).

Acknowledgements. The authors thank Escuela Superior Politécnica del Litoral (ESPOL) for institutional support.

Author contributions. M.M.R. and C.S. conceived the study and designed the method. M.M.R. wrote the software and the formal proofs, performed the analyses and wrote the first draft. C.S., C.M. and A.A. supervised the work, contributed to

the validation of the results and revised the manuscript. All authors reviewed and approved the final manuscript.

Funding. This research received no external funding.

Additional information. **Competing interests:** The authors declare no competing interests.
Correspondence and requests for materials should be addressed to M.M.R.
Supplementary information accompanies this paper (Supplementary Notes, Tables and Figures); Supplementary Note *Concluding remarks* summarises its contents.

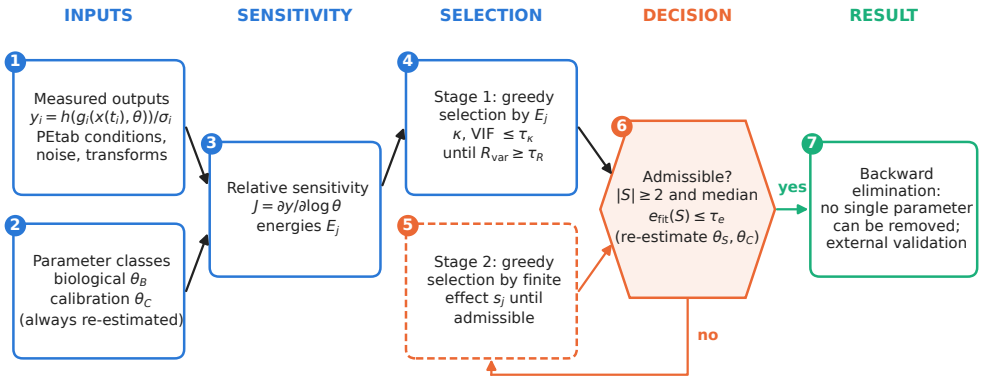


Fig. 1 Overview of the reduction method. Biological parameters are ranked by their effect on the measured outputs and selected with conditioning control (κ and VIF thresholds; stage 1, or stage 2 if needed, dashed). A subset is accepted when re-estimating it reproduces the full model; backward elimination then removes every parameter that is individually dispensable (the result is inclusion-minimal if the criterion is monotone; Theorem 2), and the reduced model is validated on held-out data.

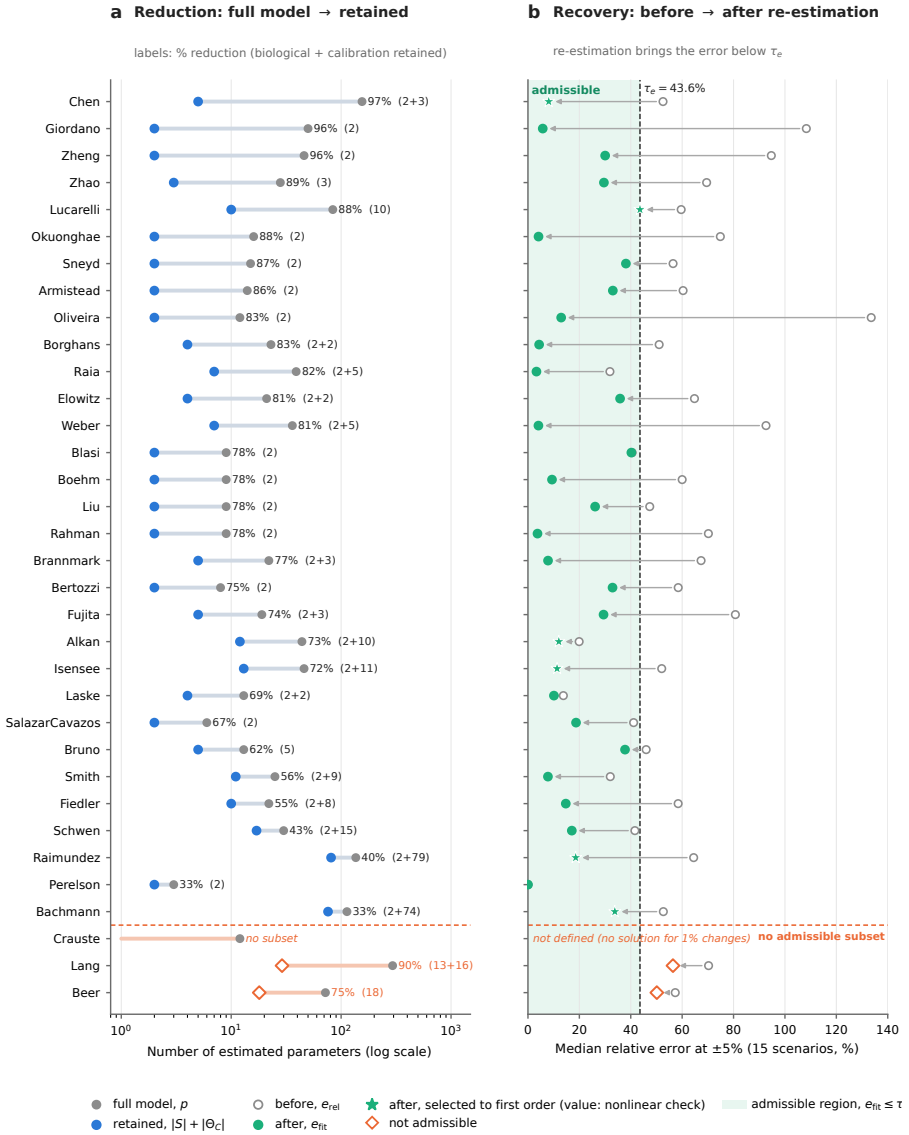


Fig. 2 Parameter reduction and recovery of the full-model output in 34 benchmark problems. Problems are ordered by the reduction achieved; the three without an admissible subset (Crauste, Lang and Beer) are at the bottom. **a**, Estimated parameters of the full model (p, grey, including the noise parameters, which are kept fixed) and parameters re-estimated in the reduced model (selected biological plus calibration parameters, blue); the percentage is the reduction relative to p. Crauste has no admissible subset, so no reduction is shown. **b**, Median relative error over 15 random scenarios at ±5% before (e_{rel}, open circles) and after re-estimating the retained parameters (e_{fit}, green; stars, subsets selected with the first-order e_{fit}, shown with the value of the nonlinear re-estimation of the final subset; for Beer and Lang, first-order value). The shaded region is the admissible region, $e_{\mathrm{fit}} \leq \tau_e = 43.6\%$.

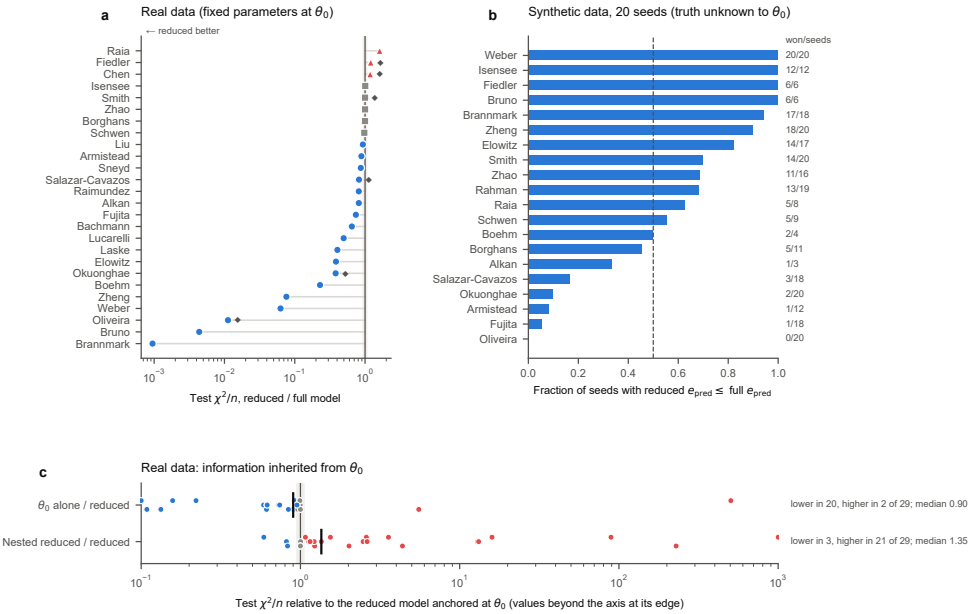


Fig. 3 Prediction of held-out experiments by the reduced and the full model. a, Real data: ratio of the test χ^2/n (mean squared noise-scaled residual on the held-out experiments) of the reduced model to that of the full model re-estimated on the same training data, on a logarithmic scale ($n = 26$ problems), with the non-selected parameters kept at θ_0. Left of the vertical line (circles) the reduced model predicts better; the grey band marks equivalence within 5% (squares); triangles mark worse predictions; ♦ flags a reduced test χ^2/n above 10, where a favourable ratio does not imply a good absolute fit. **b**, Synthetic data from a known truth, which θ_0 has not seen: for each problem, fraction of the 20 seeds in which the reduced model predicted at least as well as the full model (smaller e_{pred}), among the seeds in which the parameter change exceeded the noise (seeds won / seeds with signal on the right; problems with fewer than three such seeds are not shown). Problems are ordered by this fraction. **c**, Real data, information inherited from θ_0: test χ^2/n of the published estimate θ_0 without re-estimation, and of the nested reduced model (non-selected parameters fixed at a full-model fit to the training data instead of θ_0), each relative to the reduced model anchored at θ_0; one point per problem ($n = 29$), vertical bar at the median; blue, lower by more than 5%; red, higher by more than 5%; grey band, within 5%. Problems without a training/test split (Bertozzi, Blasi), without a full re-estimation (in **b**, Chen and Raimundez) or fitted far below the noise level (**a**: Giordano, Perelson, Rahman) are not shown. Per-problem values and scatter plots are given in Supplementary Notes *Prediction of held-out experiments* and *Nested validation without the θ_0 anchor*.

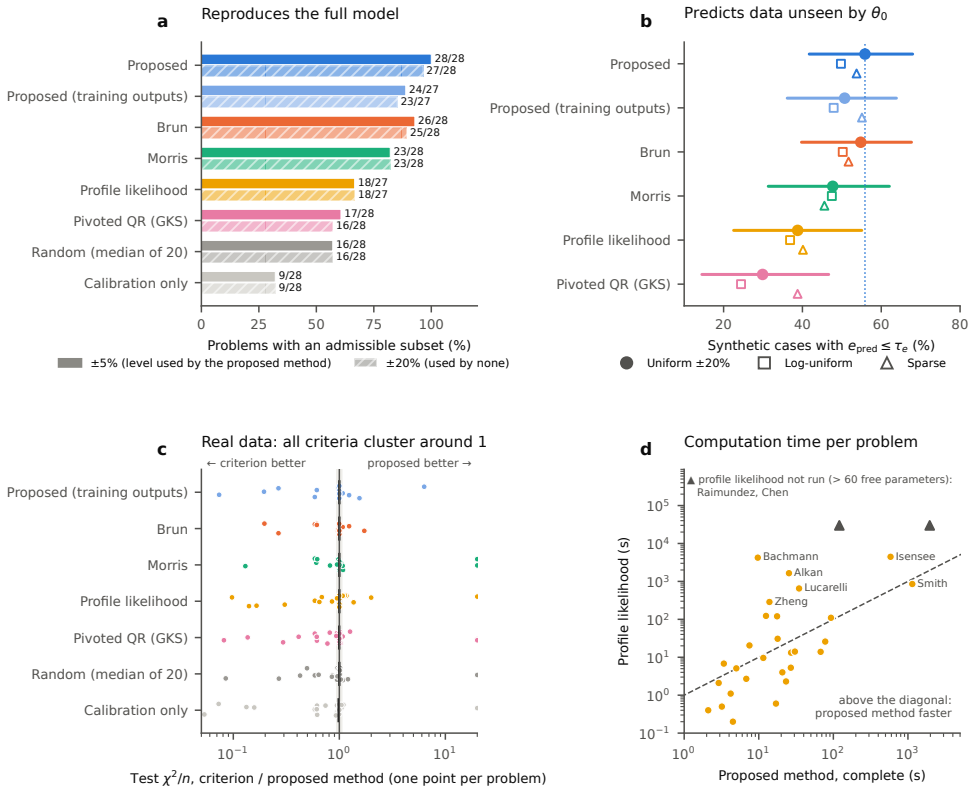


Fig. 4 The proposed method compared with four existing selection criteria and null baselines. a, Number of problems in which the subset of each criterion is admissible, i.e. its re-estimation reproduces the full model, for parameter changes of ±5% (the level used by the proposed method) and ±20% (used by none); 28 problems (27 for profile likelihood and for the proposed subsets selected on the training outputs); random subsets, median over 20 draws. **b**, Fraction of synthetic cases with signal in which the subset predicts the truth within τ_e, for three families of truths (filled circles, uniform within ±20% of θ_0, with 95% confidence intervals from resampling problems; squares, log-uniform; triangles, sparse); dotted line, the proposed method with uniform truths. **c**, Real data: ratio of the test χ^2/n of each criterion to that of the proposed method, one point per problem, vertical bar at the median (log scale; grey band, ±5%; values outside the axis are drawn at its edge); all criteria cluster around 1. **d**, Computation time of the proposed method (selection, admissibility test and backward elimination; the nonlinear check of the final subset in the six problems selected to first order is not included) and of profile likelihood per problem (log scales); points above the diagonal are faster with the proposed method; problems in which profile likelihood was not run because they have more than 60 free parameters (Chen and Raimundez) are shown at the top.

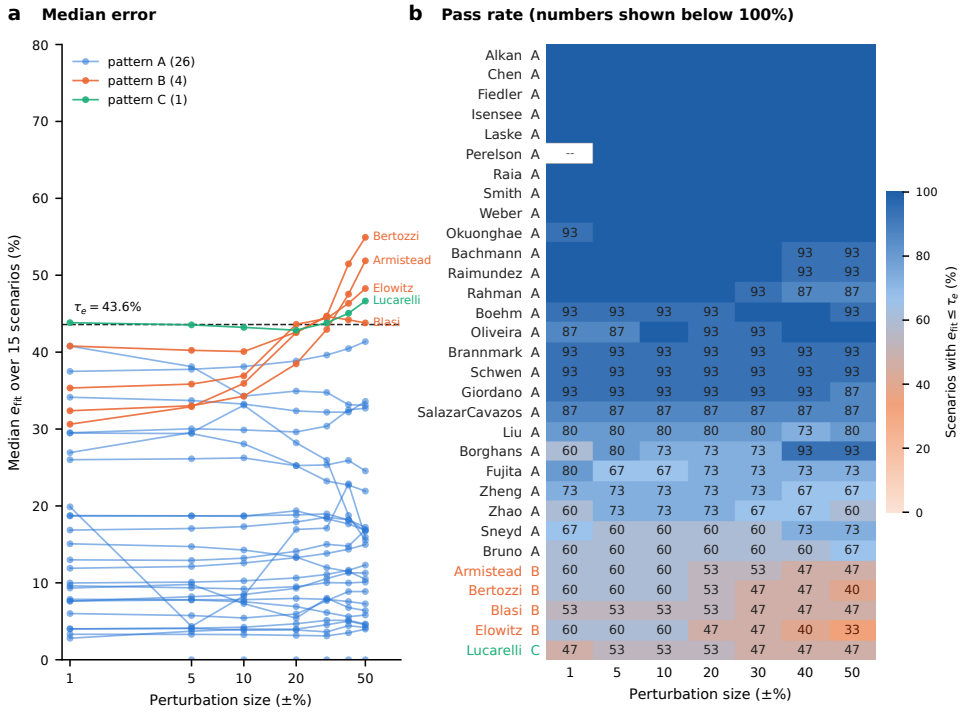


Fig. 5 Robustness of the selected subsets to the size of parameter changes. a, Median relative error after re-estimation (e_{fit}) over 15 random scenarios per level as a function of the perturbation size, for the final subset of each of the 31 admissible problems, coloured by robustness pattern (A, admissible at all levels; B, first failing at ±20% or more; C, irregular); dashed line, $\tau_e = 43.6\%$. **b**, Pass rate: percentage of the scenarios with $e_{\mathrm{fit}} \leq \tau_e$, with the value written when it is below 100%. The same fifteen random directions are used at every level. –, no valid scenario (Perelson at ±1%, where the output change is below the numerical noise floor).

Table 1 Per-problem results. p, number of parameters estimated in the PEtab problem, including noise parameters, which are kept fixed (Supplementary Table S3); m, number of measurements; $|\Theta_C|$, calibration parameters (always re-estimated); $|S|$, selected biological parameters; Stage, stage that produced the subset; e_{rel} and e_{fit}, median relative error over 15 scenarios at ±5% without and with re-estimation (admissibility: $e_{\mathrm{fit}} \leq 43.6\%$; italic, not admissible; *, subset selected with the first-order e_{fit}; the value shown is the nonlinear re-estimation of the final subset, except for Beer and Lang, which have no admissible subset (first-order value); †, not defined because the model has no solution on the measured horizon for 1% parameter changes); Pattern, behaviour of the median e_{fit} for perturbation sizes from ±1% to ±50% (A, admissible at all sizes; B, admissible up to at least ±10%, first failing at ±20% or more and not recovering; C, irregular; D, not admissible at ±10% or less; –, not applicable); Level, level of guarantee (1, $e_{\mathrm{fit}} \leq \tau_e$; 2, in addition median $\cos\Delta \geq c$; 3, in addition $R_{\mathrm{var}} \geq \tau_R$); Ident., selected parameters whose profile likelihood in the reduced problem exceeds $\Delta\chi^2 = 3.84$ on both sides within the bounds, out of $|S|$, and in parentheses those for which the hypotheses of the Lean identifiability theorem were satisfied at all evaluated profile points (a numerical check, not a certificate; Table 3); Test χ^2/n, mean squared noise-scaled residual on held-out experiments for the reduced and the full model, both re-estimated on the training experiments (values near 1 mean errors of the size of the noise; –, not evaluable or not run).

System	p	m	$\lvert\Theta_C\rvert$	$\lvert S\rvert$	Stage	e_{rel} (%)	e_{fit} (%)	Level	Pattern	Ident. (Lean)	Test χ^2/n red. / full
Alkan	44	1733	10	2	1	20	12.0*	3	A	2/2 (1)	2.75 / 3.37
Armistead	14	58	0	2	2	60	33.0	1	B	2/2 (2)	1.22 / 1.37
Bachmann	113	541	74	2	2	53	33.8*	1	A	2/2 (0)	4.77 / 7.35
Beer	72	27132	0	18	2	57	*50.1**	–	–	–	–
Bertozzi	8	22	0	2	2	58	32.9	2	B	2/2 (0)	–
Blasi	9	252	0	2	1	41	40.2	2	B	2/2 (2)	–
Boehm	9	48	0	2	2	60	9.4	2	A	2/2 (2)	2.57 / 11.22
Borghans	23	111	2	2	2	51	4.3	1	A	2/2 (0)	0.74 / 0.74
Brannmark	22	43	3	2	2	67	7.8	1	A	2/2 (0)	5.96 / 6,277
Bruno	13	77	0	5	1	46	37.8	2	A	5/5 (5)	0.91 / 208
Chen	155	120	3	2	2	53	8.1*	1	A	2/2 (0)	20.88 / 17.71
Crauste	12	21	0	–	2	–	*n.d.*†	–	–	–	–
Elowitz	21	58	2	2	1	65	35.8	1	B	2/2 (0)	1.50 / 3.89
Fiedler	22	72	8	2	2	58	14.7	1	A	2/2 (0)	11.21 / 9.32
Fujita	19	144	3	2	2	81	29.4	1	A	0/2 (0)	5.94 / 8.03
Giordano	50	313	0	2	2	108	5.8	2	A	0/2 (0)	<0.01 / <0.01
Isensee	46	687	11	2	1	52	11.3*	1	A	2/2 (0)	1.08 / 1.08
Lang	294	9600	16	13	2	70	*56.4**	–	–	–	–
Laske	13	42	2	2	1	14	10.1	2	A	2/2 (1)	0.76 / 1.87
Liu	9	100	0	2	2	47	26.1	2	A	2/2 (0)	0.95 / 1.02
Lucarelli	84	1755	0	10	2	60	43.585*	1	C	10/10 (4)	1.66 / 3.34
Okuonghae	16	92	0	2	2	75	4.1	2	A	1/2 (0)	4.7e4 / 1.2e5
Oliveira	12	120	0	2	2	134	12.9	2	A	1/2 (0)	912 / 8.1e4
Perelson	3	16	0	2	1	0	0.0	3	A	0/2 (0)	<0.01 / <0.01
Rahman	9	23	0	2	2	70	3.7	2	A	2/2 (2)	<0.01 / 0.03
Raia	39	205	5	2	2	32	3.3	2	A	2/2 (0)	8.22 / 5.10
Raimundez	136	627	79	2	2	64	18.6*	1	A	2/2 (0)	1.17 / 1.43
SalazarCavazos	6	18	0	2	2	41	18.7	2	A	2/2 (0)	102 / 124
Schwen	30	286	15	2	2	42	17.1	2	A	2/2 (0)	8.20 / 8.40
Smith	25	62	9	2	2	32	7.8	2	A	1/2 (0)	1.5e4 / 1.5e4
Sneyd	15	135	0	2	2	56	38.1	1	A	2/2 (2)	2.22 / 2.56
Weber	36	135	5	2	2	93	4.1	1	A	2/2 (0)	1.00 / 15.95
Zhao	28	82	0	3	2	70	29.6	2	A	3/3 (1)	0.35 / 0.35
Zheng	46	60	0	2	2	95	30.0	1	A	2/2 (2)	1.58 / 20.82

Table 2 Comparison of selection criteria and null baselines. All criteria choose the same number $k = |S|$ of biological parameters (except the proposed subsets selected on the training outputs, whose size the method chooses), from the training outputs except for the main proposed subsets (selected on all outputs), and are evaluated with the same protocol and the fixed scenario set. Admissible: problems in which re-estimating the subset reproduces the full model (median $e_{\mathrm{fit}} \leq \tau_e$) for parameter changes of ±5% and ±20%; for random subsets, median over the 20 draws; for calibration only, without the rule $|S| \geq 2$. Identifiable: selected parameters whose profile likelihood in the reduced problem is bounded on both sides, and in parentheses the problems in which every selected parameter is identifiable and the subset is admissible at ±5%, over the 27 problems common to all criteria (26 for the proposed subsets selected on the training outputs). Synthetic: fraction of the cases with signal in which $e_{\mathrm{pred}} \leq \tau_e$, for truths uniform within ±20% of θ_0 (281 cases common to all criteria, 24 problems, 20 seeds; difference from the proposed method with a 95% confidence interval from resampling problems), log-uniform within a factor of two (217 cases, 25 problems) and sparse (147 cases, 24 problems), ten seeds each. Real data: median ratio of the test χ^2/n to that of the proposed method (below 1, better than the proposed method) and Wilcoxon P over problems (unadjusted; exploratory). Proposed (all), (train.): subsets of the proposed method selected on all outputs and on the training outputs; Profile lik., profile likelihood; Calib., calibration only. Pivoted QR is the standard Golub–Klema–Stewart subset selection; the earlier variant is reported in the Supplementary Information. Per-problem values are given in the Supplementary Notes on the comparison of selection criteria.

	Admissible		Identifiable	Synthetic: $e_{\mathrm{pred}} \leq \tau_e$				Real data
				Proposed − other				χ^2 ratio
Criterion	±5%	±20%	Param. (both)	Uniform	(95% CI)	Log-unif.	Sparse	(P)
Proposed (all)	**28/28**	**27/28**	57/66 (**21**)	56%	–	50%	54%	–
Proposed (train.)	24/27	23/27	50/61 (16)	51%	+5 [0, 12]	48%	55%	1.00 (0.17)
Brun	26/28	25/28	56/66 (18)	55%	+1 [−1, 3]	50%	52%	1.00 (0.27)
Morris	23/28	23/28	53/66 (14)	48%	+8 [2, 16]	47%	46%	1.00 (0.46)
Profile lik.	18/27	18/27	57/66 (14)	39%	+17 [7, 28]	37%	40%	1.00 (0.44)
Pivoted QR	17/28	16/28	50/66 (7)	30%	+26 [12, 40]	24%	39%	1.00 (0.16)
Random (20)	16/28	16/28	–	–	–	–	–	1.00 (0.27)
Calib. ($k = 0$)	9/28	9/28	–	–	–	–	–	0.99 (0.003)

Table 3 What the formal verification guarantees. For each step of the method, the modules of the Lean development (numbers of Supplementary Table S4), the statement proved for any model that satisfies the stated hypotheses, the condition under which it applies to a numerical result, and the results of this study to which it applies. Levels of guarantee and identifiability per problem are given in Table 1; the formal statements are listed in Supplementary Table S7, the scope of each guarantee in Supplementary Table S5, and the correspondence between each claim of this paper and the Lean theorem that supports it in Supplementary Table S43.

Step of the method (modules)	Statement proved in Lean	Applies to a result when	In this study
Ranking by energy (1, 2, 11, 13)	The first-order output error when parameters are fixed is bounded by the discarded sensitivity energy; the bound transfers to transformed outputs	$R_{\mathrm{var}}(S) \geq \tau_R$ and the model is verified	None: Alkan and Perelson satisfy $R_{\mathrm{var}} \geq \tau_R$, but their models are not verified
Selection with κ and VIF (7–9)	The greedy subset satisfies $\kappa \leq \tau_\kappa$ and $\mathrm{VIF} \leq \tau_{\mathrm{VIF}}$; discarding a rejected candidate permanently is correct; the linearised least-squares estimates are unique and stable	The computed κ and VIF are below the thresholds (unit-length columns for the least-squares statement)	Selection: all 31 subsets (centred columns); least squares: 30 for θ_S with Θ_C fixed (not Rahman); with Θ_C re-estimated, not satisfied in 10
Admissibility threshold (3, 4, 20–22)	Theorem 1: rescaling an output change with cosine c leaves a relative error $\sqrt{1-c^2} = \tau_e$	Median $\cos\Delta \geq c$	Level 2 or 3: 17 subsets
Re-estimation (5, 12)	More re-estimated parameters never increase the fitting error; $e_{\mathrm{fit}} \leq e_{\mathrm{rel}}$; the first-order e_{fit} differs from the exact one by a second-order term	Exact minimum; the computed e_{fit} (local optimum) bounds it from above, so admissible as computed implies admissible exactly, not conversely	25 by nonlinear re-estimation; 6 first order, then checked nonlinearly at ±5%; Beer and Lang first order; Crauste not defined
Backward elimination (6)	Theorem 2: the result is admissible and no parameter can be removed	Relative to the implemented test, including $\lvert S\rvert \geq 2$	All 31 subsets (non-trivial only for Bruno, Lucarelli and Zhao)
Model hypotheses (14–18, 23–31)	The solution exists on the horizon, stays non-negative and is differentiable in the parameters	Checked by the Lean kernel at the nominal parameters and base condition (first segment of piecewise models)	28 models (27 of the 34 problems); Crauste only for $T \leq 1$; the selected initial conditions of Bertozzi and Bruno are not parameters of the Lean files
Identifiability (32)	If t_j^* is smaller than the distance to both bounds, θ_j is practically identifiable	I_j, γ and ε computed at the profile points satisfy the inequality	Hypotheses satisfied at the evaluated profile points (not a certificate): 24 parameters; all of S in 7 problems

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 2SupplementaryInformation.pdf