Supplementary Information

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

Mónica Miranda R., Carlos Salazar, César Martín, Adriana Aguirre

Contents

Supplementary Note: Selected subsets of the benchmark problems **5**

Supplementary Note: Notation of the method **7**

Supplementary Note: Prediction of held-out experiments **8**

Supplementary Note: Logical sequence of the formal proofs **10**

Modules of the formalisation . . . 11
Setting and goal . . . 14
Steps 1–13. Finite-dimensional core of the method . . . 16
Steps 14–18. Ordinary differential equations . . . 17
Steps 19–22. Certificates for ODE models . . . 17
Steps 23–31. Hypotheses of the benchmark models . . . 18
Connection with the numerical results and what the proofs do not cover . . . 19

Supplementary Note: Statements of the formal theorems **19**

Steps 1–13. Finite-dimensional core of the method (fixed sensitivity matrix) . . . 21
Steps 14–18. Ordinary differential equations: existence and sensitivities . . . 28
Steps 19–22. Certificates for ODE models . . . 31
Steps 23–31. Hypotheses of the benchmark models (kinetic models and reaction networks) . . . 33
Step 32. Identifiability of the reduced problem . . . 40

Supplementary Note: Verified benchmark models **41**

Agreement of the translation with the SBML models . . . 43
A complete example: Boehm_JProteomeRes2014 . . . 43
Benchmark problems without a verified model . . . 44
Data of each model . . . 45

Supplementary Note: Reproducibility package **57**

Contents of the package . . . 57
Three levels of checking . . . 58
What each check of the formal verification establishes . . . 58
Versions . . . 59
Steps for the formal verification . . . 59
First-order re-estimation . . . 59

Supplementary Note: Verification of the simulator **60**

Supplementary Note: Robustness to the size of parameter changes **61**

Random scenarios . . . 61
Errors in each scenario . . . 62
Summaries over the scenarios . . . 62

Results . . . 63
Independent directions at each size . . . 66

Supplementary Note: Random scenarios and large problems **68**
Construction of the random scenarios . . . 68
Computation of the problems with expensive simulations . . . 69
Agreement between machines . . . 69
Dependence of the greedy search on its path: Lang . . . 70
Settings variants of Raimundez . . . 70

Supplementary Note: Synthetic prediction experiment **71**
Design . . . 71
Results . . . 71
Size of the signal and other amplitudes . . . 73

Supplementary Note: Single-parameter subsets **75**
Design . . . 75
Results . . . 75

Supplementary Note: Minimum subset size: one parameter or two **76**

Supplementary Note: Identifiability of the selected parameters **78**
Question . . . 78
Design . . . 78
Mathematical derivation of the identifiability condition . . . 78
Results . . . 82
Subsets that are both sufficient and identifiable . . . 83
Scope . . . 84

Supplementary Note: Nested validation without the θ_0 anchor **88**
Motivation . . . 88
Design . . . 88
Results . . . 89

Supplementary Note: Pivoted QR selection **91**
Motivation . . . 91
Method: QR-pivoted selection . . . 91
Results . . . 92
Selected parameters . . . 93

Supplementary Note: Profile likelihood **94**
Purpose . . . 94
Design . . . 94
Results . . . 95

Supplementary Note: Brun and Morris selections **97**
Purpose . . . 97
Design . . . 97

Implementation differences between the criteria . . . 98

Supplementary Note: Comparison on real data **99**

Supplementary Note: Admissibility of the subsets of other criteria **101**
Purpose . . . 101
Design . . . 101
Results . . . 102

Supplementary Note: Comparison of selection criteria on synthetic data **103**
Design . . . 103
Results . . . 104

Supplementary Note: Sensitivity to the method settings **106**
Reproduction error in units of the noise . . . 107

Supplementary Note: Computational cost **109**
Time of the complete method and comparison with profile likelihood . . . 109

Supplementary Note: Computing time of the analyses **111**
How the times were measured . . . 111
What the times show . . . 111

Supplementary Note: System-by-system summary **118**
Alkan_SciSignal2018 . . . 118
Armistead_CellDeathDis2024 . . . 118
Bachmann_MSB2011 . . . 119
Beer_MolBioSystems2014 . . . 119
Bertozzi_PNAS2020 . . . 119
Blasi_CellSystems2016 . . . 120
Boehm_JProteomeRes2014 . . . 120
Borghans_BiophysChem1997 . . . 120
Brannmark_JBC2010 . . . 121
Bruno_JExpBot2016 . . . 121
Chen_MSB2009 . . . 121
Crauste_CellSystems2017 . . . 122
Elowitz_Nature2000 . . . 122
Fiedler_BMCSystBiol2016 . . . 122
Fujita_SciSignal2010 . . . 123
Giordano_Nature2020 . . . 123
Isensee_JCB2018 . . . 123
Lang_PLOSComputBiol2024 . . . 124
Laske_PLOSComputBiol2019 . . . 124
Liu_IFACPapersOnLine2025 . . . 125
Lucarelli_CellSystems2018 . . . 125
Okuonghae_ChaosSolitonsFractals2020 . . . 125
Oliveira_NatCommun2021 . . . 126

Perelson_Science1996 . . . 126
Rahman_MBS2016 . . . 126
Raia_CancerResearch2011 . . . 127
Raimundez_PCB2020 . . . 127
SalazarCavazos_MBoC2020 . . . 127
Schwen_PONE2014 . . . 128
Smith_BMCSystBiol2013 . . . 128
Sneyd_PNAS2002 . . . 128
Weber_BMC2015 . . . 129
Zhao_QuantBiol2020 . . . 129
Zheng_PNAS2012 . . . 129

Supplementary Note: Correspondence between claims and Lean theorems 130

Supplementary Note: Conditioning with unit-length columns 131

Supplementary Note: Nonlinear check of the first-order problems 132

Supplementary Note: Selection on the training outputs, null controls and standard pivoted QR 132

Supplementary Note: Identifiability over the admissible set 134

Supplementary Note: Identifiability of the subsets of other criteria 136

Supplementary Note: Admissibility with the calibration parameters fixed 139
Question . . . 139
Design . . . 139
Results . . . 139
Conclusion . . . 140
Programs and reproduction . . . 140
Computing time . . . 140

Supplementary Note: Concluding remarks 143

Supplementary Note: Selected subsets of the benchmark problems

Supplementary Table S1 gives the final subset S for each problem (after backward elimination), the parameters removed by backward elimination, and the quantities that characterise the subset at ±5%: condition number κ and largest VIF of the selected sensitivity columns (below the collinearity thresholds of the selection, centred columns: both ≤ 10), fraction of sensitivity energy kept R_{var}, median $\cos\Delta$ over the 15 scenarios, and median e_{fit} (admissibility: $\leq 43.6\%$). The last column gives the level of guarantee of each admissible subset: level 1, $e_{\text{fit}} \leq \tau_e$ (the subset reproduces the full model after re-estimation, checked numerically); level 2, in addition $\cos\Delta \geq 0.9$ (by Theorem 1 a single rescaling of the reduced response already achieves the threshold); level 3, in addition $R_{\text{var}} \geq 0.89$ (the first-order error bound G1 of Supplementary Note *Logical sequence of the formal proofs* also applies). Levels 2 and 3 are stronger guarantees, not requirements of the method; 14, 15 and 2 problems reach levels 1, 2 and 3. For the problems marked * the subset was selected with the first-order e_{fit} and the value shown is the nonlinear re-estimation of the final subset (for Beer and Lang, the first-order value). Calibration parameters, which are always re-estimated, are not listed. Crauste, Beer and Lang have no admissible subset; the subset at which the method stopped is shown. For Crauste it is not meaningful: about half of the perturbed simulations have no solution on the measured horizon (finite-time growth without bound), so the sensitivity matrix is incomplete.

Supplementary Table S1. Selected biological parameters of each benchmark problem and characteristics of the subset.

System	Selected biological parameters S	Removed by backward elimination	κ	VIF	R_{var} (%)	$\cos\Delta$	e_{fit} (%)	Level
Alkan	kt, k_rep_nhej	–	1.3	1.1	94	0.980	12.0*	3
Armistead	k_d, k00	–	1.0	1.0	58	0.807	33.0	1
Bachmann	CISRNADelay, JAK2ActEpo	–	2.8	2.5	36	0.852	33.8*	1
Beer	kdegi_typeIDT1, kdegi_typeIDT3, ksyn_typeIDT3, kdim_typeIDT1, beta_typeIDT3_ExpID2, beta_typeIDT3_ExpID3, beta_typeIDT3_ExpID5, beta_typeIDT3_ExpID1, beta_typeIDT3_ExpID4, init_Bac, beta_typeIDT1_ExpID6, beta_typeIDT1_ExpID5, beta_typeIDT1_ExpID1, beta_typeIDT1_ExpID4, beta_typeIDT1_ExpID3, kdim_typeIDwt, ksyn_typeIDT5, kdim_typeIDT5	–	7.2	10.0	71	0.823	50.1*	–
Bertozzi	R0_NY, I0_NY	–	3.9	4.3	78	0.917	32.9	2
Blasi	a_basal, a_k16_k12k16	a_k12k16_k8k12k16, a_k5k12_k5k8k12, a_k8k12k16_4ac	1.8	1.4	79	0.914	40.2	2
Boehm	k_imp_hetero, k_phos	–	1.0	1.0	68	0.927	9.4	2
Borghans	Kz, K_par	–	2.7	2.4	21	0.860	4.3	1
Brannmark	k1f, km2	–	1.9	1.4	32	0.854	7.8	1
Bruno	kb1, init_bcar2, init_zea_1, init_bcar1, init_bcry_1	szea, init_b10_1	2.3	1.6	75	0.904	37.8	2

System	Selected biological parameters S	Removed by backward elimination	κ	VIF	R_{var} (%)	$\cos\Delta$	e_{fit} (%)	Level
Chen	k106, k44	–	1.0	1.0	25	0.899	8.1*	1
Crauste	rho_E	–	1.0	1.0	0	0.535	n.d.	–
Elowitz	n_Hill, tau_mRNA	tau_prot, tau_prot_GFP, eff	2.9	2.6	70	0.887	35.8	1
Fiedler	tau2, k6	–	1.5	1.2	42	0.875	14.7	1
Fujita	init_AKT, reaction_6_k1	–	2.3	1.9	36	0.765	29.4	1
Giordano	alpha_4, alpha_28	–	3.7	3.9	74	0.967	5.8	2
Isensee	kf_RIIp_C_2__RII_C_2, xi_kf_RII_C_2__RII_2	ks_AC_cAMP, KD_cAMP, xi_AC_cAMP_Fsk, kf_RIIp_cAMP_C_2__RIIp_2, kf_RII_2__RII_C_2	1.0	1.0	68	0.866	11.3*	1
Lang	kDeE2f1, kDeCe, kDeCa2, kPhCAEB, kDpCdh, kImImSKP2_1, kImImCDKN1A_CCNE_A_B_Ser130p__1, ic_wt_Nuc__CDKN1B_CCNE_A_Thr187p_, ic_wt_Cyt__CDKN1B_CCNE_A_Thr187u_, ic_wt_Nuc__SKP2, ic_wt_Cyt__CCNE_CDKN1A_B_, ic_wt_Cyt__SKP2, ic_wt_Cyt__CCNE_CDKN1A_B_1__CDKN1A_CCNE_A_B_1_Ser130u_	–	8.7	10.0	17	0.726	56.4*	–
Laske	k_imp, k_syn_R_M	k_bind_M1	1.4	1.1	84	0.990	10.1	2
Liu	kappa, phi	–	1.3	1.1	81	0.923	26.1	2
Lucarelli	init_Rec, S2tot, S3tot, S_dephosphos, S_phos, geneF_turn, k_234, geneK_turn, S4tot, geneH_turn	–	6.5	9.8	64	0.811	43.585*	1
Okuonghae	transmission_rate_effective, gamma_i	–	1.3	1.1	53	0.998	4.1	2
Oliveira	beta_0, h_hosp_rate	–	1.2	1.0	47	0.950	12.9	2
Perelson	c, delta	–	1.9	1.5	100	1.000	0.0	3
Rahman	infected_moderate_transmission_rate, infected_normal_worsen_rate	–	5.9	9.3	60	0.967	3.7	2
Raia	SOCS3mRNA_production, init_Rec_i	–	5.5	8.1	69	0.998	3.3	2
Raimundez	kdeg_membran__MKN1, ka_MAPK__MKN1	–	1.3	1.1	42	0.786	18.6*	1
SalazarCavazos	ratio_kpkd_Y1068__FREE, ratio_kpkd_YN__FREE	–	5.4	7.9	84	0.973	18.7	2
Schwen	ini_R1, ini_R2fold	–	3.1	3.0	46	0.930	17.1	2
Smith	k8, kminus7b	–	3.2	3.1	41	0.960	7.8	2
Sneyd	l4, l2	–	1.1	1.0	44	0.877	38.1	1
Weber	s12, a21	–	3.7	4.0	25	0.407	4.1	1

System	Selected biological parameters S	Removed by backward elimination	κ	VIF	R_{var} (%)	$\cos\Delta$	e_{fit} (%)	Level
Zhao	R_Stage_I_China, R_Stage_II_Wuhan, gamma_1_Stage_I_Hubei	–	1.2	1.0	69	0.902	29.6	2
Zheng	k01_02, k13_12	–	1.4	1.1	35	0.568	30.0	1

Supplementary Note: Notation of the method

Supplementary Table S2 defines the symbols of the method used in the main text and in this Supplementary Information, with the values used in the analyses. Supplementary Table S3, at the end of this note, gives the number of parameters of every problem in each category and the number of free parameters in each analysis, so that the counts reported in different tables (for example $p = 9$ for Boehm in Table 1 of the main text and six biological parameters in its profile likelihood) can be reconciled.

Supplementary Table S2. Notation of the method.

Symbol	Meaning	Value used
$x(t;\theta) \in \mathbb{R}^n$	state trajectory of the model $\dot{x} = f(x,\theta)$	–
$\theta \in \mathbb{R}^p$, θ_0	parameter vector; nominal (published) value	PEtab nominal values
Θ_B, Θ_C	biological parameters (in the SBML model) and calibration parameters (observation scalings and offsets)	from the PEtab tables
$y_i(\theta) = h(g_i(x(t_i),\theta))/\sigma_i$	i-th scaled output: observable g_i, transformation h (lin, log, $\log_{10}$), noise s.d. σ_i, at the condition and time of measurement i	PEtab measurement table
$J = \partial y/\partial \log\theta \in \mathbb{R}^{m\times p}$	relative sensitivity matrix (central differences, step δ)	$\delta = 0.01$
$E_j = \|J_{:,j}\|^2$	sensitivity energy of parameter j	–
$R_{\text{var}}(S)$	fraction of the energy of biological parameters captured by S	$\tau_R = 0.89$
s_j	finite-perturbation effect of parameter j (stage-2 ordering)	$\eta = 0.1$
$\kappa(S)$, VIF(S)	condition number and largest variance inflation factor of the normalised columns of J_S	$\tau_\kappa = \tau_{\text{VIF}} = 10$
Δy_{full}, Δy_S	output change when all parameters, or only those in $S \cup \Theta_C$, take their scenario values	$N = 15$, $\varepsilon = 0.05$
$\cos\Delta$	cosine similarity between Δy_{full} and Δy_S	–
e_{rel}	relative error without re-estimation	–
e_{fit}	relative error after re-estimating θ_S and θ_C	–
$\tau_e = \sqrt{1-c^2}$	admissibility threshold on the median e_{fit}	$c = 0.9$, $\tau_e = 0.436$
$k_{\min}$	minimum subset size	2
χ^2/n	mean squared noise-scaled residual on test data	–
e_{pred}	relative prediction error with respect to a known truth (synthetic experiment)	–

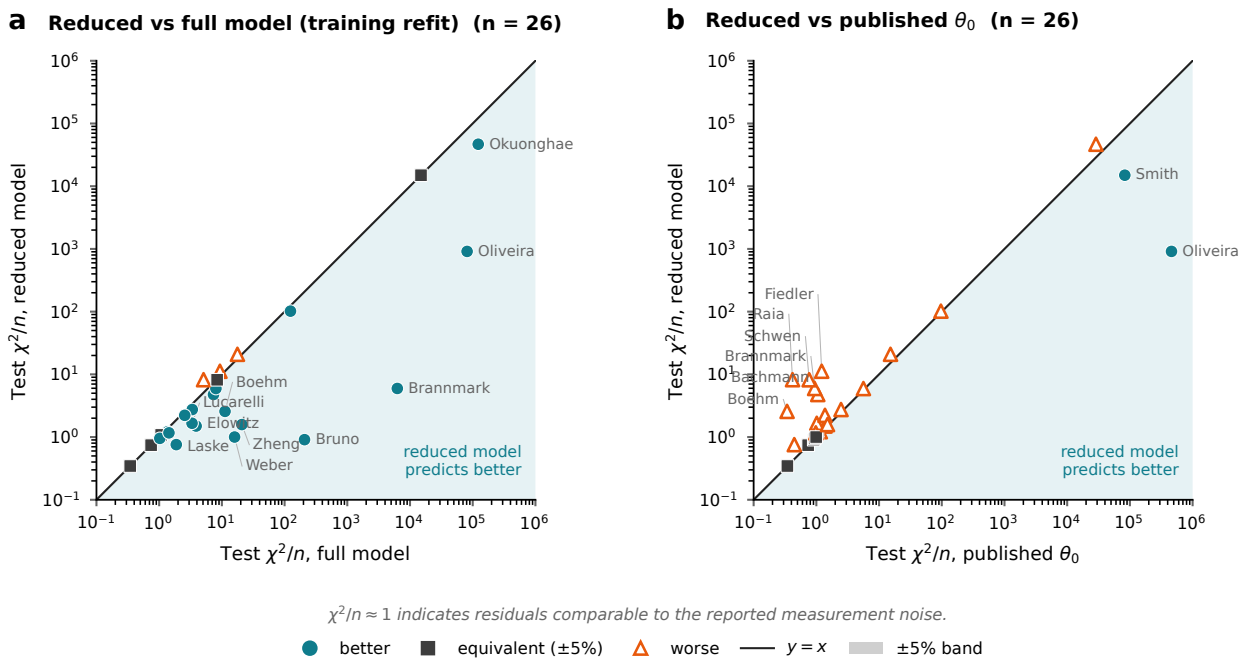
Supplementary Table S3. Parameter counts of every problem. p: parameters estimated in the PEtab problem (Table 1 of the main text), split into biological parameters (of the SBML model or of the condition table; candidates for selection), calibration parameters Θ_C (scalings and offsets of the observables, always re-estimated) and noise parameters (they enter only the standard deviation σ_i, which is evaluated at θ_0 and kept fixed). Full-model profile: biological / free parameters in the profile likelihood of the full model on the training data, which excludes the parameters that act only in test conditions (so the numbers can be smaller than those of the PEtab problem); n.e., not evaluated (not run for Chen and Raimundez because of its computational cost, or no training/test split). Reduced problem: free parameters in the reduced problem, $|S| + |\Theta_C|$.

Problem	p	Biological	Calibration	Noise	Full-model profile	Reduced problem
Alkan	44	34	10	0	34 / 44	2 + 10
Armistead	14	10	0	4	10 / 10	2 + 0
Bachmann	113	27	74	12	27 / 90	2 + 74
Beer	72	70	0	2	–	–
Bertozzi	8	6	0	2	3 / 3	2 + 0
Blasi	9	8	0	1	n.e.	2 + 0
Boehm	9	6	0	3	6 / 6	2 + 0
Borghans	23	20	2	1	20 / 22	2 + 2
Brannmark	22	15	3	4	15 / 16	2 + 3
Bruno	13	13	0	0	10 / 10	5 + 0
Chen	155	152	3	0	n.e.	2 + 3
Crauste	12	12	0	0	–	–
Elowitz	21	18	2	1	18 / 20	2 + 2
Fiedler	22	12	8	2	12 / 20	2 + 8
Fujita	19	16	3	0	16 / 19	2 + 3
Giordano	50	43	0	7	43 / 43	2 + 0
Isensee	46	32	11	3	32 / 43	2 + 11
Lang	294	261	16	17	–	–
Laske	13	6	2	5	6 / 7	2 + 2
Liu	9	7	0	2	7 / 7	2 + 0
Lucarelli	84	72	0	12	72 / 72	10 + 0
Okuonghae	16	14	0	2	14 / 14	2 + 0
Oliveira	12	12	0	0	12 / 12	2 + 0
Perelson	3	2	0	1	2 / 2	2 + 0
Rahman	9	9	0	0	9 / 9	2 + 0
Raia	39	18	5	16	18 / 23	2 + 5
Raimundez	136	57	79	0	n.e.	2 + 79
Salazar-Cavazos	6	6	0	0	6 / 6	2 + 0
Schwen	30	13	15	2	13 / 28	2 + 15
Smith	25	16	9	0	16 / 25	2 + 9
Sneyd	15	14	0	1	14 / 14	2 + 0
Weber	36	26	5	5	26 / 30	2 + 5
Zhao	28	21	0	7	21 / 21	3 + 0
Zheng	46	45	0	1	45 / 45	2 + 0

Supplementary Note: Prediction of held-out experiments

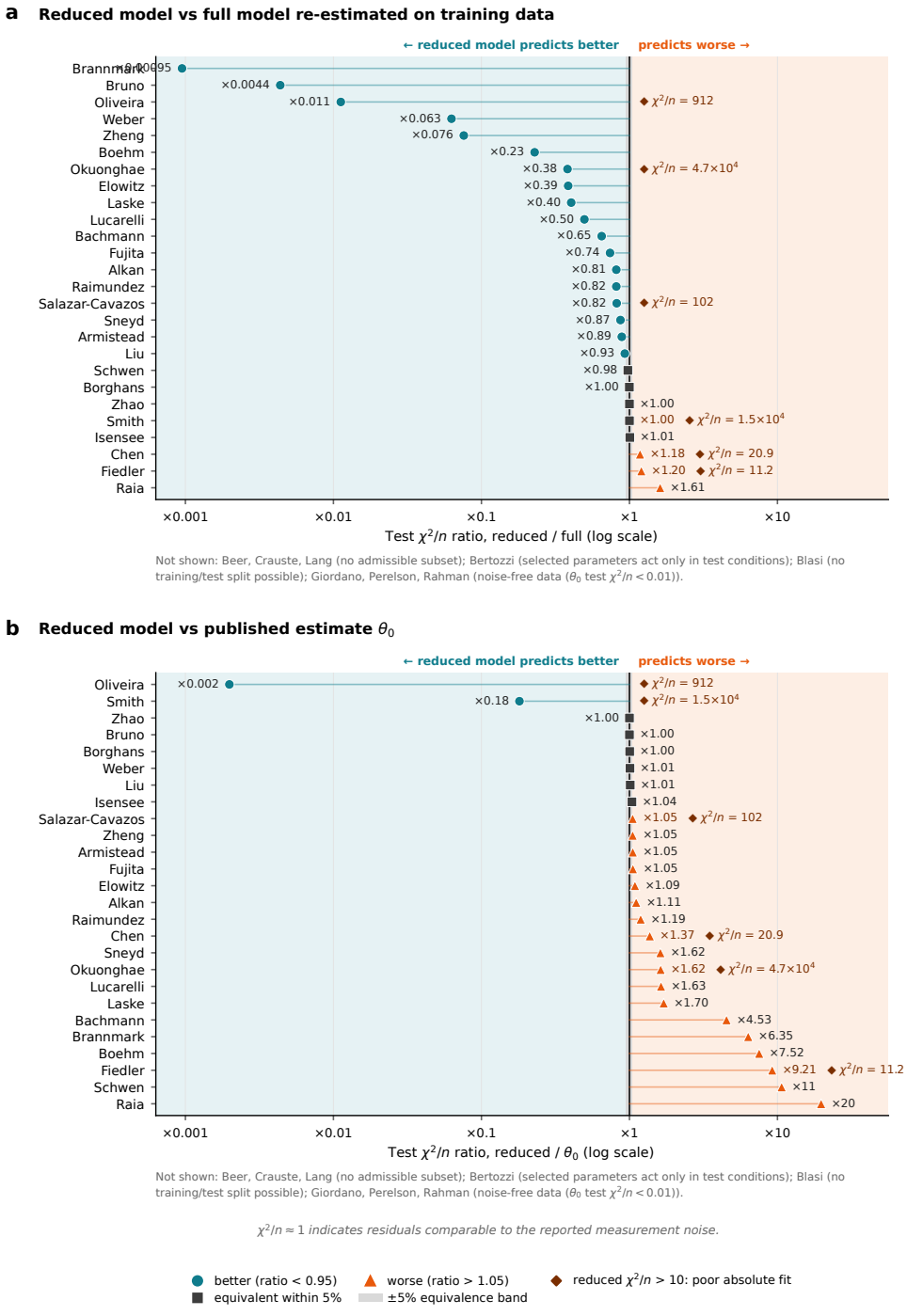
Figure 3 of the main text compares the test χ^2/n of the reduced model with that of two references through their ratio. Supplementary Fig. S1 shows the same values as a scatter plot, which also displays their absolute level: points near $\chi^2/n = 1$ predict at the level of the measurement noise, whereas points far above it (for example Okuonghae, Oliveira and Smith) are poorly fitted by every model, so that a favourable ratio there does not imply a good prediction. The values of every problem, the problems excluded from each panel and the reasons are listed in `paper/scirep/figure3/figure3_heldout_data.csv` of the repository, together with the script that produces both figures and checks every plotted point

against that table.



Supplementary Figure S1. Prediction of held-out experiments, absolute values. Test χ^2/n of the reduced model against **a**, the full model re-estimated on the same training data ($n = 26$) and **b**, the published estimate θ_0 ($n = 26$), on logarithmic axes with the same scale. Solid line, $y = x$; grey band, equivalence within ±5%; shaded region, the reduced model predicts better. Circles, better; squares, equivalent; open triangles, worse. Names are shown for points away from the diagonal. The excluded problems are those of Figure 3.

Supplementary Figure S2 shows the ratios of the reduced to the reference test χ^2/n, including the comparison with the published estimate θ_0. Because θ_0 was obtained using all data including the test experiments, it is an optimistic reference: it predicted the test data better than the reduced model in 20 of 29 problems (Supplementary Table S27).

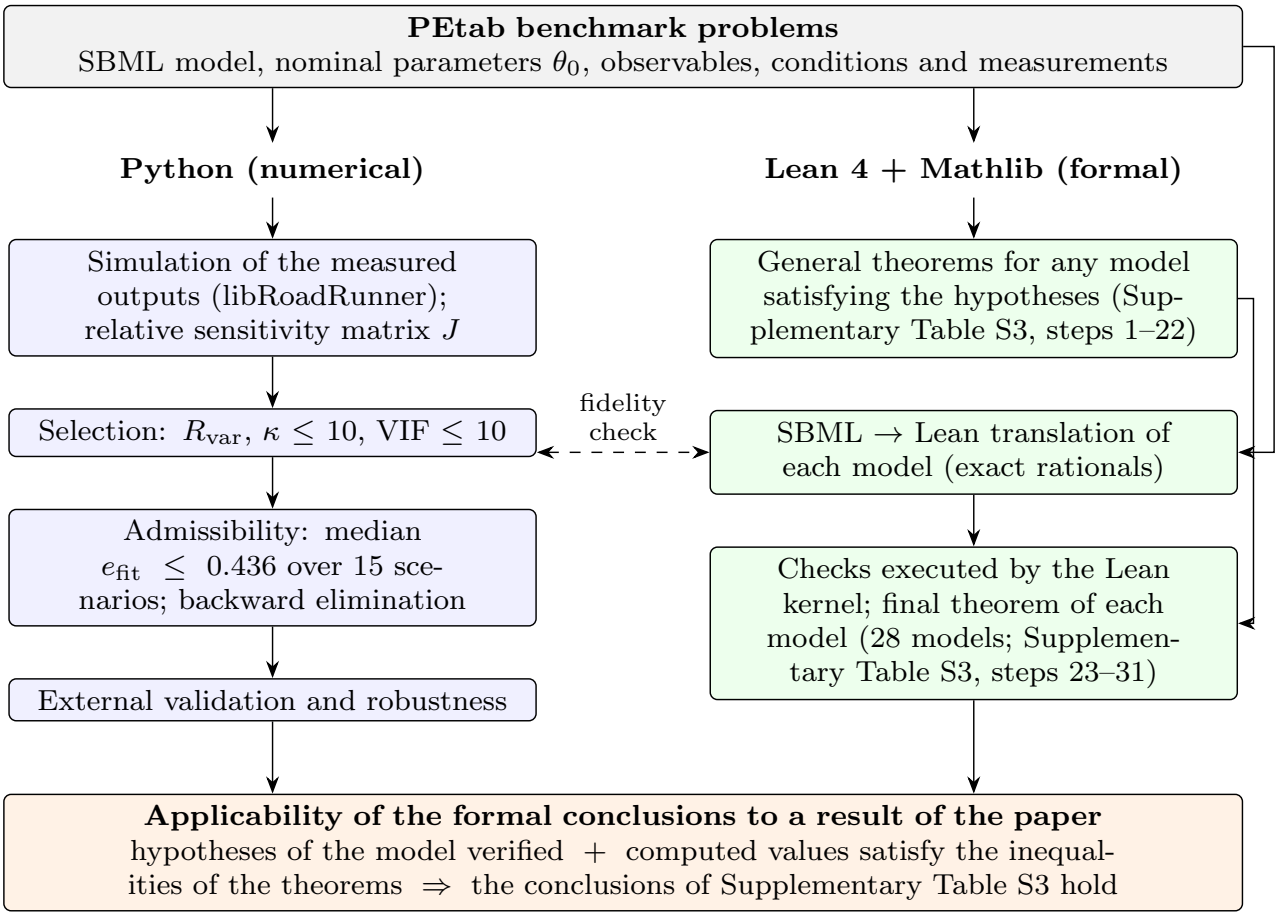


Supplementary Figure S2. Prediction of held-out experiments with real data, ratios. Ratio of the test χ^2/n of the reduced model to that of **a**, the full model re-estimated on the same training data ($n = 26$ problems), and **b**, the published estimate θ_0, obtained using all data including the test experiments ($n = 26$). Left of the vertical line (circles) the reduced model predicts better; grey band, equivalence within 5% (squares); triangles, worse; ⧫ flags a reduced test χ^2/n above 10. Problems without a training/test split (Bertozzi, Blasi) or with noise-free data (Giordano, Perelson, Rahman) are not shown.

Supplementary Note: Logical sequence of the formal proofs

This note explains, in mathematical terms, the chain of results that the Lean code proves, in the order in which they build on each other. The order is that of the import graph of `lean_proofs/`: step k only

uses steps with a smaller number (Supplementary Table S4, column 'Uses'). Lean names of the main results are given in `typewriter` font; the statement and meaning of every theorem are in Supplementary Note *Statements of the formal theorems*, and the notation in Supplementary Table S6. Supplementary Fig. S3 shows how the formal and the numerical parts of the work fit together, and Supplementary Table S5 states the scope of each formal guarantee.



Supplementary Figure S3. Architecture of the formal and numerical verification. Python (left) performs all computations on the PEtab problems: simulation, sensitivities, selection, admissibility, validation and robustness. Lean 4 with Mathlib (right) proves the general theorems behind each step of the method (Supplementary Table S4) and, for 28 benchmark models translated from SBML, checks with the Lean kernel that the hypotheses on the dynamics hold. The translation of each model is compared independently with the SBML model (dashed arrow; Supplementary Note *Verified benchmark models*). A formal conclusion applies to a result of the paper (main-text table on what the formal verification guarantees) when the hypotheses of the model are verified and the numerical quantities computed in Python satisfy the inequalities that the theorem requires (bottom); Supplementary Table S5 gives, for each guarantee, what it establishes and what it does not.

Modules of the formalisation

All formal proofs are collected in the folder `lean_proofs/` of the reproducibility package, a self-contained Lean 4 project (Lean 4.23.0, Mathlib v4.23.0) that compiles with a single command (`./verify.sh`). It contains 32 theory modules with 309 theorems and lemmas, and 28 benchmark models translated from SBML. No file contains `sorry` or `axiom`, and `#print axioms` reports only `propext`, `Classical.choice` and `Quot.sound` for every theorem. Supplementary Table S4 lists the modules in the logical order of the proofs: # is the step number, and every module only uses modules with a smaller number (column 'Uses'). n is the number of theorems and lemmas in the module (its size, not an index). Supplementary

Note *Statements of the formal theorems* gives the statement and meaning of every theorem and lemma; the file `THEOREMS.md` in the folder lists them as well.

Supplementary Table S4. Modules of the Lean formalisation in the logical order of the proofs.

#	Module	n	Uses	Content	Main theorems
Steps 1–13. Finite-dimensional core of the method (fixed sensitivity matrix)					
1	`FirstOrderErrorBound`	12	–	Energy $E_j = (J^\top J)_{jj}$, R_{var} and first-order output error when parameters are fixed	`total_eq_kept_add_removed`, `first_order_error_from_Rvar`
2	`TrajectoryErrorBound`	9	1	From the Jacobian to the trajectory: error bound when discarded parameters are fixed	`sqrt_removedParamSq_le`
3	`ScaledError`	5	–	Theorem 1: the minimum relative error after rescaling equals $\sqrt{1-\cos^2\Delta}$	`min_rel_error_eq`
4	`CosineCertificate`	5	2	Non-asymptotic bounds for finite perturbations; per-scenario certificate of $\cos\Delta$	`finite_trajectory_error`, `finite_cos_certificate`
5	`RefitMonotone`	5	–	Re-estimating more parameters never increases the fitting error; $e_{\text{fit}} \le e_{\text{rel}}$	`fitErr_anti`, `fitErr_le_erel`, `admissible_mono`
6	`Poda`	4	–	Theorem 2: backward elimination returns an admissible subset from which no parameter can be removed; inclusion-minimal for a monotone criterion	`prune_spec`, `prune_inclusion_minimal`
7	`Identifiability Conditioning`	10	–	κ and VIF bounds imply unique and stable least-squares estimates	`ls_unique`, `ls_param_kappa_stability`, `ls_vif_inverse_stability`
8	`SpectralConditioning`	22	7	κ and VIF exactly as computed numerically (spectral theorem, matrix inverse)	–
9	`GreedySelection`	7	7	Greedy selection returns a subset with $\kappa \le \tau_\kappa$ and $\text{VIF} \le \tau_{\text{VIF}}$; permanent discard is correct	`greedy_kappa_vif_guarantee`
10	`PatternMechanisms`	4	2, 9	Opposing effects imply rejection by κ and non-identifiability; zero discarded energy implies exact first-order reduction	`not_okKappa_of_kernel`, `exact_reduction_of_zero_energy`
11	`NumericalRobustness`	6	4, 9	Finite-difference error; R_{var}, κ and VIF certified from approximate values with margins	`fd_column_error`, `rvar_certified_from_approx`
12	`LinearizedFit`	3	3	The first-order e_{fit} differs from the exact one by a second-order term	`lin_le_scaled`, `fit_lower_of_lin`
13	`OutputComposition`	4	–	Error bounds transfer from states to measured outputs, including log and $\log_{10}$	`output_uniform_bound`, `log10_lipschitz`
Steps 14–18. Ordinary differential equations: existence and sensitivities					
14	`GlobalLipschitzODE`	3	–	Global existence on $[a, b]$ for globally Lipschitz vector fields	–
15	`ParamDiffODE`	8	14	Solutions depend differentiably on parameters and initial values; the derivative solves the variational equation	`hasFDerivAt_solution_param_init`
16	`LocalExistence`	9	15	Solutions exist for parameters near θ_0 when the nominal solution exists	–

#	Module	n	Uses	Content	Main theorems
17	`SensitivityLipschitz`	8	–	Explicit Lipschitz constants of trajectories and sensitivities in θ (Grönwall)	`sens_lipschitz, traj_lipschitz`
18	`EventSystems`	1	16	Systems with events at fixed times: existence and differentiability across segments	–
Steps 19–22. Certificates for ODE models					
19	`CertifiedODEReduction`	1	2, 15	Local certificate: $R_{\mathrm{var}}(S) \geq r$ bounds the trajectory error of the reduced model	`certified_parameter_reduction`
20	`CertifiedFiniteODE`	5	4, 17, 19	Certificate of $\cos\Delta$ for the ODE with an explicitly bounded Lipschitz constant	`certified_cos_ode`
21	`ExactLinear Certificate`	13	20	Certificates with the exact linear error; local version without Lipschitz constants	`finite_cos_certificate_exact, certified_cos_ode_exact`
22	`NominalCertificate`	1	16, 21	Local $\cos\Delta$ certificate from the nominal trajectory only	–
Steps 23–31. Hypotheses of the benchmark models (kinetic models and reaction networks)					
23	`KineticRegularity`	11	22	Rate laws (mass action, Michaelis–Menten, Hill, ...) define smooth vector fields on their domain	–
24	`PositivityInvariance`	15	23	Quasi-positive models keep concentrations non-negative	–
25	`GlobalExistence`	11	14, 24	Global existence under linear growth on the non-negative orthant	–
26	`StrictExistence`	12	25	Global existence with strictly positive concentrations	`exists_global_solution_strict`
27	`KineticCheck`	19	18, 24	Verified automatic checkers for translated rate laws	–
28	`KineticNetwork`	18	18, 25, 27	Reaction networks: complete check including global existence	`network_final`
29	`RiccatiNetwork`	9	28	Existence on a finite horizon under quadratic growth (Riccati bound; used for Crauste)	`riccati_final`
30	`StrictNetwork`	46	26, 28	Networks with strictly positive domains (e.g. $\beta SI/N$, Hill with real exponent)	`strict_final_init`
31	`IntervalInit`	5	30	Parameter-dependent initial values certified with rational interval arithmetic	`strict_final_initI`
Step 32. Identifiability of the reduced problem					
32	`Profile Identifiability`	18	8	Practical identifiability of the reduced problem: if the guaranteed half-width t^* from the profiled information is smaller than the distance to the bounds, the profile likelihood exceeds the threshold on both sides	`practically_identifiable_grad, practically_identifiable_rel_tail, local_injective`
	Total	309			

Verified benchmark models (28). Armistead, Bachmann, Bertozzi, Blasi, Boehm, Borghans, Brannmark, Bruno, Chen (501 states, in four parts), Crauste (horizon $T \leq 1$), Elowitz, Fiedler, Froehlich (1,228 states, in twelve parts), Fujita, Giordano, Isensee, Lang, Laske, Lucarelli, Okuonghae, Rahman, Raia, Raimundez, SalazarCavazos, Sneyd, Weber, Zhao and Zheng (folder `lean_proofs/Models/`). For each model, the final theorem establishes that the vector field is smooth on its domain, that the solution exists on the simulated horizon and that concentrations remain non-negative. The data of each model and the checks executed are given in Supplementary Note *Verified benchmark models* and

Supplementary Table S8, which also explain why the remaining seven benchmark problems were not verified.

Setting and goal

A model is an ODE $\dot{x} = f(x, \theta)$ on $[0, T]$ with $x(0) = x^0(\theta)$, $\theta \in \mathbb{R}^p$, sampled at times $t_1, \ldots, t_N$; the sampled (noise-scaled) outputs form a vector $F(\theta) \in \mathbb{R}^m$. Fixing the parameters outside a subset S at their nominal values means replacing a perturbation $\Delta\theta$ by $\Delta\theta_S$, whose components outside S are zero. The proofs establish six guarantees:

(G1) for small perturbations, the output error of the reduced model, $\|F(\theta_0 + \Delta\theta) - F(\theta_0 + \Delta\theta_S)\|$, is bounded by the sensitivity energy discarded, i.e. by $R_{\mathrm{var}}(S)$;

(G2) the cosine $\cos\Delta$ between full and reduced output changes determines the best relative error obtainable by rescaling the reduced response, $\sqrt{1 - \cos^2\Delta}$, and the linearised re-estimation of several parameters is at least as good;

(G3) the conditioning tests (κ, VIF) imply that the *linearised* least-squares problem for θ_S has a unique solution that is stable with respect to the data;

(G4) the fitting error is monotone in the re-estimated set, and backward elimination returns a subset from which no single parameter can be removed;

(G5) the hypotheses needed for (G1)–(G3) (smoothness, existence of solutions, positivity, differentiability) hold for the translated benchmark models;

(G6) a parameter of the reduced problem is practically identifiable (its profile likelihood exceeds the 95% threshold on both sides within the bounds) when a guaranteed half-width t^*, computed from the profiled information, the residual gradient and a bound on the nonlinear remainder, is smaller than the distance to the bounds.

Steps 1–22 prove (G1)–(G4) for any model satisfying explicit hypotheses; steps 23–31 prove that the 28 verified benchmark models satisfy them (G5); step 32 proves (G6) for any reduced problem. Supplementary Table S5 states, for each guarantee, what is proved, what is not, and which numerical quantity connects it with the results of the paper.

Supplementary Table S5. Scope of the formal guarantees. Numerical values refer to the final subsets of the 31 admissible benchmark problems and were computed in Python in double precision.

Guarantee	What Lean proves	What it does not prove	Numerical connection with the paper
G1. Error of the reduced model	For small $\Delta\theta$: $\|F(\theta_0+\Delta\theta) - F(\theta_0+\Delta\theta_S)\| \leq (\sqrt{(1 - R_{\mathrm{var}})E} + \eta)\|\Delta\theta\|$; an explicit version for finite $\|h\| \leq \rho_0$ with a Lipschitz constant L_c (steps 1–4, 19–20)	A small error for arbitrarily large perturbations; that the selected subsets have R_{var} close to 1	$R_{\mathrm{var}}(S)$ of the final subsets has median 0.60 and is ≥ 0.89 in 2 of 31 (level 3 of Supplementary Table S1); the bound is therefore informative only for some subsets, and admissibility is decided by e_{fit} (G2, G4); the explicit finite version could not certify any benchmark scenario because L_c is very large (see below)

Supplementary Table S5. Scope of the formal guarantees. Numerical values refer to the final subsets of the 31 admissible benchmark problems and were computed in Python in double precision.

Guarantee	What Lean proves	What it does not prove	Numerical connection with the paper
G2. Cosine and fitting error	$\min_\alpha \|a - \alpha b\|/\|a\| = \sqrt{1 - \cos^2(a, b)}$; the linearised fit of several parameters is no worse; certificates of $\cos\Delta$ for finite and small perturbations (steps 3–4, 12, 20–22)	That the nonlinear multi-parameter optimisation performed in Python reaches its optimum; the value of e_{fit} itself	Median e_{fit} over 15 scenarios at $\pm 5\%$ is $\leq \tau_e = 0.436$ in all 31 admissible problems (largest value 0.43585, Lucarelli, after nonlinear re-estimation); the median $\cos\Delta$ is ≥ 0.9 in 17 of them (levels 2–3 of Supplementary Table S1), so a subset can be admissible because re-estimating several parameters fits better than rescaling
G3. Conditioning	Uniqueness and stability of the linear least-squares estimate with normalised sensitivity columns, with $\kappa = \sqrt{\lambda_{\max}/\lambda_{\min}}$ and $\text{VIF}_j = [(Z^\top Z)^{-1}]_{jj}$ as computed (steps 7–9)	Global or structural identifiability of the nonlinear model; uniqueness of the nonlinear calibration	Centred columns of the biological parameters (the selection): $\kappa \leq 6.6$ and $\text{VIF} \leq 9.85$ for all 31 final subsets (thresholds $\tau_\kappa = \tau_{\text{VIF}} = 10$). Unit-length columns Z (the least-squares statement): in 30 of 31 subsets, not Rahman ($\kappa = 12.4$, VIF$= 39$); with the calibration parameters included, $\kappa > 10$ in ten problems (Supplementary Table S44)
G4. Re-estimation and pruning	e_{fit} never increases when parameters are added; $e_{\text{fit}} \leq e_{\text{rel}}$; pruning returns an admissible subset from which no parameter can be removed, and which is minimal with respect to inclusion when the criterion is monotone (steps 5–6)	That the result is the admissible subset of smallest size; that the numerical evaluation of admissibility is exact	Admissibility of each candidate subset is evaluated numerically (15 scenarios, median $e_{\text{fit}} \leq 0.436$)
G5. Benchmark models	For 28 translated models, at the nominal parameters and in the base condition (first segment of piecewise models): smooth vector field, existence of the solution on the horizon (every T, or $T \leq 1$ for Crauste), non-negativity, solutions for θ near θ_0 and their differentiability (steps 23–31)	That the SBML and PEtab files are correct; the accuracy of the numerical simulations; experimental conditions that override parameters or initial values (the translation uses the nominal values)	The translated vector fields, parameter values and initial values agree with the SBML models evaluated by libRoadRunner (Supplementary Note *Verified benchmark models*)

Supplementary Table S5. Scope of the formal guarantees. Numerical values refer to the final subsets of the 31 admissible benchmark problems and were computed in Python in double precision.

Guarantee	What Lean proves	What it does not prove	Numerical connection with the paper
G6. Identifiability of the selected parameters	If $\lvert\langle r_0, A\delta\rangle\rvert \le \gamma\lVert A\delta\rVert$, $I_j\delta_j^2 \le \lVert A\delta\rVert^2$ and the nonlinear remainder is at most ε on the admissible set, every point outside $\lvert\delta_j\rvert < t^*$ has $\Delta\chi^2 \ge 3.84$, so θ_j is practically identifiable when t^* is below the distance to both bounds (step 32)	Identifiability of the full model or global identifiability; a bound on the remainder over the whole admissible set, which is evaluated only at the computed profile points	The linear condition ($\gamma = \varepsilon = 0$) agrees with the profiles for all 72 evaluable selected parameters; the full hypotheses hold for 24 parameters, never for a non-identifiable one, and for all selected parameters in 7 problems (Supplementary Note *Identifiability of the selected parameters*)

Steps 1–13. Finite-dimensional core of the method

Here the sensitivity matrix $J = DF(\theta_0)$ is a fixed $m \times p$ matrix with column energies $E_j = \sum_i J_{ij}^2$, $E = \sum_j E_j$, $E_S = \sum_{j\in S} E_j$.

Steps 1–2 (error of the reduced model, G1). The total energy splits as $E = E_S + E_{S^c}$, and by Cauchy–Schwarz

$$\lVert J\Delta\theta - J\Delta\theta_S\rVert = \Big\lVert\sum_{j\notin S} J_{\cdot j}\Delta\theta_j\Big\rVert \le \sqrt{E_{S^c}}\,\lVert\Delta\theta_{S^c}\rVert \le \sqrt{(1 - R_{\mathrm{var}}(S))\,E}\,\lVert\Delta\theta\rVert$$

(`first_order_error_from_Rvar`). If F is differentiable at θ_0, the nonlinear error differs from the linear one by $o(\lVert\Delta\theta\rVert)$, so for every $\eta > 0$ and small $\Delta\theta$, $\lVert F(\theta_0+\Delta\theta) - F(\theta_0+\Delta\theta_S)\rVert \le (\sqrt{(1-r)E}+\eta)\lVert\Delta\theta\rVert$ whenever $R_{\mathrm{var}}(S) \ge r$ (`trajectory_error_from_Rvar`). This is the guarantee behind stage 1 of the selection. It is local, and it does not by itself make a subset admissible: admissibility is a separate, numerical test on finite perturbations (steps 5–6).

Step 3 (cosine and fitting error, G2). For vectors a (full response) and b (reduced response),

$$\min_{\alpha\in\mathbb{R}} \frac{\lVert a - \alpha b\rVert}{\lVert a\rVert} = \sqrt{1 - \cos^2(a,b)},$$

attained at $\alpha^* = \langle a, b\rangle/\lVert b\rVert^2$ (`min_rel_error_eq`, Theorem 1 of the paper). Hence, for $c > 0$, a relative error of at most $\tau_e = \sqrt{1-c^2}$ after rescaling is equivalent to $\lvert\cos\Delta\rvert \ge c$; in particular $\cos\Delta \ge c$ implies it (the sign matters only because α may be negative). The identity concerns a single scale factor. The method re-estimates several parameters, and step 12 shows that the linearised fit over the span of their sensitivity columns is at least as good as rescaling (`lin_le_scaled`); the exact nonlinear fit differs from the linearised one by a second-order term. Thus $\cos\Delta \ge c$ is a sufficient condition for admissibility to first order, not a necessary one, and the actual e_{fit} is computed numerically.

Step 4 (finite perturbations). If DF is L_c-Lipschitz on the ball $\bar{B}(\theta_0, \rho_0)$, Taylor's theorem gives $\lVert F(\theta_0 + h) - F(\theta_0) - Jh\rVert \le L_c\lVert h\rVert^2$, hence

$$\lVert F(\theta_0 + h) - F(\theta_0 + h_S)\rVert \le \sqrt{E_{S^c}}\,\lVert h_{S^c}\rVert + 2L_c\lVert h\rVert^2,$$

and, writing $A = \lVert Jh\rVert - L_c\lVert h\rVert^2 > 0$, the inequality $(\sqrt{E_{S^c}}\lVert h_{S^c}\rVert + 2L_c\lVert h\rVert^2)^2 \le (1-c^2)A^2$ certifies $\cos\Delta(h) \ge c$ for that perturbation (`finite_cos_certificate`).

Steps 5–6 (re-estimation and pruning, G4). With $e_{\mathrm{fit}}(S) = \inf\{r(\theta) : \theta_j = \theta_{0,j}\ \forall j \notin S\}$, enlarging S (or adding calibration parameters) never increases e_{fit}, and $e_{\mathrm{fit}} \le e_{\mathrm{rel}}$ (`fitErr_anti`, `fitErr_le_erel`);

therefore admissibility is preserved under enlargement. Backward elimination stops at an admissible $R \subseteq S$ such that $R \setminus \{x\}$ is not admissible for any $x \in R$ (`prune_spec`, Theorem 2 of the paper).

Steps 7–9 (conditioning and selection, G3). Let Z have the normalised sensitivity columns of S. If $\sigma_{\min}\|u\| \le \|Zu\| \le \sigma_{\max}\|u\|$ with $\sigma_{\min} > 0$, the linear least-squares estimate (the first-order approximation of the re-estimation of θ_S) is unique and, for two data vectors y, y',

$$\|v - v'\| \le \kappa\,\|y - y'\|, \qquad |v_j - v'_j| \le \sqrt{\mathrm{VIF}_j}\,\|y - y'\|$$

(`ls_unique`, `ls_kappa_stability`, `ls_vif_inverse_stability`). Step 8 identifies these quantities with what is computed numerically: $\kappa = \sqrt{\lambda_{\max}/\lambda_{\min}}$ from the eigenvalues of $Z^\top Z$ and $\mathrm{VIF}_j = [(Z^\top Z)^{-1}]_{jj}$ (`ls_kappa_stability_spectral`, `ls_vif_matrix_stability`); note that κ is the condition number of Z, i.e. the square root of that of $Z^\top Z$. These results concern the linearised problem; they do not establish identifiability of the nonlinear model. Step 9 proves that the greedy selection returns a subset with $\kappa \le \tau_\kappa$ and $\mathrm{VIF} \le \tau_{\mathrm{VIF}}$, and, because both tests are inherited by subsets, that a discarded candidate would still fail at the end (`greedy_kappa_vif_guarantee`).

Steps 10–13 (complements). Two parameters with proportional sensitivity columns ($Ze_a = \lambda Ze_b$) give $Z(e_a - \lambda e_b) = 0$, so no κ threshold accepts both and the estimate is not unique; zero discarded energy makes the reduction exact to first order (step 10). Values computed with finite differences and rounding certify R_{var}, κ and VIF for the exact matrix when explicit margins hold (step 11). The linearised e_{fit} used for large systems differs from the exact one by at most $M\rho^2$ (step 12). Error bounds transfer from states to measured outputs through Lipschitz observation functions, including log and $\log_{10}$ (step 13).

Steps 14–18. Ordinary differential equations

Steps 1–13 need F to be differentiable, with known derivative, for every θ near θ_0. Steps 14–18 prove this for ODE models.

Step 14. If $v(t,\cdot)$ is K-Lipschitz, Picard iteration on intervals of length at most $1/(2K)$ gives a solution on all of $[a,b]$ (`exists_solution_Icc`).

Step 15 (sensitivities). If f is C^1 on an open set U containing the nominal trajectory and x^0 is differentiable, then $x(t;\theta)$ is differentiable in θ at θ_0 for every $t \in [0,T]$, and its derivative $S(t)$ solves the variational equation

$$\dot S(t) = \partial_x f(x_0(t),\theta_0)\,S(t) + \partial_\theta f(x_0(t),\theta_0), \qquad S(0) = \partial_\theta x^0(\theta_0)$$

(`hasFDerivAt_solution_param_init`). The proof combines a uniform linearisation of f near the compact trajectory with a Grönwall estimate in a tube around it.

Step 16. Solutions for θ near θ_0 exist automatically once the nominal solution exists (`exists_solutions_near`), so this need not be assumed.

Step 17 (explicit constants). If $\|Df\| \le M_1$ and Df is M_2-Lipschitz on a convex set containing the trajectories,

$$\|x(t;\theta) - x(t;\theta')\| \le (e^{M_1 T} - 1)\|\theta - \theta'\|, \qquad \|S_\theta(t) - S_{\theta'}(t)\| \le M_2 e^{2M_1 T}\frac{e^{M_1 T} - 1}{M_1}\|\theta - \theta'\|$$

(`traj_lipschitz`, `sens_lipschitz`).

Step 18. The same holds segment by segment for systems with events at fixed times (`event_hasFDerivAt`).

Steps 19–22. Certificates for ODE models

Steps 19–22 apply steps 1–13 to $F = \mathcal{F}$, the vector of states at the sampling times, whose derivative is the sampled sensitivity matrix $\mathcal{S} = (S(t_1), \dots, S(t_N))$ given by steps 14–18.

Step 19. The energy criterion computed from $\mathcal{S}$ bounds the trajectory error of the reduced model: for every $\eta > 0$ and small $\Delta\theta$,

$$R_{\mathrm{var}}(S) \geq r \;\Rightarrow\; \|\mathcal{F}(\theta_0 + \Delta\theta) - \mathcal{F}(\theta_0 + \Delta\theta_S)\| \leq \left(\sqrt{(1-r)E} + \eta\right)\|\Delta\theta\|$$

(`certified_parameter_reduction`).

Step 20. Step 17, applied at the N sampling times, provides the Lipschitz constant needed in step 4, $L_c = \sqrt{N}\, M_2 e^{2M_1 T}(e^{M_1 T} - 1)/M_1$, so the finite-perturbation certificate of $\cos\Delta$ becomes fully explicit for ODE models (`certified_cos_ode`).

Step 21. The certificate is sharpened by using the exact linear error $\|\mathcal{S}(h - h_S)\|$ instead of its energy bound, and a local version without any Lipschitz constant is proved: $\lim_{s\to 0^+} \cos\Delta(sh) = \cos(\mathcal{S}h, \mathcal{S}h_S)$, so if the linear cosine computed from $\mathcal{S}$ exceeds c, then $\cos\Delta(sh) > c$ for all small $s > 0$ (`certified_cos_ode_exact`, `local_cos_ode`).

Step 22. Combining steps 16 and 21, the local certificate requires only that the nominal solution exist (`local_cos_from_nominal`).

Steps 23–31. Hypotheses of the benchmark models

The results of steps 14–22 assume that f is C^1 on an open set containing the trajectory and that the nominal solution exists on $[0, T]$. Steps 23–31 prove these assumptions for the benchmark models from their equations.

Step 23 (smoothness). Rate laws are represented as expressions built from constants, concentrations, parameters, $+, -, \times, \div$, powers, exp and log. Wherever an expression is well defined (non-zero denominators, positive bases of real powers and arguments of log) it is C^∞, and this set is open (`contDiffOn_field`); Michaelis–Menten and Hill laws are covered (`ok_mm`, `ok_hill`, `ok_hillR`). This yields the local certificate for every kinetic model (`kinetic_local_cos`).

Step 24 (positivity). If f is quasi-positive ($y \geq 0$, $y_i = 0 \Rightarrow f_i(y) \geq 0$), solutions starting in the non-negative orthant stay there (`nonneg_of_quasiPositive`); the proof applies Grönwall's inequality to the squared negative part.

Step 25 (global existence). If, in addition, $\sum_i w_i f_i(y) \leq a + b\sum_i w_i y_i$ on the orthant for weights $w_i > 0$, then $\sum_i w_i x_i(t)$ grows at most exponentially, the solution remains in a box $[0, B]^n$ on which f is Lipschitz, and by step 14 it exists on $[0, T]$ for every T (`exists_global_solution`).

Step 26. The same holds when some species must remain strictly positive (for example, when they appear in denominators), provided their consumption is at most proportional to their concentration (`exists_global_solution_strict`).

Steps 27–28 (automatic checks). Boolean functions that inspect the syntax of the rate laws decide well-definedness, sign and quasi-positivity, and each is proved sound (for example `qp_sound`). For a reaction network $f_i = \sum_k \nu_{ik} r_k$, the checks `checkNet` and `checkGrowth` imply the hypotheses of steps 23–25 (`checkNet_sound`, `checkGrowth_sound`), and the end-to-end theorem `network_final` gives existence, positivity and differentiability for every horizon.

Steps 29–31. Three extensions cover the remaining models: quadratic growth, $\sum_i w_i f_i \leq Q(\sum_i w_i y_i + 1)^2$, gives existence up to the Riccati time $1/(Q(\sum_i w_i x_i^0 + 1))$ (`riccati_final`, used for Crauste with $T \leq 1$); networks with strictly positive domains (`strict_final_init`); and parameter-dependent initial values checked with rational interval arithmetic (`strict_final_initI`).

Benchmark models. Each file in `lean_proofs/Models/` contains the rate laws, stoichiometry, parameter values and initial values of one model, translated from SBML with exact rational numbers. The checks of steps 27–31 are executed by the Lean kernel (`decide +kernel`), and one end-to-end theorem is then applied: `network_final` (15 models), `network_final_init` (5), `network_segments_final` (1), `strict_final_init` (5), `strict_final_initI` (1) or `riccati_final` (1). The resulting theorem states that the model satisfies the hypotheses of steps 14–22. Supplementary Note *Verified benchmark models* gives the data and the checks of each model.

Connection with the numerical results and what the proofs do not cover

Lean proves implications; it does not recompute the numerical results. The sensitivity matrices, R_{var}, κ, VIF, $\cos\Delta$ and e_{fit} reported in the paper are computed in Python, in double precision. Three levels of evidence should therefore be distinguished.

- *General mathematics (G1–G4).* Proved in Lean for every model that satisfies the stated hypotheses.
- *Hypotheses of specific models (G5).* Proved in Lean for 28 benchmark models from their translated equations; the translation was compared independently with the SBML models (Supplementary Note *Verified benchmark models*).
- *Numerical conditions.* The inequalities required by the theorems ($\kappa \leq 10$, VIF ≤ 10, median $e_{\mathrm{fit}} \leq 0.436$, and $R_{\mathrm{var}} \geq r$ for the bound of G1) were evaluated in floating-point arithmetic; Supplementary Table S5 gives their values. The certified margins of step 11 were not applied to these computations. When the explicit finite-perturbation certificates of steps 4 and 20 are evaluated with bounds M_1, M_2 computed from the vector fields, the resulting constants L_c ranged from about 10^2 to 10^{24} and no scenario at $\pm 5\%$ could be certified in any model (file `results/preliminary/finite_certificates.md` of the reproducibility package). These certificates are therefore of theoretical value; in practice the method relies on the local results (steps 19, 21 and 22) and on the numerical admissibility test. The conclusions of G1–G3 for a particular problem are therefore conditional on the computed values; for the closest cases (VIF = 9.85 in Lucarelli and $e_{\mathrm{fit}} = 0.43585$ in Lucarelli) a small numerical error could change the decision.

The sensitivities used by the method are relative, $J = \partial y / \partial \log\theta$; the theorems apply to them through the change of variables $\theta = e^u$, which is smooth for $\theta > 0$. Finally, the formal results are local, around the published estimate θ_0; the behaviour for large parameter changes is assessed empirically in the paper.

Supplementary Note: Statements of the formal theorems

This note states, in mathematical notation, every theorem and lemma of the Lean formalisation (32 modules, 309 declarations) and explains what it means, so that the formal statements can be read alongside the code. The Lean name is given as it appears in the source (prefixed by the module namespace). Hypotheses are written before $\Rightarrow$; standing assumptions of each module are given in its description. Supplementary Table S6 explains the symbols used throughout; symbols specific to a module are explained at the start of that module. Supplementary Table S7 gives the statements, in one block per module, in the order of Supplementary Table S4; the blocks are grouped as in Supplementary Note *Logical sequence of the formal proofs*. The same content, with the Lean identifiers, is in `lean_proofs/FORMULAS.md`.

Supplementary Table S6. Notation of the formalisation.

Symbol	Lean	Meaning
$n,\ p,\ N$	`n, p, N`	number of state variables, of parameters and of sampling times
$\theta \in \mathbb{R}^p,\ \theta_0$	θ, θ_0	parameter vector; nominal (published) parameter values
$\dot{x} = f(x,\theta)$	`f`	right-hand side of the ODE (vector field)
$x(t;\theta) \in \mathbb{R}^n$	`x` θ `t`	solution of the ODE at time t for parameters θ
$x_0(t)$	$\mathtt{x}_0$	nominal solution, i.e. the solution for $\theta = \theta_0$
$x^0(\theta),\ D_0$	`x0, D0`	initial value as a function of θ, and its derivative at θ_0

Supplementary Table S6. Notation of the formalisation.

Symbol	Lean	Meaning
U	`U`	open set of (state, parameter) pairs on which f is continuously differentiable (ContDiffOn ℝ 1)
K	`K`	convex set containing all trajectories considered (used for explicit constants)
$[0, T]$	`Icc 0 T`	simulation horizon
x solves the ODE	`HasDerivWithinAt (x θ) (f (x θ t) θ) (Icc 0 T) t`	x(·;θ) is differentiable on [0,T] and its derivative equals f(x,θ)
$t_1, \ldots, t_N$	`tk`	sampling (measurement) times
$\mathcal{F}(\theta) \in \mathbb{R}^{Nn}$	`sampled x tk θ`	states at all sampling times stacked into one vector
$F : \mathbb{R}^p \to \mathbb{R}^m$	`F`	generic differentiable output map (for example the stacked, noise-scaled measured outputs)
$DF(\theta),\ J$	`DF θ, J, L`	derivative of F (sensitivity matrix); J = DF(θ_0)
J_{ij}	`jacobianOf L`	entry (i, j) of the matrix of the linear map L: J_ij = (L e_j)_i
$S(t) = \partial x(t;\theta_0)/\partial\theta$	`S, S₀`	state sensitivity matrix; solves the variational equation
$\mathcal{S}$	`sampledCLM S tk, sensMatrix S tk`	sensitivity at all sampling times stacked into one Nn × p matrix
$\dot{S} = \partial_x f\, S + \partial_\theta f$	`HasDerivWithinAt S (... ∘L S t + ...)`	variational (sensitivity) equation along the nominal solution
$S \subseteq \{1, \ldots, p\}$	`S, Sset : Finset (Fin p)`	subset of parameters that are re-estimated; the others stay at θ_0
S^c	`removedSet S`	discarded (fixed) parameters
$h,\ \Delta\theta$	`h, Δθ`	parameter perturbation
h_S	`keepS S h`	perturbation restricted to S (components outside S set to 0)
$E_j = \sum_i J_{ij}^2$	`columnEnergy J j`	sensitivity energy of parameter j (diagonal entry of $J^\top J$)
$E = \mathrm{tr}(J^\top J)$	`totalColumnEnergy J`	total sensitivity energy
$E_S,\ E_{S^c}$	`keptColumnEnergy J S, removedColumnEnergy J S`	energy of the selected and of the discarded parameters
$R_{\mathrm{var}}(S) = E_S/E$	`r * totalColumnEnergy ≤ keptColumnEnergy`	fraction of sensitivity energy kept; the condition R_var(S) ≥ r is written r·E ≤ E_S
$\|h_{S^c}\|^2 = \sum_{j\notin S} h_j^2$	`removedParamSq S h`	squared size of the perturbation of the discarded parameters
$\cos(a,b) = \dfrac{\langle a,b\rangle}{\|a\|\,\|b\|}$	`cosv a b`	cosine between two vectors
$\cos\Delta(h)$	`cosv (F (θ₀+h) - F θ₀) (F (θ₀+h_S) - F θ₀)`	cosine between the output change of the full perturbation and that of the perturbation restricted to S
c	`c`	cosine threshold (the method uses c = 0.9, i.e. $\tau_e = \sqrt{(1 - c^2)} = 0.436$)
$\bar{B}(\theta_0, \rho_0)$	`closedBall θ₀ ρ₀`	closed ball of radius ρ_0 around θ_0 (range of perturbations)
L_c	`Lc`	Lipschitz constant of DF on the ball: ‖DF(θ) – DF(θ_0)‖ ≤ L_c ‖θ – θ_0‖
$M_1,\ M_2$	M_1, M_2	bounds on ‖Df‖ and on the Lipschitz constant of Df on K
Z	`Z`	linear map whose columns are the (normalised) sensitivity columns of the candidate subset
$\sigma_{\min},\ \sigma_{\max}$	`σmin, σmax`	lower and upper bounds $\sigma_{\min}\|u\| \le \|Zu\| \le \sigma_{\max}\|u\|$ (extreme singular values)

Supplementary Table S6. Notation of the formalisation.

Symbol	Lean	Meaning
$\kappa_0,\ V$	`κ₀, V, Vb`	thresholds on the condition number κ and on the variance inflation factor VIF
$\operatorname{supp} u \subseteq S$	`SupportedOn S u`	u_i = 0 for all i outside S
$v \in \mathrm{LS}(Z, y)$	`IsLSMin Z y v`	v minimises ‖Zv − y‖ (least-squares estimate)
$\forall \theta \approx \theta_0$	`∀ᶠ θ in 𝓝 θ₀`	for all θ in some neighbourhood of θ_0
$o(\|\Delta\theta\|)$	`=o[𝓝 0]`	little-o as $\Delta\theta \to 0$
$e,\ e(x, \theta)$	`e : KExpr n p, e.eval (x, θ)`	syntax tree of a rate law built from constants, x_i, θ_j, +, −, ×, ÷, integer and real powers, exp and log; its value
$\mathrm{ok}_e(x, \theta)$	`e.ok (x, θ)`	e is well defined at (x, θ): non-zero denominators, positive bases of real powers and positive arguments of log
$F_1, \ldots, F_n;\ f_F$	`F : Fin n → KExpr n p, field F x θ`	rate expressions of a kinetic model and the vector field they define
$\mathcal{D}_F$	`domain F`	set of (x, θ) where all F_i are well defined
$P,\ \theta_P > 0$	`pos, PosParams pos θ`	parameters declared positive, and the condition θ_j > 0 for those parameters
$y \geq 0$	`Nonneg y`	all components of y are non-negative
$\Sigma,\ y \in P_\Sigma$	`sx, SPos sx y`	species required to stay strictly positive; y ≥ 0 and y_j > 0 for j ∈ Σ
(r_k, ν_k)	`Rx : List (KExpr × List (Fin n × ℚ))`	reaction k with rate law r_k and stoichiometric column ν_k
$f_i = \sum_k \nu_{ik} r_k$	`netF Rx, coef`	vector field of a reaction network; ν_ik = coef
w_i	`c : Fin n → ℚ`	positive weights of the growth (Lyapunov-type) bound Σ_i w_i f_i ≤ a + b Σ_i w_i x_i
$\theta_q,\ x_q$	`θq, xq`	rational parameter values and initial values taken from the model files
$\iota(q)$	`qvec q`	rational vector viewed as a real vector
`check`$(\cdot) =$ true	`checkModel, checkNet, ... = true`	a Boolean checker, executed by Lean on the model, returns true

Steps 1–13. Finite-dimensional core of the method (fixed sensitivity matrix)

Step 1. `FirstOrderErrorBound` (12)

First-order (linearised) output error when the parameters outside S are fixed. Here J is a numeric m × p matrix.

Notation: $(J\Delta\theta)_i$ (`linResponse J Δθ i`): linear response of output i: Σ_j J_ij Δθ_j; $(J\Delta\theta_S)_i$ (`reducedResponse J S Δθ i`): linear response with only the parameters in S perturbed; $\|v\|^2$ (`sqNorm v`): sum of squares of the components of v.

Supplementary Table S7. Statements of the formal theorems and lemmas, one block per module. Step 1: `FirstOrderErrorBound`.

Lean name	Statement	Meaning
`sqNorm_nonneg`	$\|v\|^2 \geq 0$	A sum of squares is non-negative.
`columnEnergy_nonneg`	$E_j \geq 0$	Each sensitivity energy is non-negative.
`removedColumnEnergy_nonneg`	$E_{S^c} \geq 0$	The discarded energy is non-negative.
`removedParamSq_nonneg`	$\|\Delta\theta_{S^c}\|^2 \geq 0$	The size of the discarded perturbation is non-negative.

Supplementary Table S7. Statements of the formal theorems and lemmas, one block per module. Step 1: `FirstOrderErrorBound`.

Lean name	Statement	Meaning
`sum_split`	$\sum_j f_j = \sum_{j\in S} f_j + \sum_{j\notin S} f_j$	A sum over all parameters splits into selected plus discarded ones.
`total_eq_kept_add_removed`	$\mathrm{tr}(J^\top J) = E_S + E_{S^c}$	Total energy = kept energy + discarded energy.
`response_sub_reduced`	$(J\Delta\theta)_i - (J\Delta\theta_S)_i = \sum_{j\notin S} J_{ij}\Delta\theta_j$	The truncation error is exactly the contribution of the discarded columns.
`first_order_error_sq_bound`	$\|J\Delta\theta - J\Delta\theta_S\|^2 \le E_{S^c}\,\|\Delta\theta_{S^c}\|^2$	First-order error bound (squared form, Cauchy–Schwarz).
`first_order_error_bound`	$\|J\Delta\theta - J\Delta\theta_S\| \le \sqrt{E_{S^c}}\,\|\Delta\theta_{S^c}\|$	First-order error bound (norm form).
`removedParamSq_le`	$\|\Delta\theta_{S^c}\|^2 \le \|\Delta\theta\|^2$	The discarded part of a perturbation is no larger than the whole perturbation.
`first_order_error_from_removed_fraction`	$E_{S^c} \le \varepsilon E \;\Rightarrow\; \|J\Delta\theta - J\Delta\theta_S\|^2 \le \varepsilon\, E\,\|\Delta\theta\|^2$	If at most a fraction ε of the energy is discarded, the first-order error is at most $\sqrt{(\varepsilon \mathrm{E})}\|\Delta\theta\|$.
`first_order_error_from_Rvar`	$rE \le E_S \;\Rightarrow\; \|J\Delta\theta - J\Delta\theta_S\|^2 \le (1-r)\,E\,\|\Delta\theta\|^2$	Stage-1 criterion: R_var(S) ≥ r bounds the first-order error by $\sqrt{((1 – r)\mathrm{E})}\|\Delta\theta\|$.

Step 2. `TrajectoryErrorBound` **(9)**

From the linear map to the true (nonlinear) output: error of the reduced model for small perturbations.

Supplementary Table S7. (continued). Step 2: `TrajectoryErrorBound`.

Lean name	Statement	Meaning
`keepS_apply`	$(h_S)_j = \begin{cases} h_j & j \in S \\ 0 & j \notin S \end{cases}$	Definition of the restricted perturbation.
`norm_keepS_le`	$\|h_S\| \le \|h\|$	Restricting a perturbation cannot increase its size.
`clm_apply_eq_linResponse`	$(L\Delta\theta)_i = \sum_j J_{ij}\Delta\theta_j$	A linear map acts through its matrix J.
`clm_keepS_eq_reducedResponse`	$(L\Delta\theta_S)_i = \sum_{j\in S} J_{ij}\Delta\theta_j$	Same for the restricted perturbation.
`linear_part_bound`	$\|L\Delta\theta - L\Delta\theta_S\| \le \sqrt{E_{S^c}}\,\|\Delta\theta_{S^c}\|$	Linear error bound for a linear map.
`sqrt_removedParamSq_le`	$\|\Delta\theta_{S^c}\| \le \|\Delta\theta\|$	Norm form of the previous inequality on perturbations.
`remainder_isLittleO`	$\big[F(\theta_0{+}\Delta\theta) - F(\theta_0{+}\Delta\theta_S)\big] - \big[L\Delta\theta - L\Delta\theta_S\big] = o(\|\Delta\theta\|)$	If F is differentiable at θ_0 with derivative L, the nonlinear error differs from the linear one by a term of higher order.
`trajectory_error_bound`	$\forall\eta>0,\ \forall\Delta\theta\approx 0:\ \|F(\theta_0{+}\Delta\theta) - F(\theta_0{+}\Delta\theta_S)\| \le \sqrt{E_{S^c}}\,\|\Delta\theta_{S^c}\| + \eta\|\Delta\theta\|$	Local error bound of the reduced model for the true output map.
`trajectory_error_from_Rvar`	$rE \le E_S \Rightarrow \forall\eta>0,\forall\Delta\theta\approx 0:$ $\|F(\theta_0{+}\Delta\theta) - F(\theta_0{+}\Delta\theta_S)\| \le \big(\sqrt{(1-r)E} + \eta\big)\|\Delta\theta\|$	Same bound expressed with R_var(S) ≥ r.

Step 3. `ScaledError` **(5)**

Theorem 1 of the paper: the minimum relative error after rescaling the reduced response equals $\sqrt{(1 - \cos^2\Delta)}$.

Supplementary Table S7. (continued). Step 3: `ScaledError`.

Lean name	Statement	Meaning
`norm_sub_smul_sq`	$\|a - \alpha b\|^2 = \|a\|^2 - 2\alpha\langle a, b\rangle + \alpha^2\|b\|^2$	Expansion of the squared error.
`scaled_error_ge`	$\|a\|^2 - \dfrac{\langle a, b\rangle^2}{\|b\|^2} \le \|a - \alpha b\|^2 \quad \forall\alpha$	Lower bound for every scaling α (b $\neq$ 0).
`scaled_error_eq`	$\left\|a - \dfrac{\langle a, b\rangle}{\|b\|^2} b\right\|^2 = \|a\|^2 - \dfrac{\langle a, b\rangle^2}{\|b\|^2}$	The bound is attained at $\alpha^* =$ ⟨a,b⟩/‖b‖².
`scaled_error_cos`	$\|a\|^2 - \dfrac{\langle a, b\rangle^2}{\|b\|^2} = \|a\|^2\big(1 - \cos^2(a, b)\big)$	Rewriting in terms of the cosine.
`min_rel_error_eq`	$\min\limits_{\alpha} \dfrac{\|a - \alpha b\|}{\|a\|} = \sqrt{1 - \cos^2(a, b)}$	Theorem 1: with a the full response and b the reduced one, the best relative error is $\sqrt{(1 - \cos^2\Delta)}$; hence cos Δ ≥ c ⇔ error ≤ τ_e.

Step 4. `CosineCertificate` **(5)**

Non-asymptotic bounds for finite perturbations and a per-scenario certificate of cos Δ.

Supplementary Table S7. (continued). Step 4: `CosineCertificate`.

Lean name	Statement	Meaning
`cos_lower_of_rel_error`	$c^2 + \varepsilon^2 \le 1,\ \|a - b\| \le \varepsilon\|a\| \Rightarrow c\,\|a\|\,\|b\| \le \langle a, b\rangle$	A small relative error implies a large cosine.
`cos_ge_of_rel_error`	$0 \le \varepsilon < 1,\ c \ge 0,\ c^2 + \varepsilon^2 \le 1,\ \|a - b\| \le \varepsilon\|a\|,\ a \neq 0 \Rightarrow b \neq 0,\ \cos(a, b) \ge c$	Same, with the cosine itself.
`taylor_bound`	$\|F(\theta_0 + h) - F(\theta_0) - DF(\theta_0)h\| \le L_c\|h\|^2 \quad (\|h\| \le \rho_0)$	Taylor bound when DF is L_c-Lipschitz on the ball.
`finite_trajectory_error`	$\|F(\theta_0 + h) - F(\theta_0 + h_S)\| \le \sqrt{E_{S^c}}\,\|h_{S^c}\| + 2L_c\|h\|^2$	Error of the reduced model for a finite perturbation (E computed from J = DF(θ_0)).
`finite_cos_certificate`	$A = \|Jh\| - L_c\|h\|^2 > 0,\ \left(\sqrt{E_{S^c}}\|h_{S^c}\| + 2L_c\|h\|^2\right)^2 \le (1 - c^2)A^2 \Rightarrow \cos\Delta(h) \ge c$	Certificate of cos Δ for one perturbation h, computable from J, L_c and ‖h‖.

Step 5. `RefitMonotone` **(5)**

Re-estimating more parameters never increases the fitting error.

Notation: $\mathcal{B}$ (`B`): admissible parameter region (bounds); $r(\theta) \ge 0$ (`r`): fitting error of θ (for example the relative output error); feas(S) (`feas` θ_0 `B S`): $\{\theta \in \mathcal{B} : \theta_\mathrm{j} = \theta_{0,j}$ for j $\notin$ S$\}$; $e_{\mathrm{fit}}(S) = \inf_{\theta \in \mathrm{feas}(S)} r(\theta)$ (`fitErr r` θ_0 `B S`): best error when only S is re-estimated.

Supplementary Table S7. (continued). Step 5: `RefitMonotone`.

Lean name	Statement	Meaning
`feas_mono`	$S \subseteq T \Rightarrow \mathrm{feas}(S) \subseteq \mathrm{feas}(T)$	More free parameters, more candidates.

Supplementary Table S7. (continued). Step 5: `RefitMonotone`.

Lean name	Statement	Meaning
`fitErr_anti`	$S \subseteq T \Rightarrow e_{\text{fit}}(T) \leq e_{\text{fit}}(S)$	The fitting error is anti-monotone in S.
`fitErr_union_le`	$e_{\text{fit}}(S \cup C) \leq e_{\text{fit}}(S)$	Adding the calibration parameters C never hurts.
`fitErr_le_erel`	$\theta_s \in \text{feas}(S) \Rightarrow e_{\text{fit}}(S) \leq r(\theta_s)$	e_fit $\leq$ e_rel: re-estimating is never worse than keeping the perturbed values of S.
`admissible_mono`	$S \subseteq T,\ e_{\text{fit}}(S) \leq \tau \Rightarrow e_{\text{fit}}(T) \leq \tau$	Admissibility is preserved when parameters are added.

Step 6. `Poda` (4; *poda* is Spanish for pruning)

Theorem 2 of the paper: backward elimination (pruning) returns an admissible subset from which no parameter can be removed; with a monotone criterion, no proper subset of it is admissible.

Notation: $Q(S)$ (`Q`): admissibility predicate; prune(S) (`prune Q pick hpick S`): repeatedly remove an element whose removal keeps Q, until none exists.

Supplementary Table S7. (continued). Step 6: `Poda`.

Lean name	Statement	Meaning
`prune_spec`	$Q(S) \Rightarrow Q(R) \ \wedge\ R \subseteq S \ \wedge\ \forall x \in R:$ $\neg Q(R \setminus \{x\}), \quad R = \text{prune}(S)$	Pruning keeps admissibility and stops at a set from which no single parameter can be removed.
`exists_minimal_admissible`	$Q(S) \Rightarrow \exists R \subseteq S:\ Q(R) \ \wedge\ \forall x \in R:$ $\neg Q(R \setminus \{x\})$	Every admissible subset contains an admissible subset from which no single parameter can be removed.
`no_proper_admissible_of_monotone`	$(U \subseteq T \wedge Q(U) \Rightarrow Q(T)) \ \wedge\ \forall x \in R:$ $\neg Q(R \setminus \{x\}) \Rightarrow \forall U \subsetneq R:\ \neg Q(U)$	With a monotone criterion, no single removable parameter implies that no proper subset is admissible.
`prune_inclusion_minimal`	Q monotone, $Q(S) \Rightarrow Q(R) \ \wedge\ R \subseteq S \ \wedge\ \forall U \subsetneq R:\ \neg Q(U)$	Pruning returns an inclusion-minimal admissible subset when the criterion is monotone (for e_fit: RefitMonotone.admissible_mono, with optimal re-estimation).

Step 7. `IdentifiabilityConditioning` (10)

Bounds on κ and VIF imply unique and stable least-squares estimates.

Supplementary Table S7. (continued). Step 7: `IdentifiabilityConditioning`.

Lean name	Statement	Meaning
`IsLSMin.inner_residual_eq_zero`	$v \in \text{LS}(Z, y) \Rightarrow \langle Zv - y, Zw \rangle = 0 \ \ \forall w$	Normal equations: the residual is orthogonal to the range of Z.
`ls_response_stability`	$\|Z(v - v')\| \leq \|y - y'\|$	Fitted responses are 1-Lipschitz in the data (v, v′ least-squares estimates for data y, y′).
`ls_param_stability`	$\sigma_{\min}\|u\| \leq \|Zu\| \Rightarrow \|v - v'\| \leq \|y - y'\|/\sigma_{\min}$	Parameter estimates are stable when $\sigma_{\min} > 0$.
`ls_unique`	$\sigma_{\min} > 0 \Rightarrow v = v'$ (same y)	The least-squares estimate is unique.

Supplementary Table S7. (continued). Step 7: `IdentifiabilityConditioning`.

Lean name	Statement	Meaning
`one_le_kappa_mul_sigmaMin`	$\|Ze_j\| = 1,\ \sigma_{\max} \le \kappa\sigma_{\min} \Rightarrow 1 \le \kappa\,\sigma_{\min}$	With unit columns, $\sigma_{\max} \ge 1$.
`ls_kappa_stability`	$\|Ze_j\| = 1 \Rightarrow \|v - v'\| \le \kappa\,\|y - y'\|$	With normalised columns, κ bounds the sensitivity of the estimates to the data.
`ls_param_kappa_stability`	$\|D(\theta - \theta')\| \le \kappa\,\|y - y'\|$	Same for the estimates in the original scale, $\theta = \mathrm{D}^{-1}\mathrm{v}$ with a column-scaling D.
`ls_vif_coordinate_stability`	$\big(u_j = 1 \Rightarrow \|Zu\|^2 \ge 1/V\big) \Rightarrow \|v_j - v'_j\| \le \sqrt{V}\,\|y - y'\|$	A bounded VIF makes each coordinate stable.
`vif_geometric_of_inverse`	$\langle Zw, Zu\rangle = u_j\ \forall u,\ w_j \le V \Rightarrow \big(u_j = 1 \Rightarrow \|Zu\|^2 \ge 1/V\big)$	The VIF computed from the inverse Gram matrix ($\mathrm{w} = (\mathrm{Z}^\top\mathrm{Z})^{-1}\mathrm{e_j}$, $\mathrm{VIF_j} = \mathrm{w_j}$) implies the geometric condition.
`ls_vif_inverse_stability`	$w_j \le V \Rightarrow \|v_j - v'_j\| \le \sqrt{V}\,\|y - y'\|$	Coordinate stability from the VIF as computed numerically.

Step 8. `SpectralConditioning` (22)

κ and VIF exactly as computed numerically: eigenvalues of the Gram matrix and the inverse matrix.

Notation: $G = Z^\top Z$ (`gram Z`): Gram operator; $\lambda_1 \ge \cdots \ge \lambda_k$, v_i (`lam Z, evec Z`): eigenvalues (decreasing) and orthonormal eigenvectors of G; $\lambda_{\max} = \lambda_1$, $\lambda_{\min} = \lambda_k$ (`lam Z (iMax hk), lam Z (iMin hk)`): largest and smallest eigenvalue (k > 0); $\kappa = \sqrt{\lambda_{\max}/\lambda_{\min}}$: condition number of Z; M (`M : Matrix (Fin m) (Fin k) ℝ`, `matCLM M`): numeric matrix and the linear map it defines.

Supplementary Table S7. (continued). Step 8: `SpectralConditioning`.

Lean name	Statement	Meaning
`inner_gram`	$\langle Gu, v\rangle = \langle Zu, Zv\rangle$	Defining property of the Gram operator.
`gram_isSymmetric`	$G = G^\top$	The Gram operator is symmetric.
`finrank_eq`	$\dim \mathbb{R}^k = k$	Dimension (needed for the spectral theorem).
`lam_antitone`	$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k$	Eigenvalues are sorted in decreasing order.
`gram_evec`	$Gv_i = \lambda_i v_i$	Eigenvector equation.
`sq_norm_apply_evec`	$\|Zv_i\|^2 = \lambda_i$	Eigenvalues are squared singular values.
`lam_nonneg`	$\lambda_i \ge 0$	The Gram matrix is positive semidefinite.
`sq_norm_apply_eq_sum`	$\|Zu\|^2 = \sum_i \lambda_i \langle v_i, u\rangle^2$	Spectral decomposition of $\|\mathrm{Zu}\|^2$.
`sq_norm_eq_sum`	$\|u\|^2 = \sum_i \langle v_i, u\rangle^2$	Parseval identity in the eigenbasis.
`sq_norm_bounds`	$lo \le \lambda_i \le hi\ \forall i \Rightarrow lo\,\|u\|^2 \le \|Zu\|^2 \le hi\,\|u\|^2$	Eigenvalue bounds give norm bounds.
`lam_le_max`	$\lambda_i \le \lambda_{\max}$	Largest eigenvalue.
`min_le_lam`	$\lambda_{\min} \le \lambda_i$	Smallest eigenvalue.
`norm_bounds`	$\sqrt{\lambda_{\min}}\,\|u\| \le \|Zu\| \le \sqrt{\lambda_{\max}}\,\|u\|$	Extreme singular values bound $\|\mathrm{Zu}\|$.
`opNorm_eq_sqrt_lamMax`	$\|Z\| = \sqrt{\lambda_{\max}}$	The operator norm is the largest singular value.

Supplementary Table S7. (continued). Step 8: `SpectralConditioning`.

Lean name	Statement	Meaning
`sqrt_lamMin_mul_le`	$\sqrt{\lambda_{\min}}\,\|u\| \le \|Zu\|$	Lower bound by the smallest singular value.
`le_sqrt_lamMin`	$\big(\forall u:\ \sigma\|u\| \le \|Zu\|\big) \Rightarrow \sigma \le \sqrt{\lambda_{\min}}$	$\sqrt{\lambda_{\min}}$ is the best possible lower constant.
`ls_kappa_stability_spectral`	$\|Ze_j\| = 1,\ \lambda_{\min} > 0 \Rightarrow \|v - v'\| \le \sqrt{\lambda_{\max}/\lambda_{\min}}\,\|y - y'\|$	Stability of least squares with κ exactly as computed (ratio of singular values).
`matCLM_apply`	$(\mathrm{matCLM}\,M)\,u = Mu$	A matrix acts on vectors by multiplication.
`inner_matCLM`	$\langle Mu, Mv\rangle = u^\top (M^\top M) v$	Inner products through the Gram matrix.
`gram_toEuclideanLin`	$G(\mathrm{matCLM}\,M) = M^\top M$	The abstract Gram operator is the matrix $\mathrm{M}^\top\mathrm{M}$.
`vif_matrix_inverse`	$\det(M^\top M) \ne 0,\ w = (M^\top M)^{-1} e_j \Rightarrow \langle Mw, Mu\rangle = u_j\ \forall u,\ w_j = [(M^\top M)^{-1}]_{jj}$	The diagonal of the inverse Gram matrix (the VIF for normalised columns) satisfies the hypothesis of vif_geometric_of_inverse.
`ls_vif_matrix_stability`	$[(M^\top M)^{-1}]_{jj} \le V \Rightarrow \|v_j - v'_j\| \le \sqrt{V}\,\|y - y'\|$	Coordinate stability with the VIF exactly as computed numerically.

Step 9. `GreedySelection` (7)

The greedy selection algorithm with permanent discard and its guarantees.

Notation: $G = \mathrm{greedy}(\mathrm{ok}, \mathrm{stop}, l, S)$ (`greedy ok stop l S`): scan the ordered list l; add j if ok(S ∪ {j}), otherwise discard j permanently; stop when stop(S); okKappa(Z, κ_0, S) (`okKappa Z` κ_0 `S`): $\exists\ \sigma_{\min} > 0$, $\sigma_{\max} \le \kappa_0 \sigma_{\min}$ with $\sigma_{\min}\|\mathrm{u}\| \le \|\mathrm{Zu}\| \le \sigma_{\max}\|\mathrm{u}\|$ for all u supported on S ($\kappa \le \kappa_0$); okVIF(Z, V, S) (`okVIF Z V S`): for j ∈ S and every u supported on S with u_j = 1: $\|\mathrm{Zu}\|^2 \ge 1/\mathrm{V}$ (VIF_j ≤ V).

Supplementary Table S7. (continued). Step 9: `GreedySelection`.

Lean name	Statement	Meaning
`greedy_ok`	$\mathrm{ok}(S) \Rightarrow \mathrm{ok}(G)$	The result always passes the test.
`greedy_subset`	$S \subseteq G \subseteq S \cup l$	The result only contains the start set and candidates from the list.
`greedy_discard_permanent`	ok anti-monotone, $j \in l,\ j \notin G \Rightarrow \neg\,\mathrm{ok}(G \cup \{j\})\ \vee\ \mathrm{stop}(G)$	Discarding is never a mistake: a rejected candidate would still fail with the final set.
`supportedOn_mono`	$S \subseteq T,\ \mathrm{supp}\,u \subseteq S \Rightarrow \mathrm{supp}\,u \subseteq T$	Support is monotone.
`okKappa_anti`	$S \subseteq T,\ \mathrm{okKappa}(T) \Rightarrow \mathrm{okKappa}(S)$	Removing parameters never worsens the condition number.
`okVIF_anti`	$S \subseteq T,\ \mathrm{okVIF}(T) \Rightarrow \mathrm{okVIF}(S)$	Removing parameters never worsens the VIF.
`greedy_kappa_vif_guarantee`	$R = \mathrm{greedy}(\mathrm{okKappa} \wedge \mathrm{okVIF}, \mathrm{stop}, l, \emptyset):\ \kappa(R) \le \kappa_0,\ \mathrm{VIF}(R) \le V,\ R \subseteq l,\ \forall j \in l \setminus R:\ R \cup \{j\}$ fails $\vee$ stop(R)	Guarantee of the selection step: the returned subset satisfies $\kappa \le \tau_\kappa$ and VIF $\le \tau_{\mathrm{VIF}}$, and every discarded candidate would violate one of them.

Step 10. `PatternMechanisms` (4)

Mechanisms behind the robustness patterns: opposing effects and zero discarded energy.

Supplementary Table S7. (continued). Step 10: `PatternMechanisms`.

Lean name	Statement	Meaning
`not_okKappa_of_kernel`	$u \neq 0,\ \operatorname{supp} u \subseteq S,\ Zu = 0 \Rightarrow \neg\,\mathrm{okKappa}(Z, \kappa_0, S)\ \forall \kappa_0$	If two or more parameters in S have exactly compensating effects, no threshold accepts S.
`ls_not_unique_of_kernel`	$Zu = 0,\ u \neq 0,\ v \in \mathrm{LS}(Z, y) \Rightarrow v + u \in \mathrm{LS}(Z, y),\ v + u \neq v$	In that case the estimate is not unique (non-identifiability).
`opposing_pair_kernel`	$Ze_a = \lambda Z e_b,\ a \neq b \Rightarrow e_a - \lambda e_b \neq 0,\ Z(e_a - \lambda e_b) = 0$	Two proportional sensitivity columns give a kernel vector.
`exact_reduction_of_zero_energy`	$E_{S^c} = 0 \Rightarrow J\Delta\theta = J\Delta\theta_S$	If the discarded parameters have zero sensitivity, the reduction is exact to first order.

Step 11. `NumericalRobustness` (6)

From approximate numerical values (finite differences, rounding) to the exact hypotheses, with explicit margins.

Notation: $\tilde{J}$, $\tilde{Z}$ (`Jt, Zt`): approximate (computed) matrices; τ, τ_j (τ): bounds on the approximation error; $J_{\cdot j}$ (`col J j`): column j of J.

Supplementary Table S7. (continued). Step 11: `NumericalRobustness`.

Lean name	Statement	Meaning
`fd_column_error`	$\|v\| = 1,\ 0 < \delta \leq \rho_0 \Rightarrow \left\| \frac{F(\theta_0 + \delta v) - F(\theta_0)}{\delta} - DF(\theta_0)v \right\| \leq L_c\,\delta$	Error of a finite-difference column with step δ.
`columnEnergy_eq_norm_sq`	$E_j = \|J_{\cdot j}\|^2$	Energy of a column is its squared norm.
`rvar_certified_from_approx`	$\|J_{\cdot j} - \tilde{J}_{\cdot j}\| \leq \tau_j,\ r\Big(\sum_{j\in S}(\|\tilde{J}_{\cdot j}\| - \tau_j)_+^2 + \sum_{j\notin S}(\|\tilde{J}_{\cdot j}\| + \tau_j)^2\Big) \leq \sum_{j\in S}(\|\tilde{J}_{\cdot j}\| - \tau_j)_+^2 \Rightarrow rE(J) \leq E_S(J)$	$\mathrm{R_var(S)} \geq \mathrm{r}$ for the exact matrix follows from a check on the approximate one with margins.
`opNorm_le_frobenius`	$\|Du\| \leq \sqrt{\sum_j \|De_j\|^2}\,\|u\|$	Operator norm bounded by the Frobenius norm.
`kappa_certified_from_approx`	$\|(Z - \tilde{Z})u\| \leq \tau\|u\|,\ \tau < \tilde{\sigma}_{\min},\ \tilde{\sigma}_{\max} + \tau \leq \kappa_0(\tilde{\sigma}_{\min} - \tau) \Rightarrow \mathrm{okKappa}(Z, \kappa_0, S)$	$\kappa \leq \kappa_0$ for the exact matrix from bounds on the approximate one.
`vif_certified_from_approx`	$\mathrm{okVIF}(\tilde{Z}, \tilde{V}, S),\ \tau < \tilde{\sigma}_{\min} \Rightarrow \mathrm{okVIF}\Big(Z, \frac{\tilde{V}}{(1 - \tau/\tilde{\sigma}_{\min})^2}, S\Big)$	VIF bound for the exact matrix from the approximate one.

Step 12. `LinearizedFit` (3)

The linearised e_fit (used for large systems) and its relation to the exact e_fit.

Notation: $A(c)$, $F(c)$ (`A, F`): linear and exact responses as functions of the re-estimated parameters c; d (`d`): target response (perturbed full model).

Supplementary Table S7. (continued). Step 12: `LinearizedFit`.

Lean name	Statement	Meaning
`lin_le_scaled`	$b \in V,\ v = \arg\min_{w \in V} \|d - w\| \Rightarrow \|d - v\| \leq \|d\|\sqrt{1 - \cos^2(d, b)}$	The linear best fit in a subspace containing b is at least as good as rescaling b.
`fit_upper_of_lin`	$\|F(c) - A(c)\| \leq M\|c\|^2 \Rightarrow \|F(c) - d\| \leq \|A(c) - d\| + M\|c\|^2$	Upper bound: the exact error at the linear optimum exceeds the linear error by at most a second-order term.
`fit_lower_of_lin`	$\|F(c) - A(c)\| \leq M\|c\|^2\ \forall c,\ e_{\text{lin}} \leq \|A(c) - d\|\ \forall c,\ \|c\| \leq \rho \Rightarrow e_{\text{lin}} - M\rho^2 \leq \|F(c) - d\|$	Lower bound: no exact fit within radius ρ is better than e_lin – Mρ^2.

Step 13. `OutputComposition` (4)

Error bounds transfer from states to measured outputs y_i = h(g_i(x(t_i), θ))/σ_i, including log and log10 transforms.

Notation: G_i (`G i`): observation function of output i.

Supplementary Table S7. (continued). Step 13: `OutputComposition`.

Lean name	Statement	Meaning
`output_sq_bound`	$\|G_i(x_i) - G_i(x_i')\| \leq K\|x_i - x_i'\| \Rightarrow \sum_i \big(G_i(x_i) - G_i(x_i')\big)^2 \leq K^2 \sum_i \|x_i - x_i'\|^2$	Lipschitz observables transfer squared state errors to outputs.
`output_uniform_bound`	$\|x_i - x_i'\| \leq \varepsilon \Rightarrow \sum_{i \in s} \big(G_i(x_i) - G_i(x_i')\big)^2 \leq \|s\|\, K^2\varepsilon^2$	Uniform version.
`log_lipschitz`	$a, b \geq c > 0 \Rightarrow \|\log a - \log b\| \leq \|a - b\|/c$	Natural logarithm is Lipschitz away from 0.
`log10_lipschitz`	$a, b \geq c > 0 \Rightarrow \|\log_{10} a - \log_{10} b\| \leq \dfrac{\|a - b\|}{c \ln 10}$	Same for log10 (PEtab observable transformation).

Steps 14–18. Ordinary differential equations: existence and sensitivities

Step 14. `GlobalLipschitzODE` (3)

Existence of solutions on a whole interval for vector fields that are Lipschitz in the state.

Notation: $\text{Sol}_{[a,c]}(\alpha)$ (`IsSolOn v a c` α): α solves α'(t) = v(t, α(t)) on [a, c].

Supplementary Table S7. (continued). Step 14: `GlobalLipschitzODE`.

Lean name	Statement	Meaning
`isSolOn_self`	$\text{Sol}_{[a,a]}(\alpha)$	Every function solves the equation on a single point (base case).
`extend_solution`	$K(d - c) \leq \frac{1}{2},\ \text{Sol}_{[a,c]}(\alpha) \Rightarrow \exists \gamma :\ \gamma(a) = \alpha(a),\ \text{Sol}_{[a,d]}(\gamma)$	If v(t,·) is K-Lipschitz and continuous in t, a solution can be extended by a step of length at most 1/(2K) (Picard iteration).
`exists_solution_Icc`	$\forall y_0\ \exists \alpha :\ \alpha(a) = y_0,\ \text{Sol}_{[a,b]}(\alpha)$	Global existence on [a, b] for a vector field that is K-Lipschitz in the state.

Step 15. `ParamDiffODE` (8)

Differentiable dependence of ODE solutions on parameters and initial values (the derivative solves the variational equation).

Notation: ι_1, ι_2 (`ContinuousLinearMap.inl / inr`): embeddings u $\mapsto$ (u, 0) and k $\mapsto$ (0, k) of the state and parameter spaces into their product.

Supplementary Table S7. (continued). Step 15: `ParamDiffODE`.

Lean name	Statement	Meaning
`uniform_ linearization`	$\exists \delta \in (0, R]\ \forall z \in K_0,\ \|w - z\| \le \delta :$ $\|g(w) - g(z) - Dg(z)(w - z)\| \le \eta \|w - z\|$	A C^1 function is uniformly linearisable near a compact set K_0 whose R-neighbourhood lies in U, for every $\eta > 0$.
`tube_gronwall`	$\|x_\theta(t) - y_h(t)\| \le$ $\delta_0 e^{KT} + \varepsilon_g \dfrac{e^{KT} - 1}{K} \quad \forall t \in [0, T]$	Grönwall inequality in a tube: an approximate solution y_h with defect $\le \varepsilon_g$ that stays within σ of the nominal trajectory, compared with an exact solution x_θ starting δ_0 away; the exact solution does not leave the tube where v is K-Lipschitz.
`norm_comp_inl_le`	$\|L \circ \iota_1\| \le \|L\|$	Restricting a linear map on a product to the first factor does not increase its norm.
`clm_prod_apply_split`	$L(u, k) = L(\iota_1 u) + L(\iota_2 k)$	A linear map on a product splits into its state and parameter parts.
`lipschitzOn_slice`	$\|Df\| \le M$ on $\bar{B}(c, \rho) \times \{\theta\} \Rightarrow y \mapsto$ $f(y, \theta)$ is M-Lipschitz on $\bar{B}(c, \rho)$	Lipschitz constant of a slice from a bound on the derivative.
`hasFDerivAt_ solution_param_init`	$\exists S :\ S(0) = D_0,\ \dot{S}(t) =$ $\partial_x f(x_0(t), \theta_0) S(t) +$ $\partial_\theta f(x_0(t), \theta_0),\ \dfrac{\partial x(t;\theta)}{\partial \theta}\Big\|_{\theta_0} = S(t)$	Main theorem: if f is C^1 on an open U containing the nominal trajectory, solutions exist for θ near θ_0 and x(0;θ) = $x^0(\theta)$ is differentiable, then x(t;θ) is differentiable in θ for every t $\in$ [0,T] and its derivative solves the variational equation.
`hasFDerivAt_ solution_param_on`	$x(0;\theta) = x(0;\theta_0) \Rightarrow \exists S :\ S(0) = 0,\ \dot{S} =$ $\partial_x f\, S + \partial_\theta f,\ \partial_\theta x(t;\theta_0) = S(t)$	Same with an initial value that does not depend on θ.
`hasFDerivAt_ solution_param`	$f \in C^1(\mathbb{R}^n \times \mathbb{R}^p) \Rightarrow \exists S :\ S(0) = 0,\ \dot{S} =$ $\partial_x f\, S + \partial_\theta f,\ \partial_\theta x(t;\theta_0) = S(t)$	Same for a vector field that is C^1 everywhere.

Step 16. `LocalExistence` (9)

Solutions exist for all parameters near θ_0 as soon as the nominal solution exists (removes the hypothesis that perturbed solutions exist).

Notation: $\mathrm{clamp}_c(y)_i = \max(-c, \min(c, y_i))$ (`clampE c y`): truncation of each component to [–c, c].

Supplementary Table S7. (continued). Step 16: `LocalExistence`.

Lean name	Statement	Meaning
`clampE_apply`	$\mathrm{clamp}_c(y)_i = \max(-c, \min(c, y_i))$	Definition.
`abs_clamp_sub_le`	$\lvert \mathrm{clamp}_c(a) - \mathrm{clamp}_c(b) \rvert \le \lvert a - b \rvert$	Scalar truncation is 1-Lipschitz.
`dist_clampE_le`	$\|\mathrm{clamp}_c(y) - \mathrm{clamp}_c(y')\| \le \|y - y'\|$	Vector truncation is 1-Lipschitz.
`lipschitz_clampE`	clamp_c is 1-Lipschitz	Same, as a Lipschitz statement.

Supplementary Table S7. (continued). Step 16: `LocalExistence`.

Lean name	Statement	Meaning
`clampE_eq`	$\|y\| \le c \Rightarrow \mathrm{clamp}_c(y) = y$	Truncation does not change small vectors.
`clampE_zero`	$c \ge 0 \Rightarrow \mathrm{clamp}_c(0) = 0$	Truncation fixes the origin.
`norm_clampE_le`	$\|\mathrm{clamp}_c(y)\| \le (n+1)\,c$	Truncated vectors are bounded.
`exists_solutions_near`	$\exists x(\cdot;\theta):\ x(\cdot;\theta_0) = x_0,\ \forall \theta \approx \theta_0:\ x(0;\theta) = x^0(\theta),\ \dot{x} = f(x,\theta),\ (x(t;\theta),\theta) \in U$	If f is C^1 on U, the nominal solution exists on [0,T] inside U and x^0 is continuous at θ_0, then solutions exist on the same interval for all θ close to θ_0.
`hasFDerivAt_of_nominal_solution`	$\exists x(\cdot;\theta),\ \exists S:\ S(0) = D_0,\ \dot{S} = \partial_x f\, S + \partial_\theta f,\ \partial_\theta x(t;\theta_0) = S(t)$	Existence near θ_0 and differentiability, assuming only the nominal solution.

Step 17. `SensitivityLipschitz` (8)

Explicit constants: Lipschitz dependence of trajectories and sensitivities on θ (Grönwall).

Notation: $g(z) = f(x,\theta),\ Dg$ (`g, Dg`): vector field as a function of z = (x, θ) and its derivative.

Supplementary Table S7. (continued). Step 17: `SensitivityLipschitz`.

Lean name	Statement	Meaning
`gronwall_zero_init`	$u(0) = 0,\ \|u'\| \le K\|u\| + \varepsilon \Rightarrow \|u(t)\| \le \varepsilon \dfrac{e^{KT} - 1}{K}$	Grönwall inequality with zero initial value.
`norm_comp_inl_le`	$\|\partial_x g\| \le \|Dg\|$	The state part of a derivative is bounded by the whole derivative.
`norm_comp_inr_le`	$\|\partial_\theta g\| \le \|Dg\|$	Same for the parameter part.
`g_lipschitz`	$\|Dg\| \le M_1$ on convex $K \Rightarrow \|g(z) - g(z')\| \le M_1\|z - z'\|$	Mean-value inequality.
`norm_pair_sub_le`	$\|(a,b) - (a',b')\| \le \|a - a'\| + \|b - b'\|$	Norm on the product space.
`traj_lipschitz`	$\|x(t;\theta) - x(t;\theta')\| \le \|\theta - \theta'\|\,(e^{M_1 T} - 1)$	Trajectories depend Lipschitz-continuously on θ.
`sens_bound`	$\|S(t)\| \le e^{M_1 T} - 1$	Explicit bound on the sensitivity.
`sens_lipschitz`	$\|S_\theta(t) - S_{\theta'}(t)\| \le \|\theta - \theta'\|\, M_2\, e^{2M_1 T} \dfrac{e^{M_1 T} - 1}{M_1}$	Sensitivities depend Lipschitz-continuously on θ; this gives the constant L_c used in the certificates.

Step 18. `EventSystems` (1)

Model class 2: systems with events at fixed times (piecewise ODEs with reset maps).

Notation: $f_k,\ L_k,\ R_k$ (`f k, L k, R k`): vector field, length and reset map of segment k: y_{k+1}(0) = R_k(y_k(L_k), θ).

Supplementary Table S7. (continued). Step 18: `EventSystems`.

Lean name	Statement	Meaning
`event_hasFDerivAt`	$\exists Y_k(\cdot;\theta):\ Y_k(\cdot;\theta_0) = y_k,\ \dot{Y}_k = f_k(Y_k,\theta),\ Y_{k+1}(0;\theta) = R_k(Y_k(L_k;\theta),\theta),\ \theta \mapsto Y_k(s;\theta)$ differentiable at θ_0	If each f_k is C^1, each R_k is differentiable and the nominal segments exist, solutions exist near θ_0 on every segment and are differentiable in θ.

Steps 19–22. Certificates for ODE models

Step 19. `CertifiedODEReduction` (1)

Main local theorem for ODE models: the sensitivity exists and R_var(S) bounds the error of the reduced model.

Supplementary Table S7. (continued). Step 19: `CertifiedODEReduction`.

Lean name	Statement	Meaning
`certified_parameter_reduction`	$\exists S(\cdot):\ S(0) = D_0,\ \dot{S} = \partial_x f\, S + \partial_\theta f,$ and $rE(\mathcal{S}) \le E_S(\mathcal{S}) \Rightarrow \forall \eta > 0, \forall \Delta\theta \approx 0:\ \|\mathcal{F}(\theta_0{+}\Delta\theta) - \mathcal{F}(\theta_0{+}\Delta\theta_S)\| \le \left(\sqrt{(1-r)E} + \eta\right)\|\Delta\theta\|$	Hypotheses: f is C^1 on an open set U containing the nominal trajectory, solutions exist for θ near θ_0 with initial value $x^0(\theta)$ differentiable at θ_0. Conclusion: the sampled sensitivity matrix $\mathcal{S}$ exists, solves the variational equation, and the energy criterion computed from it bounds the trajectory error of fixing the parameters outside S.

Step 20. `CertifiedFiniteODE` (5)

Certificate of cos Δ for ODE models with a Lipschitz constant L_c that is itself bounded formally.

Notation: $\mathcal{S}u$ (`sampledCLM S tk u`): (S(t_1)u, …, S(t_N)u), the sampled sensitivity applied to u.

Supplementary Table S7. (continued). Step 20: `CertifiedFiniteODE`.

Lean name	Statement	Meaning
`sampledCLM_apply`	$(\mathcal{S}u)_{(k,i)} = (S(t_k)u)_i$	Component (k, i) of the sampled sensitivity is component i of S(t_k)u.
`hasFDerivAt_sampled`	$\forall k:\ \partial_\theta x(t_k;\theta) = S(t_k) \Rightarrow D\mathcal{F}(\theta) = \mathcal{S}$	If every sampled state is differentiable, so is the stacked vector, with the stacked sensitivity as derivative.
`jacobianOf_sampledCLM`	$J(\mathcal{S}) = \texttt{sensMatrix}(S,t)$	The matrix of the sampled map is the Nn × p sensitivity matrix.
`sampledCLM_sub_le`	$\|S(t_k) - S'(t_k)\| \le B\ \forall k \Rightarrow \|\mathcal{S} - \mathcal{S}'\| \le \sqrt{N}\,B$	Sampling N times multiplies a uniform bound by $\sqrt{N}$.

Supplementary Table S7. (continued). Step 20: `CertifiedFiniteODE`.

Lean name	Statement	Meaning
`certified_cos_ode`	$\big(\sqrt{E_{S^c}}\|h_{S^c}\| + 2L_c\|h\|^2\big)^2 \le (1-c^2)\big(\|\mathcal{S}_0 h\| - L_c\|h\|^2\big)^2 \Rightarrow \cos\Delta(h) \ge c, \quad L_c = \sqrt{N} M_2 e^{2M_1 T}\dfrac{e^{M_1 T}-1}{M_1}$	Hypotheses: f is C^1 on U; on a convex K $\subseteq$ U, $\|Df\| \le M_1$ and Df is M_2-Lipschitz; solutions exist for θ in a set containing the ball $\bar{B}(\theta_0, \rho_0)$, start from the same initial value and stay in K. Conclusion: the sensitivity $\mathcal{S}_0$ exists and, for every subset S and every $\|h\| \le \rho_0$ with $\|\mathcal{S}_0 h\| - L_c\|h\|^2 > 0$, the displayed inequality certifies $\cos\Delta(h) \ge c$.

Step 21. `ExactLinearCertificate` **(13)**

(A1) Certificates with the exact linear error $\|J(h - h_S)\|$ instead of its energy bound, and (A2) a local version without Lipschitz constants.

Supplementary Table S7. (continued). Step 21: `ExactLinearCertificate`.

Lean name	Statement	Meaning
`cos_certificate_of_bounds`	$0 < A \le \|a\|,\ \|a-b\| \le e,\ e^2 \le (1-c^2)A^2,\ c > 0 \Rightarrow b \ne 0,\ \cos(a,b) \ge c$	Generic certificate lemma.
`cosv_smul`	$\cos(ra, rb) = \cos(a,b) \quad (r > 0)$	The cosine is invariant under a common positive rescaling.
`continuousAt_cosv`	$(a,b) \mapsto \cos(a,b)$ is continuous at $a \ne 0,\ b \ne 0$	Continuity of the cosine.
`keepS_smul`	$(sh)_S = s\,h_S$	Restriction to S is linear.
`norm_linear_part_eq`	$\|Jh - Jh_S\| = \sqrt{\sum_i \big(\sum_{j\notin S} J_{ij} h_j\big)^2}$	The exact linear error is computable from J.
`finite_trajectory_error_exact`	$\|F(\theta_0+h) - F(\theta_0+h_S)\| \le \|J(h-h_S)\| + 2L_c\|h\|^2$	(A1) Finite-perturbation error with the exact linear term.
`exact_le_Rvar_bound`	$\|Jh - Jh_S\| \le \sqrt{E_{S^c}}\,\|h_{S^c}\|$	The exact bound is never worse than the energy bound.
`finite_cos_certificate_exact`	$\|Jh\| - L_c\|h\|^2 > 0,\ \big(\|J(h-h_S)\| + 2L_c\|h\|^2\big)^2 \le (1-c^2)\big(\|Jh\| - L_c\|h\|^2\big)^2 \Rightarrow \cos\Delta(h) \ge c$	(A1) Certificate of cos Δ with the exact linear error.
`certified_cos_ode_exact`	$\big(\|\mathcal{S}_0(h-h_S)\| + 2L_c\|h\|^2\big)^2 \le (1-c^2)\big(\|\mathcal{S}_0 h\| - L_c\|h\|^2\big)^2 \Rightarrow \cos\Delta(h) \ge c$	(A1) ODE version with the formally bounded L_c of certified_cos_ode.
`tendsto_cos_delta`	$Jh \ne 0,\ Jh_S \ne 0 \Rightarrow \lim_{s\to 0^+} \cos\Delta(sh) = \cos(Jh, Jh_S)$	(A2) For small perturbations, cos Δ tends to the linear cosine computed from J.
`eventually_cos_delta_gt`	$c < \cos(Jh, Jh_S) \Rightarrow \cos\Delta(sh) > c$ for all sufficiently small $s > 0$	(A2) Local certificate without L_c.
`exists_cos_delta_gt`	$c < \cos(Jh, Jh_S) \Rightarrow \exists s_0 > 0\ \forall s \in (0, s_0) : \cos\Delta(sh) > c$	Explicit form of the previous statement.
`local_cos_ode`	$\exists \mathcal{S}_0\ \forall S, h :\ c < \cos(\mathcal{S}_0 h, \mathcal{S}_0 h_S) \Rightarrow \exists s_0 > 0\ \forall s \in (0, s_0) :\ \cos\Delta_{\mathcal{F}}(sh) > c$	(A2) ODE version: if f is C^1 around the nominal trajectory and solutions exist near θ_0, the linear cosine computed from the sampled sensitivity certifies cos Δ for small perturbations.

Step 22. `NominalCertificate` **(1)**

Local certificate of cos Δ assuming only the nominal solution.

Supplementary Table S7. (continued). Step 22: `NominalCertificate`.

Lean name	Statement	Meaning
`local_cos_from_nominal`	$\exists x(\cdot;\theta),\ \exists \mathcal{S}_0 :\ \forall S, h :\ c < \cos(\mathcal{S}_0 h, \mathcal{S}_0 h_S) \Rightarrow \exists s_0 > 0\ \forall s \in (0, s_0) :\ \cos \Delta_{\mathcal{F}}(sh) > c$	Combines LocalExistence and ExactLinearCertificate: from the nominal solution alone, solutions near θ_0 and the sensitivity exist, and the linear cosine certifies cos Δ for small perturbations.

Steps 23–31. Hypotheses of the benchmark models (kinetic models and reaction networks)

Step 23. `KineticRegularity` **(11)**

Model class 1: rate laws (mass action, Michaelis–Menten, Hill, . . .) define infinitely differentiable vector fields on their domain.

Notation: $\mathrm{MM}(V, K, x) = \dfrac{Vx}{K + x}$ (`mm V K x`): Michaelis–Menten law; $\mathrm{Hill}_m = \dfrac{Vx^m}{K^m + x^m}$ (`hill V K x m`, `hillR V K x h`): Hill law with integer (m) or real (h) exponent; $\mathrm{noDen}(e)$ (`e.noDen`): e has no division, real power or logarithm (mass action).

Supplementary Table S7. (continued). Step 23: `KineticRegularity`.

Lean name	Statement	Meaning
`KExpr.contDiffAt_eval`	$\mathrm{ok}_e(z) \Rightarrow e \in C^k$ near z $(\forall k)$	Every well-defined rate expression is smooth.
`KExpr.isOpen_ok`	$\{z : \mathrm{ok}_e(z)\}$ is open	The set where e is well defined is open.
`KExpr.ok_of_noDen`	$\mathrm{noDen}(e) \Rightarrow \mathrm{ok}_e(z)\ \forall z$	Mass-action expressions are defined everywhere.
`KExpr.ok_mm`	$K > 0,\ x \geq 0 \Rightarrow \mathrm{ok}(\mathrm{MM}(V, K, x))$	Michaelis–Menten is well defined for positive K and non-negative x.
`KExpr.ok_hill`	$K > 0,\ x \geq 0 \Rightarrow \mathrm{ok}(\mathrm{Hill}_m(V, K, x))$	Hill law with integer exponent.
`KExpr.ok_hillR`	$K > 0,\ x > 0 \Rightarrow \mathrm{ok}(\mathrm{Hill}_h(V, K, x))$	Hill law with real exponent requires x > 0.
`isOpen_domain`	$\mathcal{D}_F$ is open	The domain of a kinetic model is open.
`contDiffOn_field`	$f_F \in C^k(\mathcal{D}_F)$ $(\forall k)$	The vector field of a kinetic model is smooth on its domain.
`domain_eq_univ`	$\forall i :\ \mathrm{noDen}(F_i) \Rightarrow \mathcal{D}_F = \mathbb{R}^n \times \mathbb{R}^p$	Pure mass-action models are defined everywhere.
`kinetic_hasFDerivAt`	$(x_0(t), \theta_0) \in \mathcal{D}_F\ \forall t \Rightarrow \exists x(\cdot;\theta), S :\ S(0) = D_0,\ \dot{S} = \partial_x f_F\, S + \partial_\theta f_F,\ \partial_\theta x(t;\theta_0) = S(t)$	Existence near θ_0 and differentiability for every kinetic model whose nominal solution stays in the domain.
`kinetic_local_cos`	$\exists x(\cdot;\theta), \mathcal{S}_0 :\ c < \cos(\mathcal{S}_0 h, \mathcal{S}_0 h_S) \Rightarrow \exists s_0 > 0\ \forall s \in (0, s_0) :\ \cos \Delta_{\mathcal{F}}(sh) > c$	Local cos Δ certificate for kinetic models.

Step 24. `PositivityInvariance` **(15)**

Quasi-positive models keep concentrations non-negative.

Notation: $y^+ = \max(y, 0)$ (`pos y`): componentwise positive part.

Supplementary Table S7. (continued). Step 24: `PositivityInvariance`.

Lean name	Statement	Meaning
`pos_apply`	$(y^+)_i = \max(y_i, 0)$	Definition.
`pos_nonneg`	$y^+ \geq 0$	The positive part is non-negative.
`pos_eq_self`	$y \geq 0 \Rightarrow y^+ = y$	Non-negative vectors are unchanged.
`dist_pos_le`	$\|y^+ - y'^+\| \leq \|y - y'\|$	The positive part is 1-Lipschitz.
`sub_pos_apply`	$(y - y^+)_i = \min(y_i, 0)$	Negative part.
`pos_apply_eq_zero`	$y_i < 0 \Rightarrow (y^+)_i = 0$	Negative components are set to 0.
`isClosed_nonneg`	$\{y : \ y \geq 0\}$ is closed	The non-negative orthant is closed.
`hasDerivAt_negSq`	$\frac{d}{da}\min(a, 0)^2 = 2\min(a, 0)$	Derivative of the squared negative part.
`nonneg_core`	$x(a) \geq 0,\ \|v(x) - v(x^+)\| \leq K\|x - x^+\|,\ \big(x_i^+ = 0 \Rightarrow v_i(x^+) \geq 0\big) \Rightarrow x(t) \geq 0\ \forall t \in [a, b]$	Core argument: the squared negative part satisfies a Grönwall inequality with zero initial value.
`nonneg_of_quasiPositive`	$\big(y \geq 0,\ y_i = 0 \Rightarrow v_i(y) \geq 0\big),\ x(0) \geq 0 \Rightarrow x(t) \geq 0\ \forall t \in [0, T]$	Quasi-positive C^1 vector fields keep the orthant invariant.
`kinetic_nonneg`	$\mathbb{R}^n_{\geq 0} \times \{\theta\} \subseteq \mathcal{D}_F,\ f_F$ quasi-positive, $x(0) \geq 0 \Rightarrow x(t) \geq 0,\ (x(t), \theta) \in \mathcal{D}_F$	Kinetic version.
`kinetic_hasFDerivAt_of_nonneg`	$\Rightarrow \exists x(\cdot;\theta), S :\ S(0) = D_0,\ \partial_\theta x(t;\theta_0) = S(t)$	Kinetic models defined on the orthant: positivity plus existence near θ_0 and differentiability.
`exampleModel_orth`	$\dot{x} = \theta_0 - \frac{\theta_1 x}{\theta_2 + x} :\ \theta_2 > 0,\ x \geq 0 \Rightarrow (x, \theta) \in \mathcal{D}$	Worked example (Michaelis–Menten degradation): well defined on the orthant.
`exampleModel_qp`	$\theta_0 \geq 0,\ x = 0 \Rightarrow \dot{x} \geq 0$	The example is quasi-positive.
`exampleModel_nonneg`	$\theta_0 \geq 0,\ \theta_2 > 0,\ x(0) \geq 0 \Rightarrow x(t) \geq 0$	The example stays non-negative.

Step 25. `GlobalExistence` (11)

Global existence of the nominal solution under a linear growth bound on the non-negative orthant.

Notation: $\mathrm{clamp}_B(z)_i = \max(0, \min(B, z_i))$ (`clampB B z`): truncation to the box $[0, \mathrm{B}]^n$; $[0, B]^n$ (`box B`): box.

Supplementary Table S7. (continued). Step 25: `GlobalExistence`.

Lean name	Statement	Meaning
`clampB_apply`	$\mathrm{clamp}_B(z)_i = \max(0, \min(B, z_i))$	Definition.
`clampB_nonneg`	$\mathrm{clamp}_B(z) \geq 0$	Truncation is non-negative.
`clampB_le`	$B \geq 0 \Rightarrow \mathrm{clamp}_B(z)_i \leq B$	Truncation is bounded by B.
`clampB_pos`	$\mathrm{clamp}_B(z^+) = \mathrm{clamp}_B(z)$	Truncation ignores the negative part.
`clampB_eq`	$0 \leq z \leq B \Rightarrow \mathrm{clamp}_B(z) = z$	Truncation fixes points of the box.
`dist_clampB_le`	$\|\mathrm{clamp}_B(y) - \mathrm{clamp}_B(y')\| \leq \|y - y'\|$	Truncation is 1-Lipschitz.
`clampB_mem`	$\mathrm{clamp}_B(z) \in [0, B]^n$	Truncation lands in the box.
`convex_box`	$[0, B]^n$ is convex	Convexity of the box.
`isCompact_box`	$[0, B]^n$ is compact	Compactness of the box.
`lipschitzOn_box`	$v \in C^1(U),\ [0, B]^n \subseteq U \Rightarrow \exists L :$ v is L-Lipschitz on $[0, B]^n$	C^1 vector fields are Lipschitz on compact boxes.

Supplementary Table S7. (continued). Step 25: `GlobalExistence`.

Lean name	Statement	Meaning
`exists_global_solution`	v quasi-positive, $\sum_i w_i v_i(z) \le a + b\sum_i w_i z_i$ $(z \ge 0)$, $x^0 \ge 0 \Rightarrow \exists x$: $x(0) = x^0$, $\dot{x} = v(x)$, $x(t) \ge 0$ $\forall t \in [0, T]$	Global existence for every horizon T: the weighted sum Σ w_i x_i grows at most exponentially, so the solution cannot blow up.

Step 26. `StrictExistence` (12)

Global existence when some species must stay strictly positive (for example denominators x_j).

Notation: $\ell_i = \varepsilon$ $(i \in \Sigma)$, 0 $(i \notin \Sigma)$ (`lo sx ε i`): lower corner of the box; $\mathrm{clamp}_Q(z)_i = \max(\ell_i, \min(B, z_i))$ (`clampQ sx ε B z`): truncation to the box Q = Π_i [ℓ_i, B]; Q (`boxQ sx ε B`): box.

Supplementary Table S7. (continued). Step 26: `StrictExistence`.

Lean name	Statement	Meaning
`lo_nonneg`	$\varepsilon \ge 0 \Rightarrow \ell_i \ge 0$	Lower corner is non-negative.
`lo_le`	$0 \le \varepsilon \le B \Rightarrow \ell_i \le B$	Lower corner is below B.
`clampQ_apply`	$\mathrm{clamp}_Q(z)_i = \max(\ell_i, \min(B, z_i))$	Definition.
`clampQ_mem`	$\mathrm{clamp}_Q(z) \in Q$	Truncation lands in the box.
`boxQ_sPos`	$\varepsilon > 0,\ z \in Q \Rightarrow z \in P_\Sigma$	The box lies in the strict region.
`clampQ_pos`	$\mathrm{clamp}_Q(z^+) = \mathrm{clamp}_Q(z)$	Truncation ignores the negative part.
`clampQ_eq`	$\ell \le z \le B \Rightarrow \mathrm{clamp}_Q(z) = z$	Truncation fixes points of the box.
`dist_clampQ_le`	$\lVert \mathrm{clamp}_Q(y) - \mathrm{clamp}_Q(y') \rVert \le \lVert y - y' \rVert$	Truncation is 1-Lipschitz.
`convex_boxQ`	Q is convex	Convexity.
`isCompact_boxQ`	Q is compact	Compactness.
`lipschitzOn_boxQ`	$v \in C^1(U),\ Q \subseteq U \Rightarrow v$ Lipschitz on Q	Lipschitz on the box.
`exists_global_solution_strict`	v quasi-positive on P_Σ, $v_i(z) \ge -C_B z_i$ $(i \in \Sigma,\ z \le B)$, $\sum_i w_i v_i \le a + b\sum_i w_i z_i$, $x^0 \in P_\Sigma \Rightarrow \exists x :\ x(t) \in P_\Sigma\ \forall t \in [0, T]$	Global existence staying in the strict region: species in Σ decay at most exponentially and never reach 0.

Step 27. `KineticCheck` (19)

Verified Boolean checkers that Lean runs on each translated model (SBML → KExpr → Lean).

Notation: isNonneg, isPos, okOrth (`isNonneg pos e, isPos pos e, okOrth pos e`): syntactic tests: e ≥ 0, e > 0, e well defined on the orthant; vanishes$_i$, qp$_i$ (`vanishes i e, qp pos i e`): e = 0 when x_i = 0; quasi-positivity of component i; checkModel, checkInit (`checkModel pos F, checkInit pos G`): all components pass okOrth and qp; initial expressions G pass okOrth and isNonneg.

Supplementary Table S7. (continued). Step 27: `KineticCheck`.

Lean name	Statement	Meaning
`KExpr'.isNonneg_sound`	$y \ge 0,\ \theta_P > 0,\ \mathrm{isNonneg}(e) \Rightarrow e(y, \theta) \ge 0$	Soundness of the non-negativity checker.
`KExpr'.isPos_sound`	$y \ge 0,\ \theta_P > 0,\ \mathrm{isPos}(e) \Rightarrow e(y, \theta) > 0$	Soundness of the positivity checker.
`KExpr'.okOrth_sound`	$y \ge 0,\ \theta_P > 0,\ \mathrm{okOrth}(e) \Rightarrow \mathrm{ok}_e(y, \theta)$	Soundness of the domain checker.

Supplementary Table S7. (continued). Step 27: `KineticCheck`.

Lean name	Statement	Meaning
`KExpr'.vanishes_sound`	$y_i = 0,\ \text{vanishes}_i(e) \Rightarrow e(y,\theta) = 0$	Soundness of the vanishing checker.
`KExpr'.qp_sound`	$y \geq 0,\ y_i = 0,\ \theta_P > 0,\ \text{qp}_i(e) \Rightarrow e(y,\theta) \geq 0$	Soundness of the quasi-positivity checker.
`checkModel_spec`	$\text{checkModel}(F) \Rightarrow \text{okOrth}(F_i) \wedge \text{qp}_i(F_i)\ \forall i$	Unfolding of the model checker.
`checkModel_sound`	$\text{checkModel}(F),\ \theta_P > 0 \Rightarrow \mathbb{R}^n_{\geq 0} \times \{\theta\} \subseteq \mathcal{D}_F\ \wedge\ \big(y \geq 0, y_i = 0 \Rightarrow f_{F,i}(y,\theta) \geq 0\big)$	A checked model is well defined and quasi-positive on the orthant.
`checked_model_hasFDerivAt`	$x_0(0) \geq 0 \Rightarrow x_0(t) \geq 0,\ (x_0(t), \theta_0) \in \mathcal{D}_F,\ \exists x(\cdot;\theta), S:\ S(0) = D_0,\ \partial_\theta x(t;\theta_0) = S(t)$	Checked model: positivity, existence near θ_0 and differentiability.
`checked_segments_hasFDerivAt`	$y_{k+1}(0) = y_k(L_k) \Rightarrow \exists Y_k(\cdot;\theta)$: segments solve and are differentiable in θ	Same for models simulated in segments (condition changes at fixed times).
`qvec_apply`	$\iota(q)_i = q_i$	Rational vectors as real vectors.
`posParams_qvec`	$\text{checkPosParams}(\theta_q) \Rightarrow \iota(\theta_q)_P > 0$	The rational parameter values are positive where declared.
`nonneg_qvec`	$\text{checkNonneg}(x_q) \Rightarrow \iota(x_q) \geq 0$	The rational initial values are non-negative.
`checked_model_final`	$\text{checkModel} \wedge \text{checkPosParams} \wedge \text{checkNonneg} \Rightarrow x_0 \geq 0,\ \text{in } \mathcal{D}_F,\ \exists x(\cdot;\theta), S:\ S(0) = 0,\ \partial_\theta x(t;\theta_q) = S(t)$	End-to-end theorem for a checked model with rational parameters and initial values.
`checked_segments_final`	same, for segmented models	End-to-end theorem for segmented models.
`checkInit_spec`	$\text{checkInit}(G) \Rightarrow \text{okOrth}(G_i) \wedge \text{isNonneg}(G_i)$	Unfolding of the initial-value checker.
`nonneg_zero`	$0 \geq 0$	The zero vector is non-negative.
`init_differentiable`	$\text{checkInit}(G) \Rightarrow \theta \mapsto G(\theta)$ differentiable at θ_0	Parameter-dependent initial values are differentiable.
`init_nonneg`	$\text{checkInit}(G) \Rightarrow G(\theta_0) \geq 0$	Parameter-dependent initial values are non-negative.
`checked_model_final_init`	$x(0;\theta) = G(\theta) :\ \ldots,\ S(0) = \partial_\theta G(\theta_q)$	End-to-end theorem with parameter-dependent initial values.

Step 28. `KineticNetwork` (18)

Reaction networks f_i = Σ_k ν_ik r_k: complete automatic check including global existence (C4).

Notation: checkNet, checkGrowth (`checkNet pos Rx, checkGrowth pos c Rx`): every reaction passes the domain and quasi-positivity test; every reaction passes the linear-growth test with weights w; $\ell_e(\theta)$ (`lowval pos θ e`): positive constant lower bound of e (from lowerpos); A_e, B_e (`linBound pos θ e`): coefficients of a linear bound e ≤ A + B Σ_j y_j.

Supplementary Table S7. (continued). Step 28: `KineticNetwork`.

Lean name	Statement	Meaning
`eval_netExpr`	$f_i(z) = \sum_k \nu_{ik}\, r_k(z)$	The network vector field is the stoichiometric sum.
`ok_netExpr`	$\forall k:\ \text{ok}_{r_k}(z) \Rightarrow \text{ok}_{f_i}(z)$	Well-defined rates give a well-defined field.
`coef_mul_nonneg`	$\text{termOK}(r_k, \nu_k),\ y \geq 0,\ y_i = 0 \Rightarrow \nu_{ik}\, r_k(y,\theta) \geq 0$	Each reaction respects quasi-positivity (consumption terms vanish when the species is absent).

Supplementary Table S7. (continued). Step 28: `KineticNetwork`.

Lean name	Statement	Meaning
`checkNet_sound`	checkNet $\Rightarrow \mathbb{R}^n_{\geq 0} \times \{\theta\} \subseteq \mathcal{D} \ \wedge \ f$ quasi-positive	Soundness of the network checker.
`eval_varfree`	e without state variables $\Rightarrow e(y,\theta) = e(y',\theta)$	Constant expressions do not depend on the state.
`ssum_nonneg`	$y \geq 0 \Rightarrow \sum_j y_j \geq 0$	Sum of a non-negative vector.
`lowerpos_sound`	lowerpos(e), $y \geq 0 \Rightarrow 0 < \ell_e(\theta) \leq e(y,\theta)$	Denominators bounded away from 0.
`linBound_sound`	linOK(e), $y \geq 0 \Rightarrow A_e, B_e \geq 0$, $e(y,\theta) \leq A_e + B_e \sum_j y_j$	Linear upper bound of a rate law.
`sum_c_coef`	$\sum_i w_i \nu_{ik} = \langle w, \nu_k \rangle$	Weighted stoichiometry.
`sum_c_netExpr`	$\sum_i w_i f_i = \sum_k \langle w, \nu_k \rangle \, r_k$	Weighted sum of the vector field.
`checkGrowth_sound`	checkGrowth $\Rightarrow \exists a, b \geq 0 :$ $\sum_i w_i f_i(y,\theta) \leq a + b \sum_i w_i y_i \ (y \geq 0)$	Soundness of the growth checker.
`checkNet_append`	checkNet$(A +\!\!+ B) =$ checkNet$(A) \wedge$ checkNet(B)	The check can be split over parts of a large model.
`checkGrowth_append`	checkGrowth$(A +\!\!+ B) =$ checkGrowth$(A) \wedge$ checkGrowth(B)	Same for the growth check.
`c_ge_one`	checkGrowth $\Rightarrow w_i \geq 1$	Weights are at least 1.
`network_exists`	checkNet, checkGrowth, $x^0 \geq 0 \Rightarrow \exists x :$ $\dot{x} = f(x,\theta_0)$, $x(t) \geq 0 \ \forall t \in [0,T]$	Global existence for checked networks.
`network_final`	$\Rightarrow \exists x_0 \geq 0$ on $[0,T]$, $\exists x(\cdot;\theta), S :$ $S(0) = 0$, $\partial_\theta x(t;\theta_q) = S(t)$	End-to-end theorem for a network with rational parameters and initial values. Used by the models Bertozzi, Blasi, Boehm, Brannmark, Bruno, Chen, Froehlich, Isensee, Lang, Laske, Raimundez, Sneyd, Weber, Zhao, Zheng.
`network_final_init`	$x(0;\theta) = G(\theta) : \ \ldots, \ S(0) = \partial_\theta G(\theta_q)$	Same with parameter-dependent initial values. Used by the models Bachmann, Fujita, Lucarelli, Raia, SalazarCavazos.
`network_segments_final`	same, for segmented models	Same for models simulated in segments. Used by the models Giordano.

Step 29. `RiccatiNetwork` (9)

Existence on a finite horizon under quadratic growth (Riccati bound); used for Crauste.

Notation: $\varphi(y) = \sum_i w_i y_i$ (`phi c y`): weighted total; Q (`Qof (tot c θq Rx)`): Riccati constant max(A, B/2, C) of the bound Σ w_i f_i $\leq$ A + Bφ + Cφ^2.

Supplementary Table S7. (continued). Step 29: `RiccatiNetwork`.

Lean name	Statement	Meaning
`exists_solution_riccati`	$\sum_i w_i v_i(z) \leq Q(\varphi(z)+1)^2$, $(1+\delta)Q(\varphi(x^0)+1)T < 1 \Rightarrow \exists x \geq 0$ on $[0,T]$	Quasi-positive field with quadratic growth: the solution exists up to the Riccati blow-up time.

Supplementary Table S7. (continued). Step 29: `RiccatiNetwork`.

Lean name	Statement	Meaning
`qval_sound`	$\mathrm{qc}(e) \Rightarrow e(y, \iota(\theta_q)) = \mathrm{qval}(e) \in \mathbb{Q}$	Constant expressions evaluate exactly in rational arithmetic.
`qnn_sound`	$\mathrm{qnn}(e) \Rightarrow e = \mathrm{qval}(e) \geq 0$	Non-negative rational constants.
`var_le_phi`	$w > 0,\ y \geq 0 \Rightarrow y_j \leq \varphi(y)/w_j$	Each species is bounded by the weighted total.
`linV_sound`	$\mathrm{linC}(e) \Rightarrow 0 \leq e \leq A + B\varphi(y),\ A, B \geq 0$	Linear bound of a rate law.
`prodV_sound`	$0 \leq a \leq A_1 + B_1\varphi,\ 0 \leq b \leq A_2 + B_2\varphi \Rightarrow 0 \leq ab \leq A + B\varphi + C\varphi^2$	Product of two linear bounds is a quadratic bound.
`quadV_sound`	$\mathrm{quadC}(e) \Rightarrow 0 \leq e \leq A + B\varphi + C\varphi^2$	Quadratic bound of a rate law.
`tot_sound`	$\sum_k \langle w, \nu_k \rangle r_k \leq A + B\varphi + C\varphi^2$	Quadratic bound of the weighted vector field (sum over reactions with positive weight).
`riccati_final`	$\mathrm{checkRiccati}:\ 1.1\,Q(\varphi(x_q)+1)T_q < 1,\ T \leq T_q \Rightarrow \exists x_0 \geq 0 \text{ on } [0,T],\ \exists x(\cdot;\theta), S:\ \partial_\theta x(t;\theta_q) = S(t)$	End-to-end theorem on a finite horizon (Crauste: T ≤ 1). Used by the models Crauste.

Step 30. `StrictNetwork` (46)

Reaction networks with strictly positive domains (e.g. βSI/N, Hill laws with real exponents): checkers and their soundness.

Notation: UB(θ) (`UB ub` θ): declared lower/upper bounds of parameters hold; PP(θ) (`PP pos ub` θ): θ_P > 0 and UB(θ); sNonneg, snp, expLe1, sPos, sOk: syntactic tests on P_Σ: e ≥ 0, e ≤ 0, 0 ≤ e ≤ 1 (exponential of a non-positive term), e > 0, e well defined; Bnd(e) (`BndB sx` θ `bx e`): ∀B ∃M: 0 ≤ e(y,θ) ≤ M on P_Σ (and on [0,B]^n if bx); Cons$_i$(e) (`ConsB sx` θ `i e`): ∀B ∃K: 0 ≤ e(y,θ) ≤ K·y_i on P_Σ ∩ [0,B]^n (consumption term proportional to y_i); checkNetS, checkGrowthS (`checkNetS pos ub sx Rx, checkGrowthS pos ub c Rx`): network checks on the strict region.

Supplementary Table S7. (continued). Step 30: `StrictNetwork`.

Lean name	Statement	Meaning
`keq_sound`	$\mathrm{keq}(a,b) \Rightarrow a = b$	Syntactic equality is sound.
`qge1_sound`	$\mathrm{qge1}(e) \Rightarrow e \geq 1$	Rational constant ≥ 1.
`qgt1_sound`	$\mathrm{qgt1}(e) \Rightarrow e > 1$	Rational constant > 1.
`addlb_sound`	$\theta_j \geq r \geq -q \Rightarrow q + \theta_j \geq 0$	Sum of a constant and a parameter with a declared lower bound.
`subub_sound`	$0 < \theta_j \leq r \leq q \Rightarrow 0 \leq q - \theta_j \leq q$	Difference of a constant and a parameter with a declared upper bound.
`ub_qvec`	$\mathrm{checkUB}(\theta_q) \Rightarrow \mathrm{UB}(\iota(\theta_q))$	The rational parameters satisfy the declared bounds.
`sign_sound`	$\mathrm{sNonneg}(e) \Rightarrow e \geq 0,\ \mathrm{snp}(e) \Rightarrow e \leq 0,\ \mathrm{expLe1}(e) \Rightarrow 0 \leq e \leq 1$	Joint soundness of the sign checkers (y ≥ 0, PP(θ)).
`sNonneg_sound`	$\mathrm{sNonneg}(e) \Rightarrow e(y,\theta) \geq 0$	Non-negativity checker.
`expLe1_sound`	$\mathrm{expLe1}(e) \Rightarrow 0 \leq e(y,\theta) \leq 1$	Exponential of a non-positive term.
`sPos_sound`	$y \in P_\Sigma,\ \mathrm{sPos}(e) \Rightarrow e(y,\theta) > 0$	Positivity checker on the strict region.
`sOk_sound`	$y \in P_\Sigma,\ \mathrm{sOk}(e) \Rightarrow \mathrm{ok}_e(y,\theta)$	Domain checker on the strict region.
`snp_sound`	$\mathrm{snp}(e) \Rightarrow e(y,\theta) \leq 0$	Non-positivity checker.
`summand_sound`	$\mathrm{summand}(X,d) \Rightarrow 0 \leq X \leq d$	X is a non-negative summand of d, so X/d ∈ [0, 1].

Supplementary Table S7. (continued). Step 30: `StrictNetwork`.

Lean name	Statement	Meaning
`bndB_const`	e constant, $e \geq 0 \Rightarrow \mathrm{Bnd}(e)$	Constants are bounded.
`bndB_div_lowerpos`	$\mathrm{Bnd}(a),\ d \geq \ell > 0 \Rightarrow \mathrm{Bnd}(a/d)$	Division by a term bounded away from 0.
`bndB_summand`	$\mathrm{summand}(a,d) \Rightarrow \mathrm{Bnd}(a/d)$	Saturating terms a/(a + ...) are bounded.
`bndB_mul`	$\mathrm{Bnd}(a), \mathrm{Bnd}(b) \Rightarrow \mathrm{Bnd}(ab)$	Products of bounded terms.
`bndB_add`	$\mathrm{Bnd}(a), \mathrm{Bnd}(b) \Rightarrow \mathrm{Bnd}(a+b)$	Sums of bounded terms.
`bndB_frac_mul_l`	$\mathrm{Bnd}(u/d), \mathrm{Bnd}(v) \Rightarrow \mathrm{Bnd}(uv/d)$	Bounded fraction times bounded factor.
`bndB_frac_mul_r`	$\mathrm{Bnd}(u), \mathrm{Bnd}(v/d) \Rightarrow \mathrm{Bnd}(uv/d)$	Same, factor on the left.
`bndB_pair`	$\mathrm{Bnd}(a_1/d_1), \mathrm{Bnd}(a_2/d_2) \Rightarrow \mathrm{Bnd}\big(a_1a_2/(d_1d_2)\big)$	Product of two bounded fractions.
`bndB_pair'`	$\mathrm{Bnd}(a_1/d_2), \mathrm{Bnd}(a_2/d_1) \Rightarrow \mathrm{Bnd}\big(a_1a_2/(d_1d_2)\big)$	Same, crossed pairing.
`consB_mul_l`	$\mathrm{Cons}_i(a), \mathrm{Bnd}(b) \Rightarrow \mathrm{Cons}_i(ab)$	Proportional to y_i times bounded.
`consB_mul_r`	$\mathrm{Bnd}(a), \mathrm{Cons}_i(b) \Rightarrow \mathrm{Cons}_i(ab)$	Same, factor on the left.
`consB_div_lowerpos`	$\mathrm{Cons}_i(a),\ d \geq \ell > 0 \Rightarrow \mathrm{Cons}_i(a/d)$	Proportional term divided by a positive term.
`bndB_congr`	$e \equiv e',\ \mathrm{Bnd}(e') \Rightarrow \mathrm{Bnd}(e)$	Bounds depend only on the value of the expression.
`consB_congr`	$e \equiv e',\ \mathrm{Cons}_i(e') \Rightarrow \mathrm{Cons}_i(e)$	Same for proportional bounds.
`eval_pair`	$\dfrac{a_1a_2}{d_1d_2} = \dfrac{a_1}{d_1} \cdot \dfrac{a_2}{d_2}$	Algebraic identity used to pair factors.
`eval_pair'`	$\dfrac{a_1a_2}{d_1d_2} = \dfrac{a_1}{d_2} \cdot \dfrac{a_2}{d_1}$	Crossed pairing.
`consB_var`	$\mathrm{Cons}_i(y_i)$	A species is proportional to itself.
`consB_npow`	$k \geq 1,\ \mathrm{Cons}_i(a),\ \mathrm{Bnd}(a) \Rightarrow \mathrm{Cons}_i(a^k)$	Powers of proportional terms.
`bndB_npow`	$\mathrm{Bnd}(a) \Rightarrow \mathrm{Bnd}(a^k)$	Powers of bounded terms.
`bndB_exp_np`	$a \leq 0 \Rightarrow \mathrm{Bnd}(e^a)$	Exponential of a non-positive term is in (0, 1].
`bndB_var`	$\mathrm{Bnd}(y_j)$ on $[0,B]^n$	Species are bounded on the box.
`bq_of`	$\big(\mathrm{Bnd}(a) \wedge d \geq \ell > 0\big) \vee \mathrm{satb} \vee \mathrm{bpair} \Rightarrow \mathrm{Bnd}(a/d)$	Assembles the three ways of bounding a fraction.
`cq_of`	$\big(\mathrm{Cons}_i(a) \wedge d \geq \ell > 0\big) \vee \mathrm{cpair} \Rightarrow \mathrm{Cons}_i(a/d)$	Assembles the ways of proving a proportional fraction.
`bnd_sound`	$\mathrm{bnd}, \mathrm{satb}, \mathrm{bpair} \Rightarrow \mathrm{Bnd};\quad \mathrm{cons}, \mathrm{cpair} \Rightarrow \mathrm{Cons}_i$	Joint soundness of the bound checkers.
`term_qp`	$\mathrm{termOKS}(r_k, \nu_k),\ y \in P_\Sigma,\ y_i = 0 \Rightarrow \nu_{ik} r_k(y,\theta) \geq 0$	Quasi-positivity of each reaction on the strict region.
`term_decay`	$i \in \Sigma \Rightarrow \forall B\ \exists K \geq 0 :\ \nu_{ik} r_k(y,\theta) \geq -K y_i\ (y \in P_\Sigma,\ y \leq B)$	Consumption of a strictly positive species is at most proportional to it.
`checkNetS_sound`	$\mathrm{checkNetS} \Rightarrow P_\Sigma \times \{\theta\} \subseteq \mathcal{D},\ f$ quasi-positive, $f_i \geq -C_B y_i\ (i \in \Sigma)$	Soundness of the strict network checker.
`term_growth`	$\mathrm{growthOKS} \Rightarrow \exists A, B \geq 0 :\ \langle w, \nu_k \rangle r_k(y,\theta) \leq A + B \sum_j y_j$	Growth bound for one reaction.
`checkGrowthS_sound`	$\mathrm{checkGrowthS} \Rightarrow \exists a, b \geq 0 :\ \sum_i w_i f_i \leq a + b \sum_i w_i y_i\ (y \in P_\Sigma)$	Growth bound for the network.
`c_ge_oneS`	$\mathrm{checkGrowthS} \Rightarrow w_i \geq 1$	Weights are at least 1.
`strict_exists`	$x^0 \in P_\Sigma \Rightarrow \exists x :\ \dot{x} = f(x,\theta_0),\ x(t) \in P_\Sigma\ \forall t \in [0,T]$	Global existence in the strict region.

Supplementary Table S7. (continued). Step 30: `StrictNetwork`.

Lean name	Statement	Meaning
`init_sPos`	$\mathrm{checkInitS}(G),\ \mathrm{PP}(\theta_0) \Rightarrow G(\theta_0) \in P_\Sigma$	Initial values lie in the strict region.
`strict_final_init`	$\Rightarrow x_0(t) \in P_\Sigma \cap \mathcal{D},\ \exists x(\cdot;\theta), S:\ S(0) = \partial_\theta G(\theta_q),\ \partial_\theta x(t;\theta_q) = S(t)$	End-to-end theorem for networks with strictly positive domains. Used by the models Armistead, Borghans, Elowitz, Okuonghae, Rahman.

Step 31. `IntervalInit` (5)

Parameter-dependent initial values certified with rational interval arithmetic.

Notation: $[l, u]$ (`ival θq e = some (l, u)`): rational interval enclosing the value of e at θ_q; checkInitI (`checkInitI sx θq G`): interval check that G(θ_q) ≥ 0 and > 0 on Σ.

Supplementary Table S7. (continued). Step 31: `IntervalInit`.

Lean name	Statement	Meaning
`mul_mem_corners`	$x \in [a_1, a_2],\ y \in [b_1, b_2] \Rightarrow \min_{i,j} a_i b_j \leq xy \leq \max_{i,j} a_i b_j$	Interval multiplication rule.
`imul_sound`	$x \in a,\ y \in b \Rightarrow xy \in a \otimes b$	Rational interval product encloses the real product.
`ival_sound`	$\mathrm{ival}(e) = [l, u] \Rightarrow \mathrm{ok}_e(y, \iota(\theta_q))\ \wedge\ l \leq e(y, \iota(\theta_q)) \leq u$	Interval evaluation is sound (real powers restricted to square roots).
`initI_sound`	$\mathrm{checkInitI}(G) \Rightarrow (0, \iota(\theta_q)) \in \mathcal{D}_G,\ G(\iota(\theta_q)) \in P_\Sigma$	The initial values are well defined and lie in the strict region.
`strict_final_initI`	$x(0;\theta) = G(\theta):\ x_0(t) \in P_\Sigma \cap \mathcal{D},\ \exists x(\cdot;\theta), S:\ S(0) = \partial_\theta G(\theta_q),\ \partial_\theta x(t;\theta_q) = S(t)$	End-to-end theorem with initial values checked by interval arithmetic. Used by the models Fiedler.

Step 32. Identifiability of the reduced problem

Step 32. `ProfileIdentifiability` (18)

Practical identifiability of the reduced problem by profile likelihood: lower bound of the cost increase, profiled information, guaranteed half-width t* of the confidence interval, local structural identifiability.

Notation: $\Delta\chi^2(\delta) = \|r(\hat{\theta} + \delta)\|^2 - \|r_0\|^2$ (`dchi2 R δ`): increase of the cost from the optimum; A (`A`): sensitivity of the noise-scaled residuals at the optimum; $I_j = 1/[(A^\top A)^{-1}]_{jj}$ (`I`): profiled information of parameter j; t^* (`tstar, tstarGrad`): guaranteed half-width of the confidence interval of log θ_j.

Supplementary Table S7. (continued). Step 32: `ProfileIdentifiability`.

Lean name	Statement	Meaning
`dchi2_lower`	$\langle r_0, A\delta\rangle = 0,\ \|r(\delta) - r_0 - A\delta\| \leq \varepsilon \leq \|A\delta\| \Rightarrow \Delta\chi^2(\delta) \geq (\|A\delta\| - \varepsilon)^2 - 2\|r_0\|\varepsilon$	Lower bound of the cost increase at a stationary point with bounded nonlinear remainder.
`info_of_inverse`	$\langle Zw, Zu\rangle = u_j\ \forall u,\ w_j > 0 \Rightarrow u_j^2/w_j \leq \|Zu\|^2\ \forall u$	Profiled information from the column w of the inverse Gram matrix.
`info_of_matrix`	$\det(M^\top M) \neq 0,\ [(M^\top M)^{-1}]_{jj} > 0 \Rightarrow u_j^2/[(M^\top M)^{-1}]_{jj} \leq \|Mu\|^2$	The same with the computed matrix inv(J.T @ J).

Supplementary Table S7. (continued). Step 32: `ProfileIdentifiability`.

Lean name	Statement	Meaning
`dchi2_ge_of_coord`	$I\varphi(\delta)^2 \le \|A\delta\|^2,\ \|\varphi(\delta)\| \ge t^* = \dfrac{\varepsilon + \sqrt{c + 2\|r_0\|\varepsilon}}{\sqrt{I}} \Rightarrow \Delta\chi^2(\delta) \ge c$	The profile exceeds the threshold c beyond t*.
`confidence_coord_lt`	$\delta \in F,\ \Delta\chi^2(\delta) < c \Rightarrow \|\varphi(\delta)\| < t^*$	The confidence region lies in the slab —δ_j— < t*.
`practically_identifiable`	$t^* < d_-,\ t^* < d_+ \Rightarrow$ PracticallyIdentifiable	If t* is smaller than the distances to the bounds, θ_j is practically identifiable within the bounds.
`tstar_linear`	$t^*(1/w, 0, c, r_0) = \sqrt{c\,w}$	With $\varepsilon = 0$, t* is the Wald half-width $\sqrt{(c\ [(A^\top A)^{-1}]_{jj})}$.
`local_injective`	$\sigma\|\delta\| \le \|A\delta\|,\ \|G(\delta) - G(0) - A\delta\| \le M\|\delta\|^2,\ 0 < \|\delta\| < \sigma/M \Rightarrow G(\delta) \ne G(0)$	Local structural identifiability: distinct parameters near the optimum give distinct outputs.
`dchi2_lower_grad`	$\|\langle r_0, A\delta\rangle\| \le \gamma\|A\delta\| \Rightarrow \Delta\chi^2(\delta) \ge (\|A\delta\| - \varepsilon)^2 - 2\gamma\|A\delta\| - 2\|r_0\|\varepsilon$	Lower bound with a residual gradient γ (numerical optimum).
`tstarGrad_zero`	$t^*_{\gamma=0} = t^*$	With $\gamma = 0$ the two half-widths coincide.
`dchi2_ge_of_coord_grad`	$\|\varphi(\delta)\| \ge \dfrac{\varepsilon + \gamma + \sqrt{\gamma^2 + c + 2\|r_0\|\varepsilon + 2\gamma\varepsilon}}{\sqrt{I}} \Rightarrow \Delta\chi^2(\delta) \ge c$	The profile exceeds c beyond t* (numerical optimum).
`practically_identifiable_grad`	$t^*_\gamma < d_-,\ t^*_\gamma < d_+ \Rightarrow$ PracticallyIdentifiable	Main theorem with a numerical optimum.
`stat_of_projection`	$\langle r_0 - p, A\delta\rangle = 0\ \forall\delta,\ \|p\| \le \gamma \Rightarrow \|\langle r_0, A\delta\rangle\| \le \gamma\|A\delta\|$	γ is the norm of the projection of r_0 onto the range of A.
`practically_identifiable_grad_tail`	$t^*_\gamma \le t_\pm < d_\pm,\ \|r(\delta) - r_0 - A\delta\| \le \varepsilon$ only for $\delta \in F,\ \varphi(\delta) \le -t_- \vee \varphi(\delta) \ge t_+ \Rightarrow$ PracticallyIdentifiable	Main theorem with the remainder bound needed only on the tail of the profile.
`dchi2_lower_rel`	$\|\langle r_0, A\delta\rangle\| \le \gamma\|A\delta\|,\ \|r(\delta) - r_0 - A\delta\| \le \eta\|A\delta\| \Rightarrow \Delta\chi^2(\delta) \ge (1-\eta)^2\|A\delta\|^2 - 2(\gamma + \eta\|r_0\|)\|A\delta\|$	Lower bound of the cost increase with a relative nonlinear remainder.
`quad_of_root`	$q > 0,\ s \ge 0,\ s^2 = b^2 + qc,\ b + s \le qa \Rightarrow c \le qa^2 - 2ba$	Elementary quadratic inequality used by the relative version.
`dchi2_ge_of_coord_rel`	$\|\varphi(\delta)\| \ge t^*_{\mathrm{rel}} = \dfrac{b + \sqrt{b^2 + (1-\eta)^2 c}}{(1-\eta)^2\sqrt{I}},\ b = \gamma + \eta\|r_0\| \Rightarrow \Delta\chi^2(\delta) \ge c$	The profile exceeds c beyond the relative half-width ($\eta < 1$).
`practically_identifiable_rel_tail`	$\eta < 1,\ t^*_{\mathrm{rel}} \le t_\pm < d_\pm,\ \|r(\delta) - r_0 - A\delta\| \le \eta\|A\delta\|$ on the tail $\Rightarrow$ PracticallyIdentifiable	Main theorem with a relative remainder bound on the tail of the profile.

Supplementary Note: Verified benchmark models

This note lists, for each of the 28 benchmark models verified in Lean, the data contained in its file in `lean_proofs/Models/`, the checks executed by the Lean kernel on those data and the final theorem that results. Each file was generated from the SBML model and the PEtab parameter table of the benchmark by `lean_translation/sbml_to_lean.py`, with every number written as an exact rational. The checks are Boolean functions proved sound in steps 27–31 (Supplementary Note *Logical sequence of the formal proofs*); each is evaluated by the Lean kernel (`decide +kernel`), so its value `true` is part of the proof. The final theorem `Models.<model>.final` then has no pending hypotheses: it states that the model satisfies the assumptions of steps 14–22 (smooth vector field on a domain containing the trajectory, existence of the nominal solution on the horizon, non-negativity, and differentiability of the solution with respect to the parameters). `#print axioms` for every final theorem reports only `propext`,

`Classical.choice` and `Quot.sound` (`CheckModelAxioms.lean`). Supplementary Table S8 summarises the models, the next subsections describe the independent check of the translation, a complete example (Boehm) and the benchmark problems that were not verified, and the last subsections give the data of each model.

Supplementary Table S8. Benchmark models verified in Lean. n: states (species, plus time when a rate law depends on it); p: parameters; R: rate terms (reactions after splitting reversible ones); $|\Sigma|$: species proved strictly positive; horizon: time interval on which existence of the solution is proved; final theorem: family of the end-to-end theorem that the model instantiates (Supplementary Note *Statements of the formal theorems*); time: compilation time of the model files (s); fidelity: largest relative difference between the Lean translation and the SBML model evaluated by libRoadRunner (vector field, parameter values and initial values).

Model	n	p	R	$\lvert\Sigma\rvert$	Horizon	Final theorem	Time	Fidelity
Armistead_Cell DeathDis2024	4	10	8	4	all $T \geq 0$	`strict_final_init`	5	0
Bachmann_MSB2011	25	27	43	–	all $T \geq 0$	`network_final_init`	12	3×10^{-15}
Bertozzi_PNAS2020	3	0	2	–	all $T \geq 0$	`network_final`	188	0
Blasi_Cell Systems2016	16	8	36	–	all $T \geq 0$	`network_final`	8	4×10^{-15}
Boehm_JProteome Res2014	9	6	10	–	all $T \geq 0$	`network_final`	7	4×10^{-16}
Borghans_Biophys Chem1997	3	20	9	3	all $T \geq 0$	`strict_final_init`	5	2×10^{-16}
Brannmark_JBC2010	9	15	14	–	all $T \geq 0$	`network_final`	6	1×10^{-13}
Bruno_JExpBot2016	7	6	6	–	all $T \geq 0$	`network_final`	4	2×10^{-16}
Chen_MSB2009	501	152	1,243	–	all $T \geq 0$	`network_final`	1080	8×10^{-14}
Crauste_Cell Systems2017	5	12	12	–	$T \leq 1$	`riccati_final`	6	0
Elowitz_Nature2000	8	18	18	8	all $T \geq 0$	`strict_final_init`	6	2×10^{-16}
Fiedler_BMCSyst Biol2016	7	12	8	6	all $T \geq 0$	`strict_final_initI`	15	0
Froehlich_Cell Systems2018	1,228	4,088	4,487	–	all $T \geq 0$	`network_final`	2021	6×10^{-14}
Fujita_SciSignal2010	9	16	13	–	all $T \geq 0$	`network_final_init`	5	1×10^{-16}
Giordano_Nature2020	10	43	96	–	6 segments	`network_segments_final`	8	4×10^{-16}
Isensee_JCB2018	25	32	34	–	all $T \geq 0$	`network_final`	8	1×10^{-14}
Lang_PLOSComput Biol2024	124	164	663	–	all $T \geq 0$	`network_final`	264	2×10^{-15}
Laske_PLOSComput Biol2019	34	6	51	–	all $T \geq 0$	`network_final`	10	0
Lucarelli_Cell Systems2018	33	69	67	–	all $T \geq 0$	`network_final_init`	26	4×10^{-15}
Okuonghae_Chaos SolitonsFractals2020	8	14	11	6	all $T \geq 0$	`strict_final_init`	5	7×10^{-16}
Rahman_MBS2016	7	9	19	4	all $T \geq 0$	`strict_final_init`	5	1×10^{-15}
Raia_Cancer Research2011	14	18	14	–	all $T \geq 0$	`network_final_init`	5	6×10^{-16}
Raimundez_PCB2020	22	57	91	–	all $T \geq 0$	`network_final`	15	1×10^{-12}

Supplementary Table S8. Benchmark models verified in Lean. n: states (species, plus time when a rate law depends on it); p: parameters; R: rate terms (reactions after splitting reversible ones); $|\Sigma|$: species proved strictly positive; horizon: time interval on which existence of the solution is proved; final theorem: family of the end-to-end theorem that the model instantiates (Supplementary Note *Statements of the formal theorems*); time: compilation time of the model files (s); fidelity: largest relative difference between the Lean translation and the SBML model evaluated by libRoadRunner (vector field, parameter values and initial values).

Model	n	p	R	$\|\Sigma\|$	Horizon	Final theorem	Time	Fidelity
SalazarCavazos_ MBoC2020	75	6	457	–	all $T \geq 0$	`network_final_ init`	318	2×10^{-13}
Sneyd_PNAS2002	6	14	16	–	all $T \geq 0$	`network_final`	5	4×10^{-15}
Weber_BMC2015	7	26	18	–	all $T \geq 0$	`network_final`	6	0
Zhao_QuantBiol2020	4	21	3	–	all $T \geq 0$	`network_final`	5	0
Zheng_PNAS2012	15	45	44	–	all $T \geq 0$	`network_final`	7	1×10^{-15}

Agreement of the translation with the SBML models

Lean checks the expressions it is given; whether these expressions are the model of the benchmark is a separate question. The script `lean_translation/verify_translation.py` answers it without using the translator: it reads the rate laws, the stoichiometry, the parameter values θ_q and the initial values directly from the Lean files, evaluates the vector field in floating point, and compares it with the derivatives computed by libRoadRunner from the original SBML model with the nominal parameters of the PEtab table, at the initial state and at five random positive states. It also compares θ_q with the PEtab nominal values and the initial values (or $G(\theta_q)$) with those of the SBML model. Over the 28 models, the largest relative difference is 1×10^{-12}, i.e. agreement to rounding error (Supplementary Table S8, last column). As a control, changing one parameter value, one initial value or one stoichiometric coefficient of the Lean file by 1% is detected (relative differences of 10^{-2}). The comparison covers the model without condition-specific overrides, which is what the Lean files describe; for models defined by segments it covers the first segment.

A complete example: Boehm_JProteomeRes2014

The model of STAT5 dimerisation by Boehm et al. is small enough to follow every step. Its file is `lean_proofs/Models/Boehm_JProteomeRes2014.lean` (about 60 lines).

1. Model. The SBML model has 8 species, and the stimulus decays as $e^{-\theta_1 \tau}$, where τ is time; the translator therefore adds τ as a ninth state with $\dot{\tau} = 1$. The six estimated parameters are $\theta_1 =$ Epo_degradation_BaF3 (0.02698), $\theta_2 =$ k_exp_hetero (1.001×10^{-5}), $\theta_3 =$ k_exp_homo (0.00617), $\theta_4 =$ k_imp_hetero (0.01637), $\theta_5 =$ k_imp_homo (9.775×10^4), $\theta_6 =$ k_phos (1.577×10^4). The vector field is $f_i = \sum_k \nu_{ik} v_k$ with the following rate terms v_k and stoichiometric coefficients ν_{ik}, listed in Supplementary Table S9 (the constants come from compartment volumes and fixed parameters of the SBML model).

Supplementary Table S9. Rate terms and stoichiometry of the Lean model of Boehm_JProteomeRes2014.

k	Rate term v_k	Stoichiometry $\nu_{\cdot k}$
1	$\frac{7}{5} \frac{1}{8000000} e^{(-\theta_1)\,\tau}\, \mathrm{STAT5A}^2\, \theta_6$	STAT5A: -10/7, pApA: 5/7
2	$\frac{7}{5} \frac{1}{8000000} e^{(-\theta_1)\,\tau}\, \mathrm{STAT5A}\, \mathrm{STAT5B}\, \theta_6$	STAT5A: -5/7, STAT5B: -5/7, pApB: 5/7
3	$\frac{9}{20}\, \theta_3\, \mathrm{nucpApA}$	STAT5A: 10/7, nucpApA: -20/9
4	$\frac{9}{20}\, \theta_2\, \mathrm{nucpApB}$	STAT5A: 5/7, STAT5B: 5/7, nucpApB: -20/9

5	$\frac{7}{5}\,\frac{1}{8000000}\,e^{(-\theta_1)\,\tau}\,\mathrm{STAT5B}^2\,\theta_6$	STAT5B: -10/7, pBpB: 5/7
6	$\frac{9}{20}\,\theta_3\,\mathrm{nucpBpB}$	STAT5B: 10/7, nucpBpB: -20/9
7	$\frac{7}{5}\,\theta_4\,\mathrm{pApB}$	pApB: -5/7, nucpApB: 20/9
8	$\frac{7}{5}\,\theta_5\,\mathrm{pApA}$	pApA: -5/7, nucpApA: 20/9
9	$\frac{7}{5}\,\theta_5\,\mathrm{pBpB}$	pBpB: -5/7, nucpBpB: 20/9
10	1	tau: 1

2. Translation into Lean. Each rate term is a `KExpr`, for example `KExpr.npow (KExpr.var 0) 2` stands for $\mathrm{STAT5A}^2$; numbers are exact rationals, and the list `Rx` pairs each term with its stoichiometric column. The initial value is the rational vector `xq` (STAT5A and STAT5B positive, all other states 0), and θ_q holds the six parameters.

3. Checks. Four Boolean checks are evaluated by the Lean kernel: `net_ok` (every rate term is well defined for $x \geq 0$, and every term that consumes a species vanishes when that species is 0, so the field is quasi-positive), `growth_ok` (with weights $w_i = 1$, $\sum_i f_i \leq a + b \sum_i x_i$: the stoichiometric columns of six terms have a negative sum, and the remaining ones, the nuclear-import terms 7–9 and $\dot{\tau} = 1$, are at most linear in the states), `theta_ok` (all $\theta_j > 0$) and `x0_ok` ($x_q \geq 0$).

4. Final theorem. `def final := network_final pos Rx net_ok c growth_ok` θ_q `theta_ok xq x0_ok` applies the end-to-end theorem of step 28 to these data. Its statement is: for every $T \geq 0$ there is a solution x_0 on $[0, T]$ with $x_0(0) = x_q$, $x_0(t) \geq 0$ and $(x_0(t), \theta_q)$ in the domain; there is a family $x(\cdot; \theta)$ of solutions for θ near θ_q with $x(\cdot; \theta_q) = x_0$; and $x(t; \theta)$ is differentiable in θ at θ_q with derivative $S(t)$, $S(0) = 0$, solving the variational equation. These are the hypotheses of steps 14–22 for this model.

5. Reproduction. From the folder `lean_proofs/`, the command

```
lake env lean Models/Boehm_JProteomeRes2014.lean
```

compiles the file in a few seconds and prints

```
'Models.Boehm_JProteomeRes2014.final' depends on axioms:  [propext,
Classical.choice, Quot.sound]
```

and, from the folder `lean_translation/` of the reproducibility package, the command

```
python3 verify_translation.py Boehm_JProteomeRes2014
```

compares the file with the SBML model (largest relative difference 4×10^{-16}).

Benchmark problems without a verified model

Of the 34 benchmark problems analysed in the paper, 27 have a verified model; Froehlich, which is verified in Lean, is the further problem analysed only on its state trajectories. The remaining seven problems were not verified, for the following reasons. They are not counted in any statement that relies on G5. Here *verified* refers only to the formal verification in Lean of the hypotheses on the dynamics (G5): the numerical analysis of these seven problems (selection, admissibility, external validation and robustness) is complete and is reported together with the other problems in the main text and in this Supplementary Information; what is missing for them is only the machine-checked proof that their equations satisfy the hypotheses of the theorems.

- **Smith**: two reactions (R16f and R17f) have rates $k_{11}\,\mathrm{Akt}\,\max(\mathrm{PIP3} - \mathrm{basal}, 0)$ and $k_{12}\,\mathrm{PKC}\,\max(\mathrm{PIP3} - \mathrm{basal}, 0)$, written in the SBML model as a piecewise expression. The maximum has a kink at PIP3 = basal, so the vector field is not C^1 and the theorem on the differentiability of the solution with respect to the parameters (steps 15–17) does not apply. Finite-difference sensitivities remain valid as long as the trajectory does not cross the kink; a proof that it never does on the simulated horizon would allow the model to be verified, but it is not implemented.

- **Liu** and **Oliveira**: they contain events triggered by the state or by a parameter-dependent condition, not at fixed times; such events are outside model class 2 (step 18).
- **Alkan**: the model translates, but the syntactic checkers cannot establish well-definedness for 22 rate terms or quasi-positivity for 6 of them; this does not mean that the properties fail, only that the automatic proof does not cover them.
- **Perelson** and **Schwen**: the quasi-positivity check fails for two rate terms each, so non-negativity of the solutions is not established by the checker.
- **Beer**: smoothness and quasi-positivity are verified on both segments of the model, but the linear growth check is not satisfied, so existence of the nominal solution on the horizon is not proved; only a conditional theorem (assuming the solution exists) is available.

Data of each model

Armistead_CellDeathDis2024

- **File:** `Models/Armistead_CellDeathDis2024.lean`.
- **States ($n = 4$):** Cer, Sphingo, S1P, Sphinga.
- **Strictly positive species Σ (4):** Cer, Sphingo, S1P, Sphinga.
- **Parameters and nominal values θ_0 ($p = 10$):** k00 = 0.2381; k0 = 1000; k_d = 0.9068; k1 = 0.3011; k2 = 1000; k3 = 0.001; k4 = 2.362; k5 = 0.001001; alpha_hai1a = 0.5; alpha_cer = -0.3275.
- **Parameters declared positive:** 9 of 10.
- **Rate terms:** 8, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the strict region P_Σ, quasi-positive, and consumption of species in Σ at most proportional to their concentration (`checkNetS`); linear growth bound on P_Σ (`checkGrowthS`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta_0) \in P_\Sigma$ (`checkInitS`); declared parameter bounds hold at θ_0 (`checkUB`).
- **Fidelity of the translation:** largest relative difference 0.
- **Final theorem** (`Models.Armistead_CellDeathDis2024.final` := `strict_final_init`): for every $T \geq 0$ the nominal solution exists on $[0, T]$ and stays in P_Σ (species in Σ strictly positive) and in the domain; solutions near θ_0 exist and are differentiable, $S(0) = \partial_\theta G(\theta_0)$.

Bachmann_MSB2011

- **File:** `Models/Bachmann_MSB2011.lean`.
- **States ($n = 25$):** EpoRJAK2, EpoRpJAK2, p1EpoRpJAK2, p2EpoRpJAK2, p12EpoRpJAK2, EpoRJAK2_CIS, SHP1, SHP1Act, STAT5, pSTAT5, npSTAT5, CISnRNA1, CISnRNA2, CISnRNA3, CISnRNA4, CISnRNA5, CISRNA, CIS, SOCS3nRNA1, SOCS3nRNA2, SOCS3nRNA3, SOCS3nRNA4, SOCS3nRNA5, SOCS3RNA, SOCS3.
- **Parameters and nominal values θ_0 ($p = 27$):** CISEqc = 432.9; CISEqcOE = 0.5303; CISInh = 7.852×10^8; CISRNADelay = 0.1448; CISRNATurn = 1000; CISTurn = 0.008399; EpoRActJAK2 = 0.2673; EpoRCISInh = 1×10^6; EpoRCISRemove = 5.43; JAK2ActEpo = 6.332×10^5; JAK2EpoRDeaSHP1 = 142.7; SHP1ActEpoR = 0.001; SHP1Dea = 0.008162; SHP1ProOE = 2.826; SOCS3Eqc = 173.6; SOCS3EqcOE = 0.6792; SOCS3Inh = 10.41; SOCS3RNADelay = 1.065; SOCS3RNATurn = 0.008309; SOCS3Turn = 1×10^4; STAT5ActEpoR = 39; STAT5ActJAK2 = 0.07811; STAT5Exp = 0.07451; STAT5Imp = 0.02689; init_EpoRJAK2 = 3.976; init_SHP1 = 26.73; init_STAT5 = 79.75.
- **Parameters declared positive:** 27 of 27.
- **Rate terms:** 43, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.

- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta)$ well defined and non-negative at θ_0 (`checkInit`).
- **Fidelity of the translation:** largest relative difference 3×10^{-15}.
- **Final theorem** (`Models.Bachmann_MSB2011.final` := `network_final_init`): same as `network_final`, with initial value $x(0) = G(\theta)$ and $S(0) = \partial_\theta G(\theta_0)$.

Bertozzi_PNAS2020

- **File:** `Models/Bertozzi_PNAS2020.lean`.
- **States ($n = 3$):** I_, R_, S_.
- **Parameters and nominal values θ_0 ($p = 0$):** .
- **Parameters declared positive:** 0 of 0.
- **Rate terms:** 2, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (2 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 0.
- **Final theorem** (`Models.Bertozzi_PNAS2020.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Blasi_CellSystems2016

- **File:** `Models/Blasi_CellSystems2016.lean`.
- **States ($n = 16$):** x_0ac, x_k8, x_k5, x_k12, x_k16, x_k5k8, x_k5k12, x_k5k16, x_k8k12, x_k8k16, x_k12k16, x_k5k8k12, x_k5k8k16, x_k5k12k16, x_k8k12k16, x_4ac.
- **Parameters and nominal values θ_0 ($p = 8$):** a_basal = 0.0668; a_k8 = 0.0273; a_k5_k5k12 = 2.062; a_k12_k5k12 = 0.5519; a_k16_k12k16 = 0.6959; a_k5k12_k5k8k12 = 0.3253; a_k12k16_k8k12k16 = 2.205; a_k8k12k16_4ac = 3.592.
- **Parameters declared positive:** 8 of 8.
- **Rate terms:** 36, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (1 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 4×10^{-15}.
- **Final theorem** (`Models.Blasi_CellSystems2016.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Boehm_JProteomeRes2014

- **File:** `Models/Boehm_JProteomeRes2014.lean`.
- **States ($n = 9$):** STAT5A, STAT5B, pApB, pApA, pBpB, nucpApA, nucpApB, nucpBpB, τ (time).

- **Parameters and nominal values θ_0 ($p = 6$):** Epo_degradation_BaF3 = 0.02698; k_exp_hetero = 1.001×10^{-5}; k_exp_homo = 0.00617; k_imp_hetero = 0.01637; k_imp_homo = 9.775×10^{4}; k_phos = 1.577×10^{4}.
- **Parameters declared positive:** 6 of 6.
- **Rate terms:** 10, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (2 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 4×10^{-16}.
- **Final theorem** (`Models.Boehm_JProteomeRes2014.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t; \theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Borghans_BiophysChem1997

- **File:** `Models/Borghans_BiophysChem1997.lean`.
- **States ($n = 3$):** Z_state, Y_state, A_state.
- **Strictly positive species Σ (3):** Z_state, Y_state, A_state.
- **Parameters and nominal values θ_0 ($p = 20$):** K2 = 0.09995; K_par = 11.41; Ka = 0.1976; Kd = 0.3925; Kf = 1.138; Kp = 0.9966; Ky = 0.2004; Kz = 0.3034; Vd = 92.87; Vm2 = 7.457; Vm3 = 22.7; Vp = 2.757; beta_par = 1.124; epsilon_par = 0.1632; init_A_state = 1; init_Y_state = 0.9993; init_Z_state = 0.08792; n_par = 4.103; v0 = 2.318; v1 = 1.005.
- **Parameters declared positive:** 20 of 20.
- **Rate terms:** 9, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the strict region P_Σ, quasi-positive, and consumption of species in Σ at most proportional to their concentration (`checkNetS`); linear growth bound on P_Σ (`checkGrowthS`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta_0) \in P_\Sigma$ (`checkInitS`); declared parameter bounds hold at θ_0 (`checkUB`).
- **Fidelity of the translation:** largest relative difference 2×10^{-16}.
- **Final theorem** (`Models.Borghans_BiophysChem1997.final` := `strict_final_init`): for every $T \ge 0$ the nominal solution exists on $[0, T]$ and stays in P_Σ (species in Σ strictly positive) and in the domain; solutions near θ_0 exist and are differentiable, $S(0) = \partial_\theta G(\theta_0)$.

Brannmark_JBC2010

- **File:** `Models/Brannmark_JBC2010.lean`.
- **States ($n = 9$):** IR, IRins, IRp, IRiP, IRi, IRS, IRSiP, X, Xp.
- **Parameters and nominal values θ_0 ($p = 15$):** k1a = 0.1772; k1aBasic = 0.0003928; k1b = 0.1747; k1c = 0.05086; k1d = 5×10^{5}; k1e = 1×10^{-6}; k1f = 5×10^{5}; k1g = 1729; k1r = 0.0267; k21 = 2.102; k22 = 666.8; k3 = 4.944×10^{-5}; k_IRSiP_DosR = 37.94; km2 = 1.16; km3 = 0.4164.
- **Parameters declared positive:** 15 of 15.
- **Rate terms:** 14, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (9 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).

- **Fidelity of the translation:** largest relative difference 1×10^{-13}.
- **Final theorem** (`Models.Brannmark_JBC2010.final := network_final`): for every $T \geq 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Bruno_JExpBot2016

- **File:** `Models/Bruno_JExpBot2016.lean`.
- **States ($n = 7$):** bcar, bcry, b10, bio, ohb10, ohbio, zea.
- **Parameters and nominal values θ_0 ($p = 6$):** k5 = 0.003087; kb1 = 0.01663; kb2 = 0.005844; kc1 = 0.001679; kc2 = 0.006976; kc4 = 0.006083.
- **Parameters declared positive:** 6 of 6.
- **Rate terms:** 6, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (5 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \leq a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 2×10^{-16}.
- **Final theorem** (`Models.Bruno_JExpBot2016.final := network_final`): for every $T \geq 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Chen_MSB2009

- **File:** `Models/Chen_MSB2009.lean` and 5 part files in `Models/Chen_MSB2009/`.
- **States:** $n = 501$ species (listed in the file).
- **Parameters:** $p = 152$, with nominal values in θ_q (exact rationals) in the file.
- **Parameters declared positive:** 152 of 152.
- **Rate terms:** 1,243, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (27 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \leq a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 8×10^{-14}.
- **Final theorem** (`Models.Chen_MSB2009.final := network_final`): for every $T \geq 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Crauste_CellSystems2017

- **File:** `Models/Crauste_CellSystems2017.lean`.
- **States ($n = 5$):** Naive, EarlyEffector, LateEffector, Memory, Pathogen.
- **Parameters and nominal values θ_0 ($p = 12$):** delta_EL = 0.5179; delta_LM = 0.02258; delta_NE = 0.01193; mu_EE = 3.914×10^{-5}; mu_LE = 1×10^{-10}; mu_LL = 8.115×10^{-6}; mu_N = 0.7399; mu_P = 1×10^{-5}; mu_PE = 1.412×10^{-10}; mu_PL = 3.634×10^{-5}; rho_E = 0.5074; rho_P = 0.1264.
- **Parameters declared positive:** 12 of 12.
- **Rate terms:** 12, each with its stoichiometric column.

- **Initial values:** constant rational vector x_q (2 non-zero entries).
- **Weights w_i of the growth bound:** values 1, 1×10^5.
- **Horizon:** $T \le T_q = 1$ (quadratic growth; Riccati bound).
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`); quadratic growth bound $\sum_i w_i f_i \le Q(\varphi+1)^2$ and $1.1\,Q(\varphi(x_q)+1)T_q < 1$ (`checkRiccati`).
- **Fidelity of the translation:** largest relative difference 0.
- **Final theorem** (`Models.Crauste_CellSystems2017.final` := `riccati_final`): for every $T \le T_q$ the nominal solution exists on $[0, T]$, is non-negative and stays in the domain; solutions near θ_0 exist and are differentiable, $S(0) = 0$.

Elowitz_Nature2000

- **File:** `Models/Elowitz_Nature2000.lean`.
- **States ($n = 8$):** X_protein, Y_protein, Z_protein, GFP, X_mRNA, Y_mRNA, Z_mRNA, GFP_mRNA.
- **Strictly positive species Σ (8):** X_protein, Y_protein, Z_protein, GFP, X_mRNA, Y_mRNA, Z_mRNA, GFP_mRNA.
- **Parameters and nominal values θ_0 ($p = 18$):** KM = 1×10^{-5}; eff = 0.03016; eff_GFP = 0.001952; init_GFP = 3.387×10^{-5}; init_GFP_mRNA = 131.4; init_X_mRNA = 2.557; init_X_protein = 30.81; init_Y_mRNA = 1000; init_Y_protein = 0.0008183; init_Z_mRNA = 19.37; init_Z_protein = 994.4; n_Hill = 1.523; tau_mRNA = 12.62; tau_mRNA_GFP = 0.0004041; tau_prot = 5.359; tau_prot_GFP = 71.93; tps_active = 0.6123; tps_repr = 1×10^{-5}.
- **Parameters declared positive:** 18 of 18.
- **Rate terms:** 18, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the strict region P_Σ, quasi-positive, and consumption of species in Σ at most proportional to their concentration (`checkNetS`); linear growth bound on P_Σ (`checkGrowthS`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta_0) \in P_\Sigma$ (`checkInitS`); declared parameter bounds hold at θ_0 (`checkUB`).
- **Fidelity of the translation:** largest relative difference 2×10^{-16}.
- **Final theorem** (`Models.Elowitz_Nature2000.final` := `strict_final_init`): for every $T \ge 0$ the nominal solution exists on $[0, T]$ and stays in P_Σ (species in Σ strictly positive) and in the domain; solutions near θ_0 exist and are differentiable, $S(0) = \partial_\theta G(\theta_0)$.

Fiedler_BMCSystBiol2016

- **File:** `Models/Fiedler_BMCSystBiol2016.lean`.
- **States ($n = 7$):** RAF, pRAF, MEK, pMEK, ERK, pERK, τ (time).
- **Strictly positive species Σ (6):** RAF, pRAF, MEK, pMEK, ERK, pERK.
- **Parameters and nominal values θ_0 ($p = 12$):** K_1 = 1×10^{-5}; K_2 = 1.511; K_3 = 0.9605; k10 = 0.02641; k11 = 163.2; k2 = 1.226; k3 = 769.2; k4 = 2.907; k5 = 577.6; k6 = 3.608; tau1 = 0.2428; tau2 = 0.02114.
- **Parameters declared positive:** 12 of 12.
- **Rate terms:** 8, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.

- **Checks executed by the kernel:** rate laws well defined on the strict region P_Σ, quasi-positive, and consumption of species in Σ at most proportional to their concentration (`checkNetS`); linear growth bound on P_Σ (`checkGrowthS`); nominal parameters positive where declared (`checkPosParams`); initial values $G(\theta_0) \in P_\Sigma$ certified by rational interval arithmetic (`checkInitI`); declared parameter bounds hold at θ_0 (`checkUB`); the real powers in G are square roots.
- **Fidelity of the translation:** largest relative difference 0.
- **Final theorem** (`Models.Fiedler_BMCSystBiol2016.final := strict_final_initI`): same as `strict_final_init`, with the initial values certified by interval arithmetic.

Froehlich_CellSystems2018

- **File:** `Models/Froehlich_CellSystems2018.lean` and 13 part files in `Models/Froehlich_CellSystems2018/`.
- **States:** $n = 1{,}228$ species (listed in the file).
- **Parameters:** $p = 4088$, with nominal values in θ_q (exact rationals) in the file.
- **Parameters declared positive:** 4088 of 4088.
- **Rate terms:** 4,487, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (0 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 6×10^{-14}.
- **Final theorem** (`Models.Froehlich_CellSystems2018.final := network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Fujita_SciSignal2010

- **File:** `Models/Fujita_SciSignal2010.lean`.
- **States ($n = 9$):** EGFR, pEGFR, pEGFR_Akt, Akt, pAkt, S6, pAkt_S6, pS6, EGF_EGFR.
- **Parameters and nominal values θ_0 ($p = 16$):** EGFR_turnover = 0.00148; init_AKT = 0.02693; init_EGFR = 3.754×10^7; init_S6 = 188.1; reaction_1_k1 = 0.003691; reaction_1_k2 = 0.002301; reaction_2_k1 = 0.0009365; reaction_2_k2 = 6.097×10^4; reaction_3_k1 = 0.4332; reaction_4_k1 = 0.03016; reaction_5_k1 = 3.273×10^{-6}; reaction_5_k2 = 0.0003985; reaction_6_k1 = 5.463×10^{-6}; reaction_7_k1 = 0.0118; reaction_8_k1 = 0.0009448; reaction_9_k1 = 0.02851.
- **Parameters declared positive:** 16 of 16.
- **Rate terms:** 13, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta)$ well defined and non-negative at θ_0 (`checkInit`).
- **Fidelity of the translation:** largest relative difference 1×10^{-16}.
- **Final theorem** (`Models.Fujita_SciSignal2010.final := network_final_init`): same as `network_final`, with initial value $x(0) = G(\theta)$ and $S(0) = \partial_\theta G(\theta_0)$.

Giordano_Nature2020

- **File:** `Models/Giordano_Nature2020.lean`.
- **States ($n = 10$):** Susceptible, Infected, Diagnosed, Ailing, Recognized, Threatened, Healed, Extinct, CumulativeDiagnosed, DiagnosedHealed.
- **Parameters and nominal values θ_0 ($p = 43$):** alpha_0 = 0.57; beta_0 = 0.011; delta_0 = 0.011; epsilon_0 = 0.171; eta_0 = 0.125; gamma_0 = 0.456; kappa_0 = 0.017; lam_0 = 0.034; mu_0 = 0.017; nu_0 = 0.027; rho_0 = 0.034; sigma_0 = 0.017; tau = 0.01; theta = 0.371; xi_0 = 0.017; zeta_0 = 0.125; alpha_4 = 0.422; alpha_22 = 0.36; alpha_28 = 0.21; beta_4 = 0.0057; beta_22 = 0.005; delta_4 = 0.0057; delta_22 = 0.005; epsilon_12 = 0.143; epsilon_38 = 0.2; eta_22 = 0.034; eta_38 = 0.025; gamma_4 = 0.285; gamma_22 = 0.2; gamma_28 = 0.11; kappa_22 = 0.017; kappa_38 = 0.02; lam_22 = 0.08; mu_22 = 0.008; nu_22 = 0.015; rho_22 = 0.017; rho_38 = 0.02; sigma_22 = 0.017; sigma_38 = 0.01; xi_22 = 0.017; xi_38 = 0.02; zeta_22 = 0.034; zeta_38 = 0.025.
- **Parameters declared positive:** 43 of 43.
- **Rate terms:** 96, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (6 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Segments:** [0, 4], [4, 12], [12, 22], [22, 28], [28, 38], [38, 50].
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`); segment lengths are non-negative.
- **Fidelity of the translation:** largest relative difference 4×10^{-16}.
- **Final theorem** (`Models.Giordano_Nature2020.final` := `network_segments_final`): the same conclusions on each segment of the piecewise-defined model, with continuity across segment boundaries.

Isensee_JCB2018

- **File:** `Models/Isensee_JCB2018.lean`.
- **States ($n = 25$):** AC, pAC, AC_Fsk, pAC_Fsk, cAMP, Csub, Csub_H89, RII_C_2, RIIp_C_2, RIIp_cAMP_C_2, RIIp_cAMP_2, Sp8_Br_cAMPS, Rp8_Br_cAMPS, Rp8_pCPT_cAMPS, Rp_cAMPS, RIIp_Rp8_Br_cAMPS_C_2, RIIp_Rp8_pCPT_cAMPS_C_2, RIIp_Rp_cAMPS_C_2, RIIp_Sp8_Br_cAMPS_C_2, RIIp_Sp8_Br_cAMPS_2, RIIp_2, RII_2, IBMX, PDE, pPDE.
- **Parameters and nominal values θ_0 ($p = 32$):** KD_Fsk = 14.28; KD_H89 = 0.04807; KD_IBMX = 11.96; KD_cAMP = 2.85; kdeg_cAMP = 1×10^{-5}; kdeg_cAMP_free = 7.384; kf_Fsk = 1000; kf_H89 = 0.0007427; kf_RII_2__RII_C_2 = 1.264; kf_RII_C_2__RII_2 = 0.01875; kf_RII_C_2__RIIp_C_2 = 0.02277; kf_RIIp_2__RII_2 = 0.0323; kf_RIIp_C_2__RII_C_2 = 0.0257; kf_RIIp_cAMP_C_2__RIIp_2 = 0.1053; kf_cAMP = 0.5015; ki_IBMX = 1000; ki_Rp8_Br_cAMPS_pAB = 6.446; ki_Rp8_pCPT_cAMPS_pAB = 9.39; ki_Rp_cAMPS_pAB = 0.01632; ki_Sp8_Br_cAMPS_AM = 0.1317; ks_AC_cAMP = 0.4232; xi_AC_cAMP_Fsk = 642.4; xi_b_Rp8_Br_cAMPS = 0.04056; xi_b_Rp8_pCPT_cAMPS = 0.02668; xi_b_Rp_cAMPS = 0.6039; xi_b_Sp8_Br_cAMPS = 21.62; xi_kf_RII_2__RII_C_2 = 0.01893; xi_kf_RII_C_2__RII_2 = 0.811; xi_KD_Rp8_Br_cAMPS = 0.03996; xi_KD_Rp8_pCPT_cAMPS = 0.1942; xi_KD_Rp_cAMPS = 0.163; xi_KD_Sp8_Br_cAMPS = 0.2232.
- **Parameters declared positive:** 32 of 32.
- **Rate terms:** 34, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (10 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 1×10^{-14}.
- **Final theorem** (`Models.Isensee_JCB2018.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Lang_PLOSComputBiol2024

- **File:** `Models/Lang_PLOSComputBiol2024.lean`.
- **States:** $n = 124$ species (listed in the file).
- **Parameters:** $p = 164$, with nominal values in θ_q (exact rationals) in the file.
- **Parameters declared positive:** 164 of 164.
- **Rate terms:** 663, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (107 non-zero entries).
- **Weights w_i of the growth bound:** values 1, 2, 3.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 2×10^{-15}.
- **Final theorem** (`Models.Lang_PLOSComputBiol2024.final := network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Laske_PLOSComputBiol2019

- **File:** `Models/Laske_PLOSComputBiol2019.lean`.
- **States ($n = 34$):** V_ex, V_end, Vp_cyt, Vp_nuc, R_C, R_C_RdRp, R_V, R_V_RdRp, R_M1, R_M2, R_M3, R_M4, R_M5, R_M6, R_M7, R_M8, P_B1, P_B2, P_PA, P_HA, P_NP, P_NA, P_M1, P_NEP, Vp_nuc_M1, Vp_cyt_M1, V_rel, P_RdRp, Cp, V_att_Hi, V_att_Lo, P_M2, B_att_Lo, B_att_Hi.
- **Parameters and nominal values θ_0 ($p = 6$):** k_imp = 0.296; k_syn_R_M = 3.06×10^4; k_syn_R_C = 1.53; k_syn_R_V = 100.9; k_bind_M1 = 1.82×10^{-6}; k_rel = 0.0011.
- **Parameters declared positive:** 6 of 6.
- **Rate terms:** 51, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (3 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 0.
- **Final theorem** (`Models.Laske_PLOSComputBiol2019.final := network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Lucarelli_CellSystems2018

- **File:** `Models/Lucarelli_CellSystems2018.lean`.
- **States ($n = 33$):** TGFb, Rec, TGFb_pRec, S2, S3, S4, S2_S4_S4, ppS2_ppS2_ppS2, ppS3_ppS3_ppS3, S4_S4_S4, pS2, pS3, ppS2, ppS3, ppS2_ppS2_S4, ppS2_ppS2_ppS3, ppS2_ppS3_ppS3, ppS3_ppS3_S4, ppS2_ppS3_S4, ppS3_S4_S4, ppS2_S4_S4, geneA, geneB, geneC, geneD, geneE, geneF, geneG, geneH, geneI, geneJ, geneK, geneL.
- **Parameters:** $p = 69$, with nominal values in θ_q (exact rationals) in the file.
- **Parameters declared positive:** 69 of 69.
- **Rate terms:** 67, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.

- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta)$ well defined and non-negative at θ_0 (`checkInit`).
- **Fidelity of the translation:** largest relative difference 4×10^{-15}.
- **Final theorem** (`Models.Lucarelli_CellSystems2018.final := network_final_init`): same as `network_final`, with initial value $x(0) = G(\theta)$ and $S(0) = \partial_\theta G(\theta_0)$.

Okuonghae_ChaosSolitonsFractals2020

- **File:** `Models/Okuonghae_ChaosSolitonsFractals2020.lean`.
- **States ($n = 8$):** susceptible, exposed, asymptomatic, symptomatic, detected, recovered, deceased, detected_cumulative.
- **Strictly positive species Σ (6):** susceptible, exposed, asymptomatic, symptomatic, detected, detected_cumulative.
- **Parameters and nominal values θ_0 ($p = 14$):** alpha = 0.5; d_0 = 0.015; d_D = 0.015; gamma_0 = 0.1398; gamma_a = 0.1398; gamma_i = 0.06667; nu = 0.5; psi = 0.0135; sigma = 0.1923; theta = 2×10^{-12}; asymptomatic_start = 188; symptomatic_start = 212; exposed_start = 441; transmission_rate_effective = 0.4236.
- **Parameters declared positive:** 14 of 14.
- **Rate terms:** 11, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the strict region P_Σ, quasi-positive, and consumption of species in Σ at most proportional to their concentration (`checkNetS`); linear growth bound on P_Σ (`checkGrowthS`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta_0) \in P_\Sigma$ (`checkInitS`); declared parameter bounds hold at θ_0 (`checkUB`).
- **Fidelity of the translation:** largest relative difference 7×10^{-16}.
- **Final theorem** (`Models.Okuonghae_ChaosSolitonsFractals2020.final := strict_final_init`): for every $T \ge 0$ the nominal solution exists on $[0, T]$ and stays in P_Σ (species in Σ strictly positive) and in the domain; solutions near θ_0 exist and are differentiable, $S(0) = \partial_\theta G(\theta_0)$.

Rahman_MBS2016

- **File:** `Models/Rahman_MBS2016.lean`.
- **States ($n = 7$):** susceptible, infected_normal, infected_moderate, infected_weak, treated_normal, treated_moderate, treated_weak.
- **Strictly positive species Σ (4):** susceptible, infected_normal, infected_moderate, infected_weak.
- **Parameters and nominal values θ_0 ($p = 9$):** infected_normal_transmission_rate_relative = 427.5; infected_moderate_transmission_rate = 0.002927; infected_weak_transmission_rate_relative = 993; infected_weak_treatment_rate = 0.0001; infected_normal_worsen_rate = 0.6654; infected_moderate_worsen_rate = 0.007976; treated_moderate_improve_rate = 0.1364; treated_weak_improve_rate = 0.7756; behavioural_change_rate = 1.338×10^{-7}.
- **Parameters declared positive:** 9 of 9.
- **Rate terms:** 19, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the strict region P_Σ, quasi-positive, and consumption of species in Σ at most proportional to their concentration (`checkNetS`); linear growth bound on P_Σ (`checkGrowthS`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta_0) \in P_\Sigma$ (`checkInitS`); declared parameter bounds hold at θ_0 (`checkUB`).

- **Fidelity of the translation:** largest relative difference 1×10^{-15}.
- **Final theorem** (`Models.Rahman_MBS2016.final` := `strict_final_init`): for every $T \geq 0$ the nominal solution exists on $[0, T]$ and stays in P_Σ (species in Σ strictly positive) and in the domain; solutions near θ_0 exist and are differentiable, $S(0) = \partial_\theta G(\theta_0)$.

Raia_CancerResearch2011

- **File:** `Models/Raia_CancerResearch2011.lean`.
- **States ($n = 14$):** Rec, Rec_i, IL13_Rec, p_IL13_Rec, p_IL13_Rec_i, JAK2, pJAK2, STAT5, pSTAT5, SOCS3mRNA, DecoyR, IL13_DecoyR, SOCS3, CD274mRNA.
- **Parameters and nominal values θ_0 ($p = 18$):** CD274mRNA_production = 0.02121; DecoyR_binding = 0.003851; JAK2_p_inhibition = 0.07814; JAK2_phosphorylation = 1.017; Kon_IL13Rec = 0.002278; Rec_intern = 0.3441; Rec_phosphorylation = 1000; Rec_recycle = 0.002098; SOCS3_accumulation = 451.8; SOCS3_degradation = 0.04338; SOCS3_translation = 15.72; SOCS3mRNA_production = 0.1434; STAT5_phosphorylation = 0.01948; init_Rec_i = 235.1; pJAK2_dephosphorylation = 0.0001759; pRec_degradation = 0.2096; pRec_intern = 0.5853; pSTAT5_dephosphorylation = 0.0002748.
- **Parameters declared positive:** 18 of 18.
- **Rate terms:** 14, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \leq a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta)$ well defined and non-negative at θ_0 (`checkInit`).
- **Fidelity of the translation:** largest relative difference 6×10^{-16}.
- **Final theorem** (`Models.Raia_CancerResearch2011.final` := `network_final_init`): same as `network_final`, with initial value $x(0) = G(\theta)$ and $S(0) = \partial_\theta G(\theta_0)$.

Raimundez_PCB2020

- **File:** `Models/Raimundez_PCB2020.lean`.
- **States ($n = 22$):** EGFR, EGFR_CET, EGFR_EGF, EGFR_EGF_2, pEGFR_EGF_2, pEGFR_EGF_2_i, MMET, MMET_2, pMMET_2, pMMET_2_i, MMET_EGFR, pMMET_pEGFR, pMMET_pEGFR_i, MMET_METinh, MMET_MMET_METinh, MMET_METinh_2, EGFR_MMET_METinh, RAS_GTP, pMAPK, pPI3K, pMPI3K, pAKT.
- **Parameters and nominal values θ_0 ($p = 57$):** KD_EGFR_CET = 0.218; KD_EGFR_EGF = 0.3208; KD_METinh = 0.04194; d_AKTtotal__MKN1_2_HS746T = 0; d_AKTtotal__fm_2_hm = 0.8229; d_MAPKtotal__MKN1_2_HS746T = 0; d_MAPKtotal__fm_2_hm = 0.4275; d_MPI3Ktotal__fm_2_hm = -1.545; d_PI3Ktotal__fm_2_hm = -0.5318; d_RAStotal__MKN1_2_HS746T = -1.06×10^{-22}; d_RAStotal__fm_2_hm = -0.5382; d_kdeg_membran__MKN1_2_HS746T = -0.6; d_kdeg_pEGFR_EGF_2_i__MKN1_2_HS746T = 3; d_kexp_pEGFR_EGF_2_i__MKN1_2_HS746T = -0.2382; d_kimp_pEGFR_EGF_2__MKN1_2_HS746T = -1.746; d_ksyn_EGFR__MKN1_2_HS746T = -3; d_ksyn_EGFR__fm_2_hm = -0.0876; d_ksyn_MMET__fm_2_hm = -0.5189; ka_AKT__MKN1 = 443.1; ka_MAPK__MKN1 = 1000; ka_PI3K__basal = 0.0003649; ka_PI3K__pEGFR_EGF_2 = 0.001701; ka_RAS__basal__MKN1 = 0.0002145; ka_RAS__pEGFR_EGF_2__MKN1 = 1.47×10^{-5}; kbin_EGFR_CET = 0.5853; kbin_EGFR_EGF = 8.882; kdeg_membran__MKN1 = 0.002259; kdeg_pEGFR_EGF_2_i__MKN1 = 0.0668; kdim_EGFR_EGF = 0.03239; kdim_MMET = 2.61×10^{-5}; kdim_MMET_EGFR = 0.001882; kdim_MMETinh = 0.03127; kexp_pEGFR_EGF_2_i__MKN1 = 0.7517; ki_AKT__MKN1 = 18.47; ki_MAPK = 2.177; ki_PI3K__MKN1 = 1.316; ki_RAS__MKN1 = 0.327; kimp_pEGFR_EGF_2__MKN1 = 0.7469; kpho_EGFR_EGF = 1.344; kpho_MMET = 2.543; kpho_MMET_EGFR = 4.563; ksyn_MMET__HS746T_fm = 0.02635; xi_ka_PI3K_pMMET_2 = 301.3; xi_ka_PI3K_pMMET_pEGFR = 0.7655; xi_ka_RAS_pMMET_2 = 0.4494; xi_ka_RAS_pMMET_pEGFR = 0.2536; xi_kdeg_pMMET_2_i = 0.7837; xi_kdeg_pMMET_pEGFR_i = 3.533; xi_kdim_MMET = 0.9748; xi_kdim_MMET_EGFR = 1; xi_kexp_pMMET_2_i = 1.243; xi_kexp_pMMET_pEGFR_i = 0.2864; xi_ki_MPI3K = 0.00322;

xi_kimp_pMMET_2 = 0.07715; xi_kimp_pMMET_pEGFR = 0.00796; xi_kpho_MMET = 0.008328; xi_kpho_MMET_EGFR = 0.06967.

- **Parameters declared positive:** 45 of 57.
- **Rate terms:** 91, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (0 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 1×10^{-12}.
- **Final theorem** (`Models.Raimundez_PCB2020.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

SalazarCavazos_MBoC2020

- **File:** `Models/SalazarCavazos_MBoC2020.lean`.
- **States:** $n = 75$ species (listed in the file).
- **Parameters and nominal values θ_0 ($p = 6$):** GRB2_total__FREE = 1.699×10^5; SHC1_total__FREE = 6.494×10^5; kdephosY1068__FREE = 1.659; kdephosYN__FREE = 0.01718; ratio_kpkd_Y1068__FREE = 0.1575; ratio_kpkd_YN__FREE = 0.4448.
- **Parameters declared positive:** 6 of 6.
- **Rate terms:** 457, each with its stoichiometric column.
- **Initial values:** expressions $G(\theta)$ in the parameters (SBML initial assignments).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); parameter-dependent initial values $G(\theta)$ well defined and non-negative at θ_0 (`checkInit`).
- **Fidelity of the translation:** largest relative difference 2×10^{-13}.
- **Final theorem** (`Models.SalazarCavazos_MBoC2020.final` := `network_final_init`): same as `network_final`, with initial value $x(0) = G(\theta)$ and $S(0) = \partial_\theta G(\theta_0)$.

Sneyd_PNAS2002

- **File:** `Models/Sneyd_PNAS2002.lean`.
- **States ($n = 6$):** IPR_O, IPR_R, IPR_I1, IPR_S, IPR_A, IPR_I2.
- **Parameters and nominal values θ_0 ($p = 14$):** k1 = 3.727; k2 = 1×10^5; k3 = 15.75; k4 = 9.994×10^4; k_1 = 0.9239; k_2 = 0.001002; k_3 = 1.915; k_4 = 3079; l2 = 0.9401; l4 = 2.858; l6 = 1×10^5; l_2 = 0.3477; l_4 = 0.01388; l_6 = 0.001.
- **Parameters declared positive:** 14 of 14.
- **Rate terms:** 16, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (1 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 4×10^{-15}.
- **Final theorem** (`Models.Sneyd_PNAS2002.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Weber_BMC2015

- **File:** `Models/Weber_BMC2015.lean`.
- **States ($n = 7$):** PKD, PKDDAGa, PI4K3B, PI4K3Ba, CERTERa, CERT, CERTTGNa.
- **Parameters and nominal values θ_0 ($p = 26$):** a11 = 0.1825; a12 = 26.67; a21 = 1.877; a22 = 0.0001; a31 = 0.1353; a32 = 0.0001; a33 = 0.0001; m11 = 1×10^{10}; m22 = 2154; m31 = 9.97×10^9; m33 = 4.187×10^7; p11 = 4.244; p12 = 0.005936; p13 = 0.00246; p21 = 5.629; p22 = 21.79; p31 = 2538; p32 = 16.77; p33 = 2.166×10^4; pu3 = 1×10^8; pu4 = 3.313×10^7; pu5 = 33.7; pu6 = 113.4; s12 = 8.846×10^4; s21 = 2.961×10^6; s31 = 4.328×10^6.
- **Parameters declared positive:** 26 of 26.
- **Rate terms:** 18, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (7 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 0.
- **Final theorem** (`Models.Weber_BMC2015.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Zhao_QuantBiol2020

- **File:** `Models/Zhao_QuantBiol2020.lean`.
- **States ($n = 4$):** Susceptible, Unquarantined_Infected, Quarantined_Infected, Confirmed_Infected.
- **Parameters and nominal values θ_0 ($p = 21$):** R_Stage_I_Wuhan = 4.709; gamma_1_Stage_I_Wuhan = 0.063; gamma_2_Stage_I_Wuhan = 0.05; R_Stage_II_Wuhan = 0.7575; gamma_1_Stage_II_Wuhan = 0.3917; gamma_2_Stage_II_Wuhan = 0.0643; R_Stage_III_Wuhan = 0.4797; gamma_1_Stage_III_Wuhan = 0.6185; gamma_2_Stage_III_Wuhan = 0.322; R_Stage_I_Hubei = 5.934; gamma_1_Stage_I_Hubei = 0.05; gamma_2_Stage_I_Hubei = 0.05; R_Stage_II_Hubei = 0.6079; gamma_1_Stage_II_Hubei = 0.488; gamma_2_Stage_II_Hubei = 0.1914; R_Stage_I_China = 1.528; gamma_1_Stage_I_China = 0.1941; gamma_2_Stage_I_China = 0.05; R_Stage_II_China = 0.5753; gamma_1_Stage_II_China = 0.5157; gamma_2_Stage_II_China = 0.2189.
- **Parameters declared positive:** 21 of 21.
- **Rate terms:** 3, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (3 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 0.
- **Final theorem** (`Models.Zhao_QuantBiol2020.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Zheng_PNAS2012

- **File:** `Models/Zheng_PNAS2012.lean`.
- **States ($n = 15$):** K27me0K36me0, K27me0K36me1, K27me1K36me0, K27me0K36me2, K27me1K36me1, K27me2K36me0, K27me0K36me3, K27me1K36me2, K27me2K36me1, K27me3K36me0, K27me1K36me3, K27me2K36me2, K27me3K36me1, K27me2K36me3, K27me3K36me2.

- **Parameters and nominal values θ_0 ($p = 45$):** inflowp = 0.03048; k00_01 = 997.3; k00_10 = 0.234; k01_00 = 479.8; k01_02 = 1.913; k01_11 = 2.651; k02_01 = 0.06996; k02_03 = 1×10^{-5}; k02_12 = 1×10^{-5}; k03_02 = 1×10^{-5}; k03_13 = 988.1; k10_00 = 1×10^{-5}; k10_11 = 1×10^{-5}; k10_20 = 0.2692; k11_01 = 1×10^{-5}; k11_10 = 1×10^{-5}; k11_12 = 11.85; k11_21 = 1×10^{-5}; k12_02 = 1×10^{-5}; k12_11 = 0.107; k12_13 = 0.03563; k12_22 = 0.01062; k13_03 = 450.4; k13_12 = 0.1953; k13_23 = 1×10^{-5}; k20_10 = 1×10^{-5}; k20_21 = 1×10^{-5}; k20_30 = 0.1507; k21_11 = 1×10^{-5}; k21_20 = 0.0354; k21_22 = 999.3; k21_31 = 1×10^{-5}; k22_12 = 0.05338; k22_21 = 426.1; k22_23 = 0.1517; k22_32 = 1×10^{-5}; k23_13 = 0.7912; k23_22 = 1×10^{-5}; k30_20 = 1×10^{-5}; k30_31 = 0.2813; k31_21 = 1.001×10^{-5}; k31_30 = 1×10^{-5}; k31_32 = 0.3536; k32_22 = 0.633; k32_31 = 1×10^{-5}.
- **Parameters declared positive:** 45 of 45.
- **Rate terms:** 44, each with its stoichiometric column.
- **Initial values:** constant rational vector x_q (15 non-zero entries).
- **Weights w_i of the growth bound:** all equal to 1.
- **Checks executed by the kernel:** rate laws well defined on the non-negative orthant and quasi-positive (`checkNet`); linear growth bound $\sum_i w_i f_i \le a + b \sum_i w_i x_i$ with the weights w (`checkGrowth`); nominal parameters positive where declared (`checkPosParams`); initial values non-negative (`checkNonneg`).
- **Fidelity of the translation:** largest relative difference 1×10^{-15}.
- **Final theorem** (`Models.Zheng_PNAS2012.final` := `network_final`): for every $T \ge 0$ the nominal solution from x_q exists on $[0, T]$, is non-negative and stays in the domain of the rate laws; solutions exist for all θ near θ_0, and $x(t;\theta)$ is differentiable in θ with sensitivity $S(t)$, $S(0) = 0$.

Supplementary Note: Reproducibility package

All programs, results and formal proofs of this work are collected in a reproducibility package (folder `reproducibility/` of the repository `https://github.com/monicammr/archivos`, archived on Zenodo, DOI `10.5281/zenodo.23338174`). Its file `README.md` is a step-by-step manual; this note summarises what can be checked and how.

Contents of the package

- `analysis/`: the method and every numerical analysis of the paper in Python (simulation of the measured outputs of a PEtab problem, selection, external and nested validation, synthetic experiment, robustness, single-parameter subsets, settings, cost, simulator check, comparison with other criteria, identifiability), including the additional analyses described in the Supplementary Notes on robustness (independent directions), on the synthetic experiment (true signal and other amplitudes), on the method settings (error in units of the noise) and on random scenarios and large problems; all settings and seeds are in `analysis/config.py`.
- `results/`: the results of these analyses, one file per benchmark problem, with a data dictionary (`DATA_DICTIONARY.md`).
- `figures_tables/`: the programs that produce every figure, table and Supplementary Note that contains numbers from those results.
- `lean_proofs/`: the Lean 4 development (32 modules, 309 theorems and lemmas, 28 benchmark models, and the non-vacuity witnesses `NonVacuity.lean`).
- `lean_translation/`: the translator of the SBML models into Lean and the independent check of the translation.

The PEtab problems are downloaded from the benchmark collection at commit `fcbddf1` by `scripts/get_benchmark.sh`.

Three levels of checking

The package can be checked at three levels, which answer different questions and are independent of each other: a reader can carry out any one of them without the others.

Level A: do the figures, tables and numbers of the paper follow from the stored results? The command `python figures_tables/build_all.py` reads the result files of `results/` (one per problem and analysis) and regenerates every figure, Table 1, Table 2 and every Supplementary Table and Note that contains numbers; each program also prints the counts quoted in the text (for example the number of admissible problems), so that they can be compared with the paper. No simulation is run. The README contains a table that relates every figure, table and number of the paper to the program that produces it and to the result files it reads, and `DATA_DICTIONARY.md` explains every field of those files. This level shows that the numbers, tables and figures of the paper are produced by these programs from the stored results; it does not show that the results themselves are correct, which is the purpose of level C. It takes a few minutes on a laptop and needs Python and a LaTeX distribution.

Level B: are the mathematical statements proved, and do they assume nothing unproved? The command `./verify.sh` in `lean_proofs/` compiles the whole Lean development (32 theory modules with 309 theorems and lemmas, the non-vacuity witnesses and the 28 benchmark models) from scratch and stops with an error unless (i) no file contains `sorry` or an `axiom` declaration, (ii) every proof is accepted by the Lean kernel and (iii) every theorem and every model theorem depends only on Lean's standard axioms. The independent checker `lean4checker` can be run on the compiled modules as a second, separate verification. This level shows that each theorem follows from its hypotheses; whether a statement says what the paper claims must be judged by reading it, for which Supplementary Note *Statements of the formal theorems* gives every statement in mathematical notation, and whether its hypotheses hold for a numerical result is stated in Supplementary Table S5. It takes about 20 minutes on a many-core machine and one to two hours on a laptop (most of it for the largest models, Chen and Froehlich), and needs only the Lean version manager `elan`, which installs Lean 4.23.0 and the precompiled Mathlib.

Level C: do the analyses, re-run from the original benchmark problems, give the stored results? After downloading the PEtab problems (`scripts/get_benchmark.sh`, commit `fcbddf1`), the programs of `analysis/` are run in the order given in the README, writing to a separate folder (variable `REPRO_RESULTS`), and `scripts/compare_results.py` compares the new results with the stored ones; the additional analyses (fixed scenario set, selection on the training outputs, standard pivoted QR, other families of synthetic truths, nonlinear check of the first-order problems and identifiability over the admissible set) are run by `scripts/run_second_review.sh` and `scripts/run_identifiability_second_review.sh`. With the fixed seeds the results are deterministic; small differences can arise only where a simulation fails to integrate in one run and not in another (some scenarios of Chen and Isensee) and from different versions of the numerical libraries; for example, the selection of Lang and Raimundez gave the same subsets with libRoadRunner 2.9.2 and 2.10.0, with e_{fit} agreeing to within 10^{-8} (Supplementary Note *Random scenarios and large problems*). This level shows that the stored results are those that the code produces from the public data. Its cost depends on how much is repeated: the method itself (selection of all 34 problems) takes minutes per problem, and about 1.3 h of processor time in total, whereas all the analyses of the paper together took about 697 h of processor time (Supplementary Note *Computing time of the analyses*), that is, about a week on four cores or about a day on a 32-core server. A reader can therefore repeat the selection of every problem, or all analyses of a few problems, on a laptop, while repeating everything for the largest problems (Chen, Isensee, Raimundez, Smith, Bachmann) needs a many-core server. It needs Python with the packages of `requirements.txt` and the benchmark (about 80 MB).

What each check of the formal verification establishes

- *Compilation* establishes that every proof is accepted by the Lean kernel, i.e. that each theorem follows from its hypotheses.

- *Axiom inspection* (`#print axioms`) establishes that no proof relies on an unproved assumption: the only axioms are `propext`, `Classical.choice` and `Quot.sound`, which are part of Lean's standard logic.
- *The fidelity check* of Supplementary Note *Verified benchmark models* establishes that the data on which the model checks run (rate laws, stoichiometry, parameter values and initial values) are those of the benchmark models.

The statements themselves should be compared with the mathematics they are meant to encode; Supplementary Note *Statements of the formal theorems* gives each statement in mathematical notation for this purpose.

Versions

Lean 4.23.0 (file `lean_proofs/lean-toolchain`) and Mathlib at tag `v4.23.0` (file `lean_proofs/lakefile.lean`). The Python programs use Python 3.11 with libRoadRunner 2.10, libSBML 5.21, NumPy, SciPy, pandas, SymPy and Matplotlib at the versions listed in `requirements.txt`.

Steps for the formal verification

1. Install the Lean version manager `elan` (`https://github.com/leanprover/elan`); it reads the Lean version from `lean-toolchain`.
2. In the folder `lean_proofs/`, run `./verify.sh`. The script (i) checks that no file contains `sorry` or an `axiom` declaration, (ii) downloads the precompiled Mathlib (`lake exe cache get`), (iii) compiles the 32 theory modules, the witnesses `NonVacuity.lean` and the 28 models (`lake build`) and (iv) prints the axioms of every theorem (`CheckAxioms.lean`) and of the final theorem of every model (`CheckModelAxioms.lean`). The theory compiles in a few minutes; the models take longer, the largest (Chen and Froehlich) about 30 minutes each. A large stack is needed (`ulimit -s unlimited`; the build passes `--tstack=4000000` to the model files).
3. To inspect one model only: `lake env lean Models/<model>.lean`. The last line of each model file prints the axioms of its final theorem.
4. To locate a theorem: Supplementary Table S4 gives the module of every main result, and `lean_proofs/FORMULAS.md` lists every declaration with its module and statement; the module `X` is the file `lean_proofs/X.lean`.
5. To check the translation of the models (needs the benchmark): in the folder `lean_translation/`, `python sbml_to_lean.py` regenerates the model files from the SBML models (the files are identical to those of `lean_proofs/Models/`), `python verify_translation.py` compares every model file with its SBML model and writes `results/sbml_to_lean/translation_fidelity.json`, and `python mutation_control.py` shows that a 1% change of one number of a model file is detected.

The complete build log, with the compilation time of every module, is `lean_proofs/BUILD_LOG.md`.

First-order re-estimation

For the eight problems treated to first order, the first-order re-estimation uses the columns of the sensitivity matrix of the selected and of the calibration parameters, $J_{S\cup\Theta_C}$; the program computes the columns of both kinds of parameters. Run from scratch, without any cached result, it reproduces exactly the subsets and e_{fit} reported for all eight problems (for example Bachmann: CISRNADelay and JAK2ActEpo, $e_{\mathrm{fit}} = 0.337$; Chen: k106 and k44, 0.082).

Supplementary Note: Verification of the simulator

The simulation of the measured outputs (`analysis/petab_problem.py`, libRoadRunner with CVODE, relative and absolute tolerances 10^{-7} and 10^{-10}) was compared, at the nominal parameters θ_0, with the reference simulations distributed with the benchmark (file `simulations.tsv` of each problem), matching every measurement by observable, experimental condition and time. For measurement i the relative difference is $|\hat{y}_i - y_i^{\mathrm{ref}}| / \max(|y_i^{\mathrm{ref}}|, 10^{-12})$; Supplementary Table S10 gives its median and maximum for every problem.

In all problems with a comparable reference the median difference is below 10^{-5} except Chen (1.1×10^{-4}), Crauste (0.8%, a very stiff model), Lang, Perelson and Zheng. In Lang, an oscillating cell-cycle model simulated to $t = 136{,}472$, the median difference is 2.8% and the peaks differ by an accumulated phase shift. In Perelson the difference is 17%; the initial state of the SBML model is stationary and the simulation respects it, which suggests, but does not prove, that the reference was generated with other values. In Zheng the reference file contains the measured data rather than a simulation. Problems without a comparable reference have no file in the benchmark or use different condition identifiers; for them (Fujita, Isensee, Okuonghae, Oliveira, Raia, Blasi and Zhao) this check cannot exclude an implementation error.

Because all criteria are evaluated on noise-scaled outputs, the relevant size of a discrepancy is its value in units of the recorded noise, $|h(\hat{y}_i) - h(y_i^{\mathrm{ref}})| / \sigma_i$. In these units the discrepancy is negligible in Perelson (at most 2×10^{-6}, because its recorded noise is large compared with the outputs), in Crauste (at most 0.17) and in Sneyd (median 3×10^{-7}; only the duplicated rows exceed one), but not in Lang: the median is 0.29, 22% of the measurements differ by more than one noise unit and the largest difference is 18. The results of Lang should therefore be read with this caveat. In Zheng the same quantity compares the simulation with the measured data (median 0.47), so it is a residual rather than a simulation error.

Supplementary Table S10. Agreement of the simulation of the measured outputs with the reference simulations of the benchmark at θ_0: number of measurements and conditions, estimated parameters p and those that affect the outputs, and median and maximum relative difference.

Problem	Meas.	Cond.	p	p_y	Median	Max	Note
Alkan	1733	73	44	44	3.1×10^{-8}	3.1×10^{-5}	agrees
Armistead	58	2	14	10	0	6.1×10^{-7}	agrees
Bachmann	541	36	113	101	7.1×10^{-8}	2.2×10^{-5}	agrees
Beer	27132	19	72	70	5.9×10^{-8}	3.8×10^{-4}	agrees
Bertozzi	22	2	8	6	1.6×10^{-7}	1.3×10^{-6}	agrees
Blasi	252	1	9	8	–	–	reference file has 288 rows for 252 measurements
Boehm	48	1	9	6	1.1×10^{-7}	8.6×10^{-7}	agrees
Borghans	111	1	23	22	2.2×10^{-6}	1.6×10^{-4}	agrees
Brannmark	43	8	22	18	2.3×10^{-7}	1.3×10^{-5}	agrees
Bruno	77	6	13	13	1.3×10^{-7}	2.3×10^{-6}	agrees
Chen	120	4	155	155	1.1×10^{-4}	2.8×10^{-3}	agrees
Crauste	21	1	12	12	7.8×10^{-3}	7.9×10^{-2}	median difference below 1% (very stiff model)
Elowitz	58	1	21	20	1.6×10^{-6}	1.7×10^{-5}	agrees
Fiedler	72	3	22	20	1.1×10^{-7}	9.9×10^{-7}	agrees
Fujita	144	6	19	19	–	–	reference uses other condition identifiers
Giordano	313	1	50	43	8.9×10^{-7}	2.0×10^{-6}	agrees
Isensee	687	123	46	43	–	–	reference uses other condition identifiers
Lang	9600	1	294	277	2.8×10^{-2}	2.3×10^{0}	oscillator simulated to $t = 136{,}472$: medians agree, peaks differ by an accumulated phase shift

Problem	Meas.	Cond.	p	p_y	Median	Max	Note
Laske	42	3	13	8	1.3×10^{-7}	4.5×10^{-7}	agrees
Liu	100	2	9	7	2.9×10^{-8}	6.1×10^{-7}	agrees
Lucarelli	1755	16	84	72	0	1.5×10^{-6}	agrees
Okuonghae	92	1	16	14	–	–	no reference file in the benchmark
Oliveira	120	1	12	12	–	–	no reference file in the benchmark
Perelson	16	1	3	2	1.7×10^{-1}	4.5×10^{-1}	discrepancy; the initial state is stationary ($\dot{V}(0) = 0$) and the simulation respects it, so the reference was probably generated with other values
Rahman	23	1	9	9	6.5×10^{-8}	6.0×10^{-7}	agrees
Raia	205	4	39	23	–	–	reference uses other condition identifiers
Raimundez	627	170	136	136	9.3×10^{-6}	3.3×10^{-5}	agrees
Salazar-Cavazos	18	4	6	6	7.5×10^{-9}	1.2×10^{-7}	agrees
Schwen	286	19	30	28	2.0×10^{-8}	2.1×10^{-7}	agrees
Smith	62	35	25	25	2.6×10^{-8}	7.2×10^{-7}	agrees
Sneyd	135	9	15	14	2.0×10^{-7}	5.3×10^{-1}	median exact; the maximum comes from duplicated rows at the same time
Weber	135	2	36	31	1.0×10^{-10}	1.5×10^{-8}	agrees
Zhao	82	7	28	21	–	–	reference file has 169 rows
Zheng	60	1	46	45	1.1×10^{-1}	9.0×10^{-1}	reference file contains the measured data, not a simulation

Froehlich is not included: its 9,169 experimental conditions, each requiring a simulation of a model with 1,228 species, made the analysis of its measured outputs infeasible with the resources available (Supplementary Note *Computational cost*).

Supplementary Note: Robustness to the size of parameter changes

This note describes in detail how the robustness of the selected subsets to the size of the parameter changes was assessed (Fig. 5 of the main text) and gives, for every admissible problem, the full distribution of the fitting error over the random scenarios, not only its median.

Random scenarios

For a problem with p estimated parameters, fifteen random directions $u^{(1)}, \dots, u^{(15)} \in [-1, 1]^p$ are drawn once, with independent components $u_j^{(k)} \sim \mathcal{U}(-1, 1)$, from a pseudo-random generator with fixed seed (NumPy `default_rng(42)`). For a perturbation size ε, scenario k changes all parameters simultaneously:

$$\theta^{(k)}(\varepsilon) = \theta_0 \odot \left(\mathbf{1} + \varepsilon\, u^{(k)}\right), \qquad \theta_j^{(k)}(\varepsilon) = \theta_{0,j}\left(1 + \varepsilon\, u_j^{(k)}\right), \quad j = 1, \dots, p,$$

so that every parameter is multiplied by a factor between $1-\varepsilon$ and $1+\varepsilon$. The same fifteen directions are used at every size $\varepsilon \in \{0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50\}$; only ε changes (paired design). Differences between sizes are therefore due to the size of the change and not to different random draws, and the scenarios at $\varepsilon = 0.05$ are exactly those used to select the subset. The directions of every problem are stored in the result files (`results/robustness/<problem>.json`, key `u`), and the script checks that they are identical at every size.

Errors in each scenario

Let $y(\theta) \in \mathbb{R}^m$ be the vector of noise-scaled measured outputs (main text, Methods) and S the selected biological parameters together with the calibration parameters Θ_C. For scenario k,

$$\Delta y^{(k)}_{\text{full}} = y\big(\theta^{(k)}\big) - y(\theta_0), \qquad \Delta y^{(k)}_{S} = y\big(\theta^{(k)}_{S}\big) - y(\theta_0), \qquad \big(\theta^{(k)}_{S}\big)_j = \begin{cases} \theta^{(k)}_j, & j \in S \\ \theta_{0,j}, & j \notin S, \end{cases}$$

that is, Δy_S is the output change when only the retained parameters take their scenario values. The error without re-estimation and the cosine are

$$e^{(k)}_{\text{rel}} = \frac{\|\Delta y^{(k)}_{\text{full}} - \Delta y^{(k)}_{S}\|}{\|\Delta y^{(k)}_{\text{full}}\|}, \qquad \cos\Delta^{(k)} = \frac{\langle \Delta y^{(k)}_{\text{full}}, \Delta y^{(k)}_{S} \rangle}{\|\Delta y^{(k)}_{\text{full}}\|\,\|\Delta y^{(k)}_{S}\|}.$$

The error after re-estimation is the smallest relative distance between the reduced model and the output of the full model in the scenario, when the retained parameters are free and all others are fixed at θ_0:

$$e^{(k)}_{\text{fit}} = \min\Big\{ e^{(k)}_{\text{rel}},\ \min_{v \in B} \frac{\|y(\theta_0[S \leftarrow v]) - y(\theta^{(k)})\|}{\|\Delta y^{(k)}_{\text{full}}\|} \Big\}, \qquad B = \prod_{j \in S} \Big[\tfrac{|\theta_{0,j}|}{10},\ 10\,|\theta_{0,j}| \Big],$$

where $\theta_0[S \leftarrow v]$ is θ_0 with the components in S replaced by v. The inner minimisation is a bounded nonlinear least-squares problem in $\log|v|$ (trust-region reflective method, at most $100(|S|+1)$ function evaluations, finite-difference step 10^{-3}), started from the scenario values (clipped to B) and from θ_0; the smaller result is kept. Since $v = \theta^{(k)}_S$ is admissible when it lies in B, $e_{\text{fit}} \le e_{\text{rel}}$ always holds (this is also proved in Lean, step 5). For the problems whose simulations are too slow for repeated nonlinear fits (Alkan, Bachmann, Chen, Isensee, Lucarelli, Raimundez), the first-order version is used,

$$e^{(k)}_{\text{fit,lin}} = \min\Big\{ e^{(k)}_{\text{rel}},\ \min_{c \in \mathbb{R}^{|S|}} \frac{\|\Delta y^{(k)}_{\text{full}} - J_S\, c\|}{\|\Delta y^{(k)}_{\text{full}}\|} \Big\},$$

with J_S the columns of the relative sensitivity matrix for S (central differences, step 1%); it differs from the exact value by a second-order term (Lean, step 12). A scenario is discarded when the simulation fails or when the full-model change is negligible, $\|\Delta y^{(k)}_{\text{full}}\| < \max(10^{-10}\|y(\theta_0)\|, 10^{-6})$, because the relative errors are then undefined; the number $n \le 15$ of valid scenarios is reported for every problem and size. For Chen some simulations fail in a run-dependent way, so the set of valid scenarios (7 to 13 of 15) differs slightly between runs; the medians of this rerun differ from those of the main text by at most 1.2 percentage points and the pattern is unchanged. At ±5% the scenarios are those of the selection. For the six problems decided to first order (Alkan, Bachmann, Chen, Isensee, Lucarelli, Raimundez) this analysis uses the first-order e_{fit}, whereas Table 1 of the main text gives the nonlinear re-estimation of the final subset; the two differ by at most 0.2 percentage points except in Isensee (9.8% here to first order, 11.3% nonlinear in Table 1). In the other problems the median e_{fit} agrees with Table 1 to within 10^{-5}. Isensee is, with Chen, the problem in which most simulations fail (only 1 to 6 of the 15 scenarios per level can be simulated), so the median is taken over very few scenarios and depends on which of them succeed; both values are far below τ_e.

Summaries over the scenarios

For each problem and size ε, with $e^{(1)}, \dots, e^{(n)}$ the values of e_{fit} in the valid scenarios:

- median $\tilde{e}_{\text{fit}}$, first and third quartiles Q_1, Q_3 (interquartile range $Q_3 - Q_1$), and the 5th and 95th percentiles P_5, P_{95}, computed by linear interpolation between order statistics (with $n = 15$, P_5 and P_{95} lie between the two smallest and the two largest values);
- maximum $\max_k e^{(k)}_{\text{fit}}$, the worst scenario;

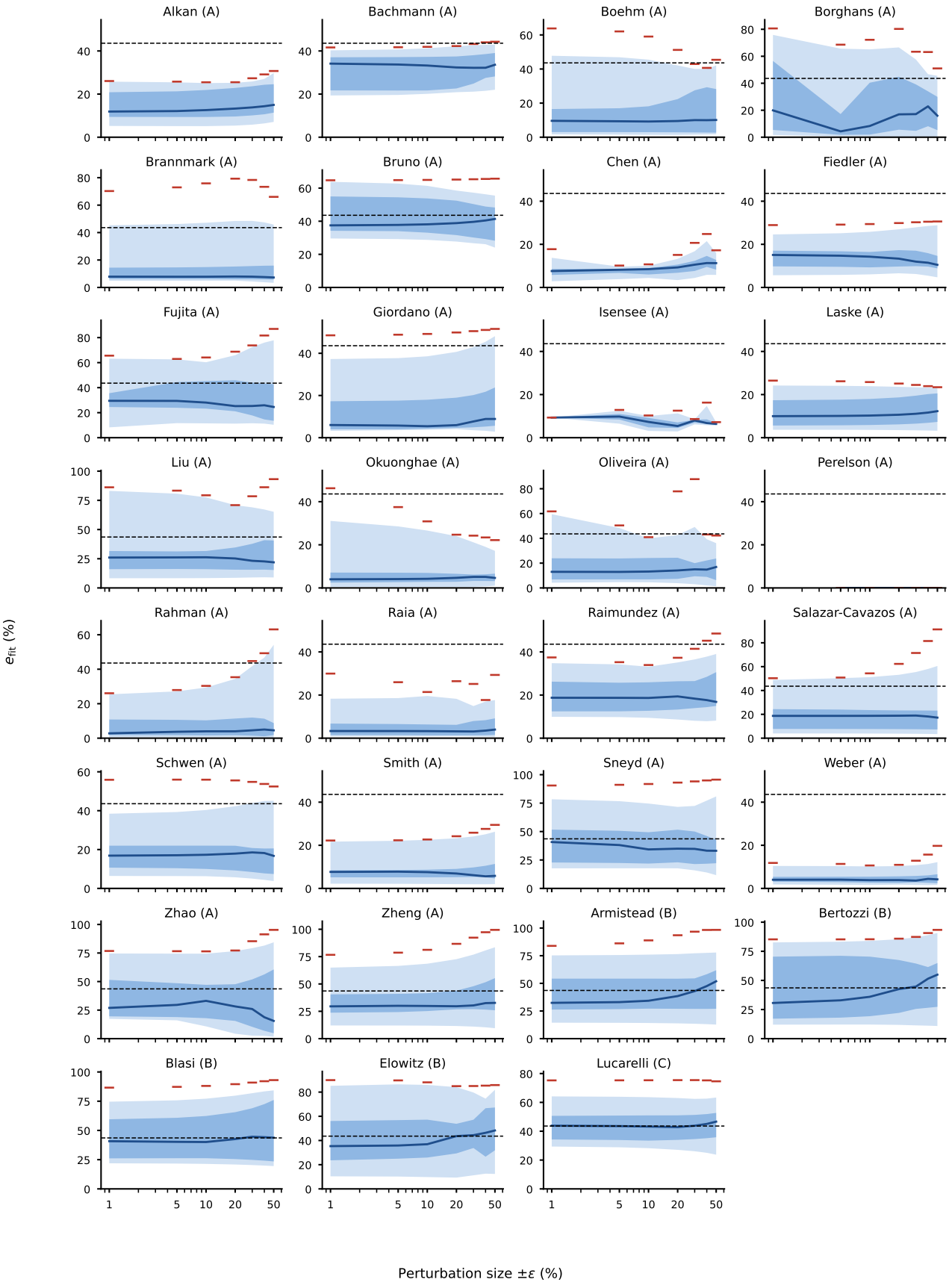
- pass rate, the fraction of scenarios in which the reduced model reproduces the full one,

$$\pi(\varepsilon) = \frac{\#\{k:\ e_{\mathrm{fit}}^{(k)} \le \tau_e\}}{n}, \qquad \tau_e = \sqrt{1 - 0.9^2} = 0.436.$$

Admissibility at ±5% and the robustness patterns of the main text use the median: pattern A, $\tilde{e}_{\mathrm{fit}} \le \tau_e$ at every size; B, admissible up to at least ±10%, first failing at ±20% or more and not recovering; C, irregular (fails and is admissible again at a larger size); D, not admissible at ±10% or less. The median requires that at least half of the scenarios are reproduced; the pass rate and the maximum show how many, and how badly the others fail.

Results

At ±5%, the size used for selection, all 31 admissible problems reproduce the full model in at least half of the scenarios by construction, and 13 of them in all scenarios. Supplementary Fig. S4 shows, for every problem, the median, the interquartile range, the 5–95% range and the worst scenario at each size; Supplementary Table S11 gives the pass rate and Supplementary Table S12 the full summary at ±5% and ±50%.



Supplementary Figure S4. Fitting error over the fifteen scenarios at each perturbation size. For every admissible problem (pattern in parentheses): median $\tilde{e}_{\mathrm{fit}}$ (line), interquartile range (dark band), 5–95% range (light band) and worst scenario (red marks); dashed line, $\tau_e = 43.6\%$. The same fifteen directions are used at every size.

Supplementary Table S11. Pass rate $\pi(\varepsilon)$: percentage of valid scenarios with $e_{\text{fit}} \leq \tau_e$, and number of valid scenarios when it is below 15 (in parentheses).

Problem	±1%	±5%	±10%	±20%	±30%	±40%	±50%	Pattern
Alkan	100	100	100	100	100	100	100	A
Bachmann	100	100	100	100	100	93	93	A
Boehm	93	93	93	93	100	100	93	A
Borghans	60	80	73	73	73	93	93	A
Brannmark	93	93	93	93	93	93	93	A
Bruno	60	60	60	60	60	60	67	A
Chen	100 (11)	100 (8)	100 (9)	100 (13)	100 (12)	100 (8)	100 (8)	A
Fiedler	100	100	100	100	100	100	100	A
Fujita	80	67	67	73	73	73	73	A
Giordano	93	93	93	93	93	93	87	A
Isensee	100 (1)	100 (4)	100 (6)	100 (6)	100 (3)	100 (5)	100 (2)	A
Laske	100	100	100	100	100	100	100	A
Liu	80	80	80	80	80	73	80	A
Okuonghae	93	100	100	100	100	100	100	A
Oliveira	87	87	100	93	93	100	100	A
Perelson	–	100 (2)	100 (11)	100	100	100	100	A
Rahman	100	100	100	100	93	87	87	A
Raia	100	100	100	100	100	100	100	A
Raimundez	100	100	100	100	100	93	93	A
Salazar-Cavazos	87	87	87	87	87	87	87	A
Schwen	93	93	93	93	93	93	93	A
Smith	100	100	100	100	100	100	100	A
Sneyd	67	60	60	60	60	73	73	A
Weber	100	100	100	100	100	100	100	A
Zhao	60	73	73	73	67	67	60	A
Zheng	73	73	73	73	73	67	67	A
Armistead	60	60	60	53	53	47	47	B
Bertozzi	60	60	60	53	47	47	40	B
Blasi	53	53	53	53	47	47	47	B
Elowitz	60	60	60	47	47	40	33	B
Lucarelli	47	53	53	53	47	47	47	C

Supplementary Table S12. Summary of e_{fit} (%) over the scenarios at ±5% and ±50%: median, interquartile range $[Q_1, Q_3]$, 5–95% range $[P_5, P_{95}]$, maximum and pass rate π.

	±5%					±50%				
Problem	Median	$[Q_1, Q_3]$	$[P_5, P_{95}]$	Max	π	Median	$[Q_1, Q_3]$	$[P_5, P_{95}]$	Max	π
Alkan	12	[9, 21]	[5, 25]	26	100	15	[11, 25]	[7, 30]	31	100
Bachmann	34	[22, 37]	[20, 41]	42	100	34	[28, 39]	[22, 43]	44	93
Boehm	9	[3, 17]	[1, 47]	62	93	10	[3, 28]	[2, 42]	45	93
Borghans	4	[2, 17]	[1, 66]	69	80	16	[5, 30]	[0, 46]	51	93
Brannmark	8	[6, 15]	[5, 46]	73	93	7	[6, 16]	[3, 46]	66	93
Bruno	38	[34, 54]	[29, 63]	65	60	41	[28, 48]	[24, 55]	66	67
Chen	8	[7, 9]	[4, 10]	10	100	11	[8, 13]	[6, 16]	17	100
Fiedler	15	[10, 17]	[6, 25]	29	100	10	[9, 15]	[5, 29]	31	100
Fujita	29	[24, 44]	[12, 63]	63	67	25	[14, 43]	[10, 78]	87	73
Giordano	6	[4, 18]	[4, 38]	49	93	9	[6, 24]	[3, 48]	52	87
Isensee	10	[8, 11]	[6, 13]	13	100	6	[6, 7]	[6, 7]	7	100
Laske	10	[6, 18]	[4, 24]	26	100	12	[7, 21]	[3, 23]	23	100
Liu	26	[16, 31]	[8, 81]	83	80	22	[15, 41]	[9, 65]	93	80
Okuonghae	4	[3, 7]	[0, 29]	37	100	5	[3, 7]	[1, 17]	22	100
Oliveira	13	[7, 24]	[5, 48]	50	87	17	[6, 24]	[1, 36]	42	100
Perelson	0	[0, 0]	[0, 0]	0	100	0	[0, 0]	[0, 0]	0	100

	±5%					±50%				
Problem	Median	$[Q_1, Q_3]$	$[P_5, P_{95}]$	Max	π	Median	$[Q_1, Q_3]$	$[P_5, P_{95}]$	Max	π
Rahman	4	[2, 11]	[1, 27]	28	100	5	[2, 9]	[0, 54]	63	87
Raia	3	[1, 7]	[1, 19]	26	100	4	[2, 9]	[1, 18]	29	100
Raimundez	19	[12, 26]	[10, 34]	35	100	17	[15, 31]	[8, 39]	49	93
Salazar-Cavazos	19	[8, 24]	[4, 50]	51	87	17	[7, 23]	[3, 61]	91	87
Schwen	17	[10, 22]	[6, 39]	56	93	17	[7, 21]	[4, 45]	52	93
Smith	8	[5, 9]	[2, 22]	22	100	6	[5, 11]	[2, 26]	29	100
Sneyd	38	[22, 51]	[18, 77]	91	60	33	[22, 43]	[12, 81]	96	73
Weber	4	[3, 5]	[2, 10]	11	100	4	[2, 7]	[2, 12]	20	100
Zhao	30	[19, 49]	[16, 74]	77	73	16	[5, 61]	[2, 84]	95	60
Zheng	30	[24, 41]	[12, 66]	79	73	33	[26, 55]	[10, 84]	99	67
Armistead	33	[27, 54]	[14, 76]	86	60	52	[27, 62]	[13, 78]	98	47
Bertozzi	33	[18, 71]	[12, 83]	85	60	55	[27, 65]	[11, 91]	93	40
Blasi	40	[26, 61]	[22, 76]	87	53	44	[24, 76]	[20, 84]	93	47
Elowitz	36	[25, 57]	[10, 86]	90	60	48	[32, 67]	[12, 82]	86	33
Lucarelli	44	[34, 51]	[29, 64]	75	53	47	[36, 53]	[24, 63]	75	47

Independent directions at each size

Purpose. The paired design above reuses the same fifteen directions at every perturbation size, so that the differences between sizes are due only to the size of the change. Its limitation is that the sizes above ±5% are not tested with new random draws: a subset could appear robust because the particular fifteen directions happen to be favourable, or fragile because a few unfavourable directions are repeated at every size. To separate the effect of the size from that of the particular directions, the whole evaluation was repeated with independent directions at each size.

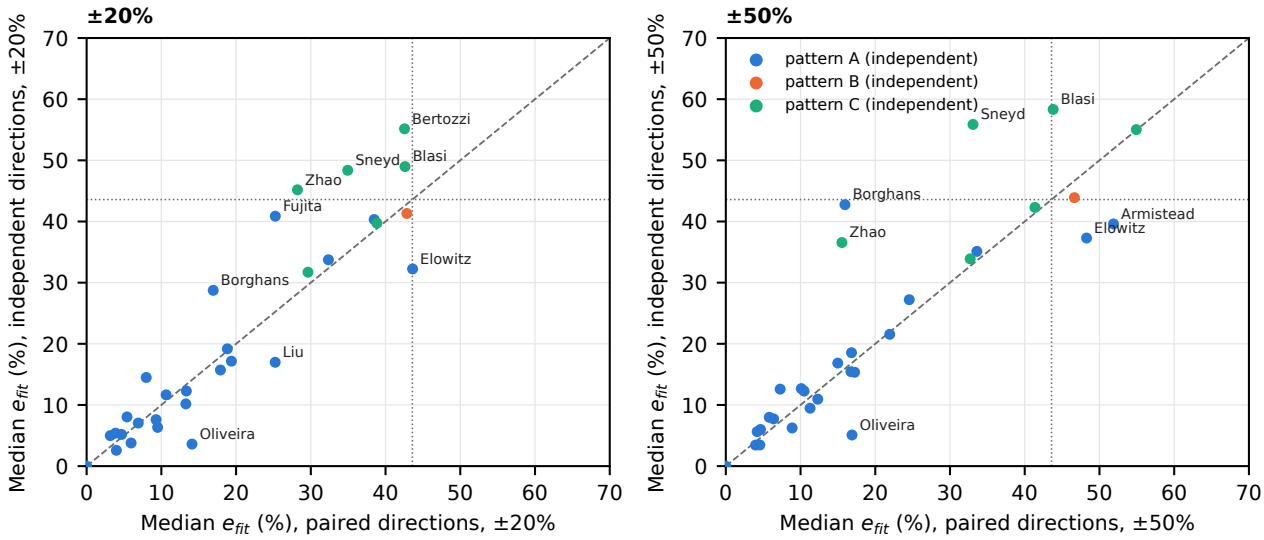
Design. For each admissible problem, with its final subset fixed and its calibration parameters re-estimated as above, the fifteen directions at size ε were drawn with seed $42 + 1000\,i$, where i is the position of ε in the list of sizes, except at ±5%, which keeps seed 42 and therefore the scenarios of the selection. Everything else (the discarding of scenarios, the re-estimation, the first-order e_{fit} of the six large admissible problems, the median and the classification into patterns) is identical to the paired design (`robustness.py` with `ROBUST_INDEP=1` in the reproducibility package; results in `results/robustness_independent/`).

Results. Supplementary Table S13 gives the pattern of every problem with both designs and the median e_{fit} with independent directions at every size, and Supplementary Figure S5 compares the medians of the two designs at ±20% and ±50%. Of the 31 admissible problems, the paired design classifies 26 as pattern A, 4 as B and 1 as C, and the independent design 24, 1 and 6. Twenty-two problems are of pattern A in both designs, and their medians with the two designs lie close to each other and far below the threshold. Nine problems change: Bruno, Sneyd, Zhao and Zheng (A → C), Armistead and Elowitz (B → A), Bertozzi and Blasi (B → C) and Lucarelli (C → B). All nine have medians within about ten percentage points of τ_e at the larger sizes in at least one design, and the median at a given size can differ by up to 27 percentage points between the two sets of directions (Borghans at ±50%, which nevertheless keeps pattern A). Pattern C, which requires the median to cross the threshold more than once, is the typical outcome when the median fluctuates around τ_e, and it becomes more frequent with independent directions because each size then adds its own sampling fluctuation.

Interpretation. The statement that most subsets remain adequate for parameter changes of up to ±50% does not depend on the particular directions: it holds for the 22 problems of pattern A in both designs. For the nine problems whose pattern changes, fifteen scenarios per size are not enough to establish where admissibility is lost, and their pattern should be read as "close to the threshold at large perturbations" rather than as a definite collapse or recovery. Admissibility at ±5%, on which the selection and all other analyses rely, is not affected, because both designs use the same scenarios at that size.

Supplementary Table S13. Robustness with independent directions at each perturbation size. Pattern with the paired design (the same fifteen directions rescaled at every size, as in the main text) and with independent directions (new directions at every size; in bold when the pattern changes); median e_{fit} (%) over the valid scenarios at each size with independent directions; and pass rate π (%) at ±50%. At ±5% both designs use the scenarios of the selection.

Problem	Paired	Indep.	±1%	±5%	±10%	±20%	±30%	±40%	±50%	π_{50}
Alkan	A	A	24	12	16	10	17	21	17	93
Armistead	B	**A**	35	33	33	40	31	35	40	53
Bachmann	A	A	32	34	31	34	29	33	35	87
Bertozzi	B	**C**	49	33	30	55	36	27	55	27
Blasi	B	**C**	44	40	33	49	37	46	58	27
Boehm	A	A	7	9	11	6	20	9	13	93
Borghans	A	A	10	4	8	29	32	23	43	53
Brannmark	A	A	18	8	12	14	9	16	13	100
Bruno	A	**C**	38	38	41	40	42	46	42	53
Chen	A	A	6	8	10	8	8	11	9	100
Elowitz	B	**A**	33	36	25	32	33	31	37	60
Fiedler	A	A	11	15	10	12	8	8	12	100
Fujita	A	A	23	29	39	41	35	29	27	73
Giordano	A	A	4	6	3	4	9	5	6	100
Isensee	A	A	6	10	6	8	7	6	8	100
Laske	A	A	13	10	15	12	9	8	11	100
Liu	A	A	11	26	24	17	28	26	22	80
Lucarelli	C	**B**	41	44	35	41	44	51	44	40
Okuonghae	A	A	4	4	5	5	11	4	6	93
Oliveira	A	A	9	13	14	4	17	17	5	93
Perelson	A	A	–	0	0	0	0	0	0	100
Rahman	A	A	3	4	6	3	9	2	3	100
Raia	A	A	6	3	3	5	3	3	3	100
Raimundez	A	A	23	19	17	17	18	18	19	100
Salazar-Cavazos	A	A	14	19	31	19	20	18	15	93
Schwen	A	A	23	17	15	16	18	22	15	100
Smith	A	A	8	8	9	7	7	8	8	100
Sneyd	A	**C**	38	38	54	48	44	36	56	20
Weber	A	A	4	4	5	5	4	3	6	100
Zhao	A	**C**	28	30	40	45	22	44	37	53
Zheng	A	**C**	35	30	27	32	23	45	34	73



Supplementary Figure S5. Median e_{fit} with paired and with independent directions. Each point is one admissible problem, at ±20% (left) and ±50% (right): median e_{fit} over the scenarios with the fifteen directions of the selection rescaled to that size (x) and with fifteen new directions drawn for that size (y), coloured by the pattern obtained with independent directions. Dashed, identity; dotted, $\tau_e = 43.6\%$. Problems far from the identity or on different sides of the threshold are labelled.

Supplementary Note: Random scenarios and large problems

This note describes in detail three aspects of the computations that the main text and the other notes only summarise: how the random scenarios treat parameters whose nominal values are negative, zero or extremely small; how the problems with the most expensive simulations were computed, and how the results were checked across machines; and why the greedy search of the method does not reach an admissible subset in Lang although one exists.

Construction of the random scenarios

Every random scenario used by the method (selection, admissibility test, robustness, the admissibility test applied to the other criteria and the error in units of the noise) has the form

$$\theta_j^{(k)} = \theta_{0,j}\left(1 + \varepsilon\, u_j^{(k)}\right), \qquad u_j^{(k)} \sim \mathcal{U}(-1,1), \quad k = 1,\dots,15,$$

with the directions $u^{(k)}$ drawn from a generator with fixed seed (42), so that all these analyses use exactly the same scenarios and their comparisons are paired. Because $|\varepsilon u_j^{(k)}| \le \varepsilon \le 0.5$, the factor $1 + \varepsilon u_j^{(k)}$ lies in $[1-\varepsilon, 1+\varepsilon] \subset (0,2)$: every parameter keeps its sign and changes by at most a fraction ε of its value. No clipping or lower bound is applied to the perturbed values, so that the relative size of the change is the same for every parameter whatever the magnitude or sign of its nominal value. Three groups of parameters deserve comment:

- *Negative nominal values.* Eleven estimated parameters have negative nominal values: `alpha_cer` in Armistead, a factor of the ceramide feedback, and ten parameters of Raimundez that describe differences between the two cell lines on a linear scale. They are perturbed by $\pm\varepsilon$ of their value like the others, and in every re-estimation their sign is fixed and $\log|\theta_j|$ is optimised within $[\log(|\theta_{0,j}|/10), \log(10|\theta_{0,j}|)]$.
- *Very small positive values.* Lang has 44 estimated parameters with nominal values below 10^{-12} (initial amounts and rate constants estimated on a logarithmic scale, with lower bound 10^{-24} in

the benchmark). They are perturbed relatively, like all others, so their perturbations have the same relative size.

- *Zero nominal values.* Six estimated parameters have nominal value zero (four in Lang, two in Raimundez). The simulator represents them by 10^{-12}; their relative perturbations, and therefore their columns of the relative sensitivity matrix J and their finite-perturbation effects, are negligible, so the proposed method never selects them. Pivoted QR, which normalises every column of its sensitivity matrix to unit length and uses an absolute finite-difference step for very small values, can select them: one of its two parameters in Raimundez (`d_MAPKtotal__MKN1_2_HS746T`) has nominal value zero.

A scenario is discarded when the simulation fails or when the change of the full model is negligible ($\|\Delta y_{\text{full}}\| < \max(10^{-10}\|y(\theta_0)\|, 10^{-6})$), and the number of valid scenarios is reported in every table.

Computation of the problems with expensive simulations

Eight problems have simulations slow enough to make the repeated nonlinear re-estimation of the admissibility test impractical (Alkan, Bachmann, Beer, Chen, Isensee, Lang, Lucarelli and Raimundez); for them e_{fit} is computed to first order (Methods), and six of them are admissible. These problems, and the analyses that need many fits of a large model, were computed on shared servers with 128 and 256 cores; the others on an 8-core workstation. The following implementation details matter for reproducing the results:

- *Parallelisation.* The columns of J and the stage-2 ordering by finite effects are distributed over processes; in the nested validation the finite-difference Jacobian of each fit of the full model is distributed as well (70 processes in Raimundez, which reduced the computation from hours to about a minute), and in the identifiability analysis the points of the profiles are computed in parallel.
- *Caches.* For problems with more than 200 parameters, J, the stage-2 ordering and the e_{fit} of the subsets already evaluated are stored on disk so that an interrupted run can resume. Each setting of the method (Supplementary Note *Sensitivity to the method settings*) uses its own cache, because these quantities depend on the finite-difference step, the integrator tolerances and the perturbation size.
- *Full model.* The full model is re-estimated on real data in every admissible problem (up to 155 free parameters, in Chen), and in the synthetic experiment only when it has at most 100 free parameters once those acting only in test conditions are removed, which excludes Chen and Raimundez. All real-data numbers of the paper come from a single run, the nested validation (Supplementary Note *Nested validation without the θ_0 anchor*).
- *Checkpoints.* Long fits write checkpoints identified by the problem and the number of free parameters. When an analysis is repeated with a different subset of the same size, the results and checkpoints of that problem must be deleted first; otherwise an earlier fit would be resumed. Likewise, the admissibility test of the subsets of the other criteria must run after their selections (pivoted QR, Brun, Morris, profile likelihood) have been computed.

Agreement between machines

The selection of Lang and Raimundez was computed independently on the workstation and on the 128-core server, with the same code and NumPy (2.4.6) and SciPy (1.17.1) but different versions of the simulator (libRoadRunner 2.9.2 on the workstation and 2.10.0 on the server): both runs gave the same subsets, and the median e_{fit} agreed to within 10^{-8}. All the analyses of Armistead were also computed on both machines; the test χ^2/n of the reduced model in the nested validation agreed to within 10^{-8} (1.2181).

Dependence of the greedy search on its path: Lang

In Lang, an oscillating cell-cycle model with 294 estimated parameters (277 of which affect the outputs) simulated to $t = 136{,}472$, the method stops without an admissible subset. Stage 1 does not reach the energy threshold τ_R (the 13 parameters accepted by the conditioning test keep 17% of the sensitivity energy), so stage 2 orders the parameters by their finite effect. The first parameter is `kDeE2f1` (degradation of E2F1) and the second `kDeCe`; the pair is not admissible, and the search keeps adding the next parameters in this order as long as the conditioning test is satisfied, until no further parameter can be added ($\kappa = 8.7$, VIF $= 9.99$, just below the limit of 10). The resulting 13 parameters have a median first-order e_{fit} of 56%, above τ_e. Because adding parameters can never increase the first-order e_{fit} (Lean, `fitErr_union_le`), no subset of these 13 parameters is admissible either.

Admissible pairs nevertheless exist outside this path. With the same scenarios and only the conditioning threshold changed (Supplementary Table S14), the search accepts a different second parameter and finds {`kDeE2f1`, `kSySkp2`} ($e_{\text{fit}} = 13\%$, with $\tau_\kappa = \tau_{\text{VIF}} = 5$) and {`kDeE2f1`, `kSyCb1`} (16%, with 30); with finite-difference steps of 0.1% or 5% it finds {`kDeE2f1`, `kAspEB55`} and {`kDeE2f1`, `kSyCb1`}. With perturbations of ±1%, or with coarser integrator tolerances ($10^{-6}/10^{-8}$), the pair {`kDeE2f1`, `kDeCe`} itself becomes admissible (8% and 13%); this last result shows that the first-order e_{fit} of this long oscillatory simulation is sensitive to the accuracy of the integration, which agrees with the discrepancy between the simulation of Lang and the reference simulation of the benchmark (Supplementary Note *Verification of the simulator*). Lang is therefore non-admissible under the method as defined, but the failure reflects the greedy nature of the search, which commits to the second parameter with the largest finite effect, and not the absence of a small admissible subset; a search that explored several candidates for each position (for example a beam search) would find one. The method was kept unchanged and Lang is reported as not admissible.

Settings variants of Raimundez

Raimundez, with 136 estimated parameters of which 79 are calibration parameters, selects {`kdeg_membran__MKN1`, `ka_MAPK__MKN1`} in every computed variant of the settings, with a median e_{fit} between 18% and 19%; the selection of this problem is insensitive to the settings. Two variants of Lang (±10% and tolerances $10^{-9}/10^{-12}$) and one of Raimundez ($\tau_e = 0.10$) did not finish in the available computing time.

Supplementary Table S14. Settings variants of Lang and Raimundez, computed as in Supplementary Table S33 (one setting changed at a time, all others at their baseline values, the same scenarios). Cell: selected subset and median e_{fit}; =, the same subset as the baseline; ×, not admissible (number of parameters at which the search stopped and its e_{fit}). Not computed: variants that did not finish in the available computing time.

Setting	Lang	Raimundez
Baseline	× (13 par., 56%)	`kdeg_membran__MKN1`, `ka_MAPK__MKN1` (19%)
$\delta = 0.1\%$	`kDeE2f1`, `kAspEB55` (11%)	= (19%)
$\delta = 5\%$	`kDeE2f1`, `kSyCb1` (16%)	= (19%)
tolerances $10^{-9}/10^{-12}$	not computed	= (19%)
tolerances $10^{-6}/10^{-8}$	`kDeE2f1`, `kDeCe` (13%)	= (19%)
perturbation ±1%	`kDeE2f1`, `kDeCe` (8%)	= (19%)
perturbation ±10%	not computed	= (18%)
$\tau_R = 0.80$	× (13 par., 56%)	= (19%)
$\tau_R = 0.85$	× (13 par., 56%)	= (19%)
$\tau_R = 0.95$	× (13 par., 56%)	= (19%)
$\tau_\kappa = \tau_{\text{VIF}} = 5$	`kDeE2f1`, `kSySkp2` (13%)	= (19%)

Setting	Lang	Raimundez
$\tau_\kappa = \tau_{\mathrm{VIF}} = 30$	`kDeE2f1`, `kSyCb1` (16%)	= (19%)
$\tau_e = 0.10$	× (13 par., 56%)	not computed
$\tau_e = 0.20$	× (13 par., 56%)	= (19%)
$\tau_e = 0.30$	× (13 par., 56%)	= (19%)
$\tau_e = 0.50$	× (13 par., 56%)	= (19%)

Supplementary Note: Synthetic prediction experiment

With real data the true parameters are unknown and, in addition, the published estimate θ_0 at which the non-selected parameters are fixed was obtained using the test experiments (Supplementary Note *Nested validation without the θ_0 anchor*). To measure the prediction error itself, without either limitation, the external validation was repeated with synthetic data generated from a known truth that θ_0 has not seen.

Design

1. *Truth.* For each seed $s \in \{42, \dots, 61\}$, a vector $u \in [-1, 1]^p$ with components $u_j \sim \mathcal{U}(-1, 1)$ is drawn (NumPy `default_rng`(s)) and $\theta^*_j = \theta_{0,j}(1 + 0.2\, u_j)$ for every parameter that can be estimated from the training data (parameters acting only in test conditions stay at $\theta_{0,j}$, since no method could learn them).
2. *Data.* $\tilde{y}_i = y_i(\theta^*) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$ in noise-scaled units, i.e. Gaussian noise with the standard deviation σ_i recorded in the benchmark, drawn from the same generator.
3. *Split and fits.* The same training/test split as for the real data. The reduced model (re-estimating $S \cup \Theta_C$) and the full model (re-estimating all parameters) are fitted to the synthetic training data from θ_0, with the same least-squares method and bounds. The full model was not re-estimated for Chen and Raimundez, which have more than 100 free parameters after removing those that act only in the test conditions (Bachmann has 101 parameters that affect the outputs but 84 free ones, and its full model was re-estimated).
4. *Prediction error.* On the test measurements,

$$e_{\mathrm{pred}} = \frac{\| y_{\mathrm{test}}(\hat{\theta}) - y_{\mathrm{test}}(\theta^*) \|}{\| y_{\mathrm{test}}(\theta^*) - y_{\mathrm{test}}(\theta_0) \|},$$

   the error of the prediction relative to the change of the true response; $e_{\mathrm{pred}} = 1$ for θ_0, and $e_{\mathrm{pred}} \leq \tau_e = 0.436$ is used as the success criterion.
5. *Signal and noise.* When the change from θ_0 to θ^* moves the test outputs less than the noise, the fits are dominated by the noise of the training data and e_{pred}, whose denominator is then small, is unstable. The signal is considered larger than the noise when the test χ^2/n of θ_0 on the synthetic data is at least 2.

The full-model fits, with many free parameters, can end in different local optima when they are repeated with other library versions, which does not change any qualitative conclusion.

Results

Over all seeds, 620 problem–seed combinations were planned (31 admissible problems × 20 seeds). Forty could not be evaluated because Bertozzi and Blasi have no training/test split, eight failed in Isensee because the synthetic truth θ^* could not be simulated, and in two the prediction error is undefined (Chen, seed 51, in which the fitted reduced model could not be simulated on the test conditions, and

Perelson, seed 50). The remaining 570 cases, 320 with signal and 250 without, are the denominators below. Among the cases with signal, the reduced model reached $e_{\text{pred}} \leq \tau_e$ in 178 of 320 (56%; between 31% and 76% depending on the seed) and predicted at least as well as the full model in 160 of 282 (57%; 40–73%). Without signal, the reduced model predicted at the noise level (test $\chi^2/n \leq 2$) in 197 of 250 cases.

The outcome is stable within each problem (Supplementary Table S15 and main-text Fig. 3b) and divides the benchmark into two groups. The reduced model predicts better than the full model in almost every seed where the full model is over-parameterised for the available data: Weber (20 of 20 seeds), Isensee (12 of 12), Fiedler and Bruno (6 of 6), Brannmark (17 of 18), Zheng (18 of 20) and Elowitz (14 of 17). It predicts worse in almost every seed of the other group, in which the full model, re-estimated on the training data, predicts well (the reason is not established; the full-model profiles of these problems show few identifiable parameters, Supplementary Note *Profile likelihood*): Oliveira (0 of 20), Armistead (1 of 12), Fujita (1 of 18), Okuonghae (2 of 20) and Salazar-Cavazos (3 of 18). Beating the full model does not imply an accurate prediction: in Brannmark the reduced model reaches the threshold in only 1 of 18 seeds, and in Raimundez, Zhao and Schwen it rarely does. Supplementary Table S16 gives the counts for every seed.

Supplementary Table S15. Synthetic prediction experiment, reduced and full model, 20 seeds (42–61). Seeds: valid seeds; signal: seeds in which the test χ^2/n of θ_0 on the synthetic data is at least 2. Median e_{pred} over valid seeds. Last two columns, over the seeds with signal: reduced model with $e_{\text{pred}} \leq \tau_e$, and reduced e_{pred} no larger than that of the full model (–, full model not re-estimated: more than 100 free parameters).

Problem	Seeds	Signal	Reduced e_{pred}	Full e_{pred}	Reduced $\leq \tau_e$	Reduced $\leq$ full
Alkan	20	3	0.306	0.326	3/3	1/3
Armistead	20	12	0.467	0.227	7/12	1/12
Bachmann	20	2	2.85	4.65	0/2	2/2
Bertozzi	*not evaluable*					
Blasi	*not evaluable*					
Boehm	20	4	1.51	5.65	1/4	2/4
Borghans	20	11	0.648	0.494	2/11	5/11
Brannmark	20	18	2.1	13.4	1/18	17/18
Bruno	20	6	0.429	11	4/6	6/6
Chen	19	18	0.144	–	17/18	–
Elowitz	20	17	0.293	0.454	11/17	14/17
Fiedler	20	6	1.15	3.51	1/6	6/6
Fujita	20	18	0.345	0.183	14/18	1/18
Giordano	20	0	346	515	0/0	–
Isensee	12	12	0.134	0.839	9/12	12/12
Laske	20	0	1.91	8.45	0/0	–
Liu	20	1	0.863	0.913	0/1	0/1
Lucarelli	20	0	1.01	3.27	0/0	–
Okuonghae	20	20	0.156	0.014	17/20	2/20
Oliveira	20	20	0.392	0.0111	12/20	0/20
Perelson	19	1	20.9	20.9	0/1	1/1
Rahman	20	19	0.118	0.372	15/19	13/19
Raia	20	8	0.453	0.467	7/8	5/8
Raimundez	20	20	0.571	–	3/20	–
Salazar-Cavazos	20	18	0.334	0.213	12/18	3/18
Schwen	20	9	2.63	2.53	0/9	5/9
Smith	20	20	0.0954	0.179	18/20	14/20
Sneyd	20	1	0.598	0.453	1/1	1/1
Weber	20	20	0.359	3.77	11/20	20/20
Zhao	20	16	1	1	0/16	11/16
Zheng	20	20	0.41	0.681	12/20	18/20

Supplementary Table S16. Synthetic prediction experiment, reduced and full model: counts per seed.

Seed	Problems	With signal	Reduced $\leq \tau_e$	Compared with full	Reduced $\leq$ full
42	28	13	8	11	6
43	28	19	12	17	9
44	29	16	7	14	7
45	28	14	9	12	7
46	29	17	9	15	10
47	28	14	8	12	6
48	29	12	6	11	6
49	28	15	10	13	9
50	27	17	6	15	11
51	28	18	9	17	10
52	29	20	11	18	11
53	29	17	8	15	6
54	29	15	9	13	6
55	29	17	10	15	11
56	29	16	5	14	6
57	29	14	9	12	6
58	28	17	13	15	10
59	28	17	10	15	8
60	29	16	10	14	7
61	29	16	9	14	8

Size of the signal and other amplitudes

Purpose. The experiment above has three limitations. First, it defines the cases with signal through the test χ^2/n of θ_0 on the synthetic data, a quantity that contains the noise realisation; a case could be classified by the particular noise drawn rather than by the true change. Second, e_{pred} divides by the true change of the test outputs, so when that change is small compared with the noise the measure is unstable and partly reflects the noise of the training data. Third, the truth was generated at a single distance from θ_0 ($\pm 20\%$), so the experiment says nothing about how the result depends on that distance. Two additional computations address these points.

Design. (i) *True signal.* For every problem and seed, the change of the true response, $y(\theta^*) - y(\theta_0)$, was computed without any fit or noise. Because the outputs are divided by the recorded noise standard deviations, its root mean square over the test measurements, $\|y_{\text{test}}(\theta^*) - y_{\text{test}}(\theta_0)\|/\sqrt{n_{\text{test}}}$, is the size of the signal in units of σ per measurement (`synthetic_seeds.py` with the switch `SINT_SOLO_SENAL=1`, signal only; results in `results/synthetic_signal_a005`, `_a020` and `_a050`). (ii) *Other amplitudes.* The complete experiment (truth, noisy data, the same split and the same fits of the reduced and the full model) was repeated with truths $\theta^* = \theta_0(1 + a\,u)$ for $a = 0.05$ and $a = 0.50$, with seeds 42–51, and compared with $a = 0.20$ on the same seeds (amplitude set by `SINT_AMPLITUD`; results in `results/synthetic_seeds_a005` and `_a050`). As in the main experiment, the full model was not re-estimated in Chen and Raimundez.

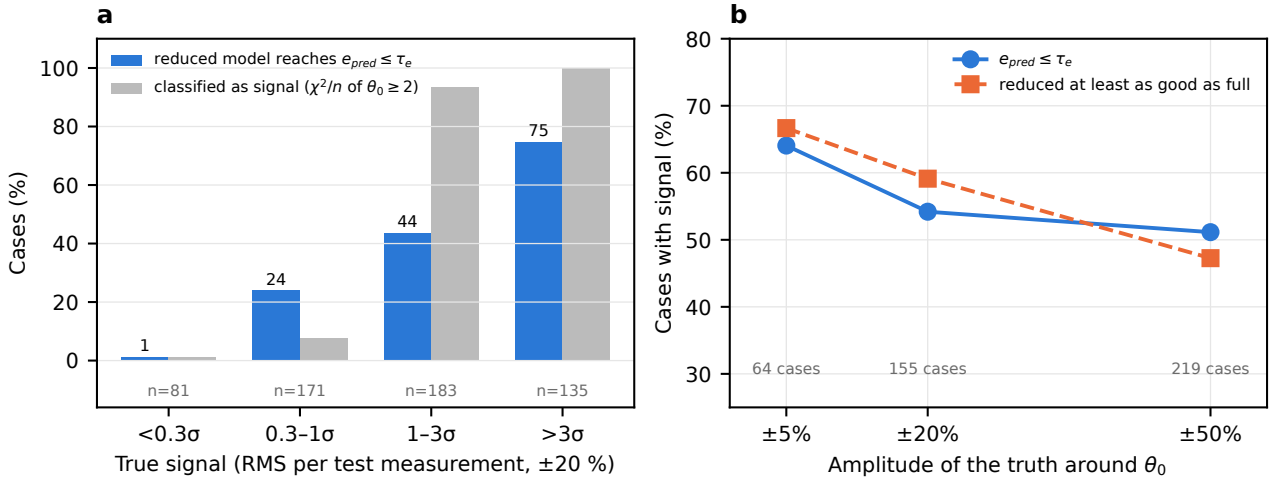
Results. Supplementary Table S17 gives the signal and the success per problem at the three amplitudes, and Supplementary Figure S6 summarises them. At $a = 0.20$ the median signal over all problems and seeds is 1.25σ per measurement (interquartile range 0.50–2.84); it is 0.31σ at $a = 0.05$ and 2.63σ at $a = 0.50$, and it varies by orders of magnitude between problems, from below $10^{-3}\sigma$ in Giordano and Perelson, whose recorded noise is far larger than the outputs, to tens or hundreds of σ in Okuonghae and Oliveira. The success of the reduced model grows steadily with the true signal: $e_{\text{pred}} \leq \tau_e$ in 1% of the 81 cases with a signal below 0.3σ, 24% of the 171 between 0.3σ and σ, 44% of the 183 between σ and 3σ and 75% of the 135 above 3σ. Below about one noise unit per measurement, e_{pred} therefore reflects mainly the noise of the training data. The criterion used above to define the cases with signal agrees closely with the true signal: it classifies as cases with signal 1% and 8% of the cases whose true

signal is below 0.3σ and between 0.3σ and σ, and 93% and 100% of those between σ and 3σ and above 3σ; the counts reported above are thus essentially the counts of cases with a true signal of at least one noise unit.

At the three amplitudes, the number of cases with signal grows from 64 (±5%) to 155 (±20%) and 219 (±50%), as the truth moves further from θ_0. Among these cases, the reduced model reached $e_{\mathrm{pred}} \leq \tau_e$ in 64%, 54% and 51%, and predicted at least as well as the full model in 67% (42 of 63), 59% (81 of 137) and 47% (95 of 201). The reduction is local: it predicts best when the truth is close to the reference θ_0 around which the subset was selected, its advantage over the full model disappears for changes of ±50%, where the full model, which can move all its parameters, more often recovers the truth, and yet it still reaches the prediction threshold in about half of the cases. This is consistent with the robustness analysis, in which most subsets remain admissible up to ±50%: reproducing the full model at large perturbations and predicting new data from noisy measurements at large distances from θ_0 are different tasks, and the second depends also on the noise and on the training data.

Supplementary Table S17. True signal and success of the reduced model in the synthetic experiment at three amplitudes of the truth, $\theta^* = \theta_0(1 + a\,u)$. Signal: median over the seeds of the root mean square of $y(\theta^*) - y(\theta_0)$ per test measurement, in units of the recorded noise σ (seeds 42–61 at $a = 0.20$, 42–51 otherwise). Success: seeds with $e_{\mathrm{pred}} \leq \tau_e$ among those with signal (test χ^2/n of $\theta_0 \geq 2$), seeds 42–51 at every amplitude.

Problem	Signal ±5%	±20%	±50%	Success ±5%	±20%	±50%
Alkan	0.22	0.65	2.2	–	3/3	9/10
Armistead	0.27	1.3	3.1	–	2/5	4/8
Bachmann	0.14	0.53	1.5	–	0/2	0/10
Boehm	0.15	0.46	1.7	0/1	1/3	4/6
Borghans	0.9	1.2	1.2	0/3	1/6	1/5
Brannmark	0.72	1.8	5.9	0/3	1/10	0/10
Bruno	0.16	0.72	1.5	–	1/1	4/8
Chen	0.51	2.5	6.2	1/1	8/8	7/8
Elowitz	0.62	2.3	4	1/1	5/9	7/10
Fiedler	0.23	0.89	2.1	–	0/4	5/10
Fujita	0.65	2.4	5.6	2/2	7/8	5/10
Giordano	8.2×10^{-5}	4.4×10^{-4}	9.0×10^{-4}	–	–	–
Isensee	0.66	2.7	6.5	–	4/4	3/5
Laske	0.095	0.31	0.92	–	–	2/4
Liu	0.068	0.31	0.59	–	–	0/1
Lucarelli	0.096	0.39	0.95	–	–	1/4
Okuonghae	59	2.1×10^2	4.0×10^2	9/10	8/10	8/10
Oliveira	1.4×10^2	6.1×10^2	9.7×10^2	5/10	4/10	2/10
Perelson	1.8×10^{-7}	7.3×10^{-7}	1.7×10^{-6}	–	–	–
Rahman	1	4.9	11	5/5	8/9	9/10
Raia	0.17	0.65	1.5	–	3/3	5/6
Raimundez	0.51	2	5.5	–	2/10	4/10
Salazar-Cavazos	0.44	1.9	4.4	1/3	5/9	5/9
Schwen	0.2	0.91	2.4	0/1	0/4	1/9
Smith	15	64	1.6×10^2	10/10	10/10	10/10
Sneyd	0.14	0.55	1.5	–	–	4/7
Weber	1.6	3.7	7.1	3/6	5/10	7/10
Zhao	0.35	1.7	6.8	0/1	0/7	0/9
Zheng	1.3	4.6	10	4/7	6/10	5/10



Supplementary Figure S6. Signal and amplitude in the synthetic experiment. a, Cases at ±20% (all seeds) grouped by the true signal, the root mean square of $y(\theta^*) - y(\theta_0)$ per test measurement in noise units: percentage in which the reduced model reaches $e_{\text{pred}} \le \tau_e$ (blue) and percentage classified as cases with signal by the criterion of the experiment, test χ^2/n of $\theta_0 \ge 2$ (grey); n, number of cases. **b**, Seeds 42–51 at the three amplitudes of the truth: among the cases with signal, percentage in which the reduced model reaches the threshold (circles) and in which it predicts at least as well as the full model (squares).

Supplementary Note: Single-parameter subsets

The method requires at least $k_{\min} = 2$ biological parameters. This note examines whether a single parameter could suffice.

Design

For every admissible problem the candidates are the ten biological parameters with the largest sensitivity energy E_j and the ten with the largest finite effect s_j (the two orderings of the method). Each candidate j is tested alone, with the calibration parameters Θ_C re-estimated together with it as in the method: $\{j\}$ is admissible if the median e_{fit} over the same 15 scenarios at ±5% is at most τ_e (without the requirement $|S| \ge 2$). For the best admissible candidate j^* (smallest median e_{fit}) the median e_{fit} was also computed at ±10% and its external validation, real and synthetic, exactly as for the selected subset S (Supplementary Note *Synthetic prediction experiment*). Bachmann, Chen, Isensee and Raimundez were not included because the repeated nonlinear re-estimations are too slow.

Results

A single parameter was admissible at ±5% in 15 of the 27 problems examined. In the synthetic prediction experiment the selected subset S predicted better than the best single parameter (smaller e_{pred}) in 9 of the 15 problems where both could be evaluated (Supplementary Table S18), but on real data the single parameter had the lower test χ^2/n in 10 of the 15 (Supplementary Table S18). These two comparisons are mixed; the test of Supplementary Note *Minimum subset size: one parameter or two* shows that the second parameter acts as a margin of safety: a single parameter is as identifiable and predictive, but fails in more individual scenarios, especially for large changes.

Supplementary Table S18. Problems in which a single biological parameter j^* is admissible: median e_{fit} (%) at ±5% and ±10%, test χ^2/n on real data and e_{pred} in the synthetic experiment, for $\{j^*\}$ and for the selected subset S (both with Θ_C re-estimated).

		e_{fit} ±5%		e_{fit} ±10%		Real χ^2/n		
Problem	j^*	j^*	S	j^*	S	j^*	S	e_{pred} j^* / S
Alkan	kt	21	12	13	12	2.74	2.75	0.13 / 0.13
Borghans	v0	10	4	14	8	0.74	0.74	1.01 / 1.04
Brannmark	km2	9	8	9	8	1.63	5.96	1.19 / 0.92
Fiedler	k10	15	15	15	14	11.68	11.21	0.16 / 0.23
Fujita	reaction_2_k1	20	29	22	28	5.96	5.94	0.53 / 0.95
Giordano	gamma_4	11	6	11	5	1.1×10^{-8}	1.0×10^{-7}	1.78 / 1,027
Laske	k_imp	16	10	16	10	0.45	0.76	3.27 / 0.50
Okuonghae	gamma_0	5	4	5	4	4.1×10^{4}	4.7×10^{4}	0.30 / 0.08
Oliveira	delta_	11	13	12	13	2.0×10^{4}	912	0.38 / 0.29
Rahman	infected_ moderate_ transmission_ rate	11	4	12	4	1.5×10^{-3}	2.0×10^{-3}	0.35 / 0.07
Raia	init_Rec_i	4	3	4	3	7.42	8.22	0.43 / 0.48
Salazar-Cavazos	ratio_kpkd_Y1068_ _FREE	30	19	29	19	100.00	102	0.79 / 0.17
Schwen	ini_R2fold	19	17	19	17	8.08	8.20	1.89 / 2.03
Smith	k30r	8	8	8	8	7.1×10^{4}	1.5×10^{4}	1.00 / 0.07
Weber	a21	6	4	6	4	0.99	1.00	0.94 / 0.63

Supplementary Note: Minimum subset size: one parameter or two

Question. The method adds biological parameters until re-estimating them reproduces the full model and then removes those that are not needed, but the admissibility rule requires $|S| \geq 2$, so the search cannot stop at one parameter and backward elimination cannot go below two. This is why 28 of the 31 admissible subsets have exactly two biological parameters: in those problems the size is set by the rule, not by the data, and in 15 of the 27 problems tested a single parameter already reaches the threshold (Supplementary Note *Single-parameter subsets*). This note asks whether requiring a second parameter has a measurable benefit or is an arbitrary choice.

Design. For each of these 15 problems, the best single parameter j^* (the admissible single parameter with the smallest median e_{fit} at ±5%, re-estimated together with the calibration parameters) was put through the same tests as the pair selected by the proposed method: robustness to the size of the parameter changes, from ±1% to ±50% with 15 scenarios per size and with paired and with independent directions (Supplementary Note *Robustness to the size of parameter changes*); admissibility at ±20% with the fixed scenario set; prediction of synthetic data with 20 seeds (Supplementary Note *Synthetic prediction experiment*); and practical identifiability by the profile likelihood of the reduced problem (Supplementary Note *Identifiability of the selected parameters*). The rule $|S| \geq 2$ is not applied to $\{j^*\}$. Pair and single parameter are compared problem by problem with the exact sign test (`scripts/run_test_kmin.sh`, `analysis/compare_single_vs_pair.py`; 27 min of elapsed time on an 88-core server and 5.7 h of processor time, 4.7 h of it for Smith; Supplementary Note *Computing time of the analyses*).

Results (Supplementary Table S19).

- *The median criterion does not distinguish one parameter from two.* With one parameter, the median e_{fit} stays below τ_e at every perturbation size in 14 of the 15 problems (pattern A; with the pair, 15), and both are admissible at ±20% in all 15 problems. The exception is Borghans, whose single parameter fails at ±1% and, with independent directions, at several larger sizes.

- *Identifiability and prediction are the same.* Every selected parameter is identifiable in 12 of the 15 problems with the single parameter and in 10 with the pair ($P = 0.5$): a single parameter cannot be compensated by another selected parameter. The synthetic success rate is 55% (105 of 190 cases with signal) with the single parameter and 59% (113 of 190) with the pair; the pair is better in 3 problems and the single parameter in 4 ($P = 1$).
- *The pair fails in fewer individual scenarios and degrades less for large changes.* Over all perturbation sizes, 94% of the scenarios reproduce the full model with the pair and 88% with the single parameter with paired directions (the pair is better in 10 problems, the single parameter in 1; $P = 0.012$), and 92% against 87% with independent directions (11 against 1; $P = 0.006$). At ±50% with independent directions the median e_{fit} is 10.9% with the pair and 18.7% with the single parameter, larger with the single parameter in 14 of the 15 problems ($P < 0.001$; with paired directions 10.5% against 11.7%, 13 of 15, $P = 0.007$).

Interpretation. A second parameter is not needed for the reduced model to be admissible in the median, identifiable or predictive: in about half of the problems a single biological parameter, re-estimated with the calibration parameters, already does all three. What the second parameter adds is a margin of safety: the reduction fails in fewer individual parameter changes and its error grows less when the parameters change by large amounts. The minimum size $k_{\min} = 2$ is therefore a conservative choice, supported by this margin, rather than a requirement of the method; users who prefer the smallest possible model can set $k_{\min} = 1$ (`config.py`), at the cost of a reduction that is less robust to individual large changes.

Supplementary Table S19. Best single parameter against the pair selected by the proposed method, in the 15 problems in which one parameter already reaches the admissibility threshold at ±5%. Pass rate: percentage of the scenarios, over all perturbation sizes from ±1% to ±50%, with $e_{\text{fit}} \leq \tau_e$, with paired (P) and independent (I) directions; e_{fit} at ±50%: median over the 15 scenarios with independent directions (%); ±20%: admissible at ±20%; Synthetic: cases with signal in which $e_{\text{pred}} \leq \tau_e$ over the 20 seeds (identifiability is reported in the text). Each cell gives pair / single.

Problem	Single parameter	Pair	Pass rate P (%)	Pass rate I (%)	e_{fit} ±50%	±20%	Synthetic
Alkan	kt	kt, k_rep_nhej	100 / 81	96 / 83	17 / 26	yes / yes	3/3 / 3/3
Borghans	v0	Kz, K_par	78 / 68	66 / 51	43 / 47	yes / yes	2/11 / 1/11
Brannmark	km2	k1f, km2	93 / 91	99 / 99	13 / 19	yes / yes	1/18 / 1/18
Fiedler	k10	tau2, k6	100 / 100	99 / 99	12 / 13	yes / yes	1/6 / 1/6
Fujita	reaction_2_k1	init_AKT, reaction_6_k1	72 / 93	70 / 86	27 / 15	yes / yes	14/18 / 16/18
Giordano	gamma_4	alpha_4, alpha_28	92 / 80	97 / 86	6 / 33	yes / yes	– / –
Laske	k_imp	k_imp, k_syn_R_M	100 / 80	98 / 80	11 / 27	yes / yes	– / –
Okuonghae	gamma_0	transmission_rate_effective, gamma_i	99 / 99	96 / 94	6 / 7	yes / yes	17/20 / 18/20
Oliveira	delta_	beta_0, h_hosp_rate	94 / 91	90 / 89	5 / 17	yes / yes	12/20 / 12/20
Rahman	infected_moderate_transmission_rate	infected_moderate_transmission_rate, infected_normal_worsen_rate	95 / 84	99 / 83	3 / 22	yes / yes	15/19 / 17/19
Raia	init_Rec_i	SOCS3mRNA_production, init_Rec_i	100 / 97	98 / 94	3 / 5	yes / yes	7/8 / 7/8
Salazar-Cavazos	ratio_kpkd_Y1068__FREE	ratio_kpkd_Y1068__FREE, ratio_kpkd_YN__FREE	87 / 67	78 / 62	15 / 28	yes / yes	12/18 / 3/18
Schwen	ini_R2fold	ini_R1, ini_R2fold	93 / 89	97 / 94	15 / 19	yes / yes	0/9 / 0/9
Smith	k30r	k8, kminus7b	100 / 100	100 / 99	8 / 10	yes / yes	18/20 / 19/20
Weber	a21	s12, a21	100 / 100	100 / 100	6 / 6	yes / yes	11/20 / 7/20

Supplementary Note: Identifiability of the selected parameters

Question

The method selects a subset S that is *sufficient*: re-estimating S together with the calibration parameters Θ_C, with the other parameters at θ_0, reproduces the outputs of the full model. Sufficiency does not imply that the data determine the values of the selected parameters. A selected parameter could, for example, compensate for a scaling factor of the measurement, so that only their product is determined. This note examines whether the selected parameters are also practically identifiable in the reduced problem, that is, with $S \cup \Theta_C$ free and the remaining parameters at their published values.

Design

1. *Local check.* At θ_0, the columns of the sensitivity matrix for $S \cup \Theta_C$, normalised to unit length and not centred (the Fisher information of the reduced problem up to scaling), were examined: for each selected parameter, the variance inflation factor (VIF) within $S \cup \Theta_C$ and the column of Θ_C or S most similar to it.
2. *Profile likelihood of the reduced problem.* With all measurements of the problem and the noise levels of the benchmark, the reduced problem was fitted from θ_0 (minimum $\hat{\theta}$). For each selected parameter θ_j, $\log|\theta_j|$ was fixed at displacements of 0.05 to 4.6 from $\log|\hat{\theta}_j|$ and the other parameters of $S \cup \Theta_C$ were re-estimated, with the same least-squares method and bounds $[|\theta_0|/10, 10|\theta_0|]$ as in the rest of the study. If a profile point improved the minimum, the reduced problem was re-fitted from that point and the profile repeated (at most three rounds). Only points within the bounds of θ_j were used. As in Raue et al. (2009), θ_j is *identifiable* if $\Delta\chi^2$ exceeds 3.84 (95%, one degree of freedom) on both sides within the bounds, *practically non-identifiable* if on one side only, and *flat* otherwise. Because several problems have recorded noise levels that do not match the scatter of the data, a robust variant scales the threshold by $\max(1, \chi^2_{\min}/\text{d.f.})$, which is equivalent to rescaling the noise so that the fit has $\chi^2/\text{d.f.} = 1$.
3. *Condition of the formal theorem.* For each selected parameter, the following quantities, which appear in the theorem described below, were computed: the profiled information $I_j = 1/[(A^\top A)^{-1}]_{jj}$, where A is the sensitivity matrix of the residuals with respect to $\log|\theta|$ for $S \cup \Theta_C$ at $\hat{\theta}$; the residual gradient $\gamma = \|Pr_0\|$, the norm of the projection of the residual at $\hat{\theta}$ onto the range of A; and the nonlinear remainder $\|r(\hat{\theta}+\delta) - r_0 - A\delta\|$ at the computed profile points.

Five problems with expensive simulations (Bachmann, Chen, Isensee, Raimundez, Smith) were computed on a server (Raimundez on a 128-core server, the others on a 256-core server) and the other 26 on an 8-core workstation, so the profiles cover all 31 problems with an admissible subset.

Mathematical derivation of the identifiability condition

This subsection states the result that is verified in Lean (module `ProfileIdentifiability`, step 32 of Supplementary Table S4) and gives the complete argument, with the reasoning behind each step. All statements are deterministic inequalities in a real inner-product space: no probabilistic assumption is used in the proofs, and the statistical meaning enters only through the threshold c.

Setting

Let k be the number of free parameters of the reduced problem, the selected biological parameters S and the calibration parameters Θ_C, all other parameters being fixed at θ_0. The analysis uses the displacement $\delta \in \mathbb{R}^k$ of their logarithms, $\log|\theta| = \log|\hat{\theta}| + \delta$, from the optimum $\hat{\theta}$ of the reduced fit.

Let

$$r(\delta) = \Big(\frac{y_i(\hat{\theta}+\delta) - \tilde{y}_i}{\sigma_i}\Big)_{i=1,\dots,m} \in \mathbb{R}^m$$

be the vector of noise-scaled residuals over all measurements (y_i the model output, $\tilde{y}_i$ the measurement and σ_i its recorded noise level, all after the transformation of the observable), so that the weighted least-squares cost is $\chi^2(\delta) = \|r(\delta)\|^2$. Write $r_0 = r(0)$ and

$$\Delta\chi^2(\delta) = \|r(\delta)\|^2 - \|r_0\|^2,$$

the increase of the cost when the parameters move by δ from the optimum. Let $A = \partial r/\partial\delta\,(0) \in \mathbb{R}^{m\times k}$ be the sensitivity matrix of the residuals at the optimum and

$$\rho(\delta) = r(\delta) - r_0 - A\delta$$

the nonlinear remainder of the first-order expansion. Fix a selected parameter j and denote by δ_j the j-th coordinate of δ. $F \subseteq \mathbb{R}^k$ is the admissible set (the bounds $[|\theta_0|/10, 10|\theta_0|]$ of every free parameter), and d_-, $d_+ > 0$ are the distances from $\log|\hat{\theta}_j|$ to the lower and upper bound of θ_j.

The profile likelihood of θ_j is $\mathrm{PL}_j(t) = \min\{\Delta\chi^2(\delta) : \delta \in F,\ \delta_j = t\}$. By Wilks' theorem, $\{t : \mathrm{PL}_j(t) < c\}$ with $c = \chi^2_{1;0.95} = 3.84$ is the 95% profile-likelihood confidence interval of $\log|\theta_j| - \log|\hat{\theta}_j|$ (Raue et al., 2009); this is the only place where a statistical argument enters.

Definition 1 (Practical identifiability within the bounds)**.** θ_j is *practically identifiable* if there are $0 < t_- < d_-$ and $0 < t_+ < d_+$ such that $\Delta\chi^2(\delta) \geq c$ for every $\delta \in F$ with $\delta_j \leq -t_-$ or $\delta_j \geq t_+$.

Equivalently, $\mathrm{PL}_j(t) \geq c$ for $t \leq -t_-$ and for $t \geq t_+$: the profile exceeds the threshold on both sides before reaching the bounds, so the confidence interval is contained in $(-t_-, t_+)$ and is finite. This is the definition applied numerically to the computed profiles, and it is the predicate `PracticallyIdentifiable` in Lean.

Hypotheses

Hypothesis 1 (approximate stationarity)**.** There is $\gamma \geq 0$ with $|\langle r_0, A\delta\rangle| \leq \gamma\|A\delta\|$ for all δ.

At an exact unconstrained minimum the normal equations give $A^\top r_0 = 0$, so $\gamma = 0$. A numerical optimiser stops at a tolerance, and an optimum can lie on a bound, so the hypothesis allows a residual gradient. Proposition 2 shows how γ is computed.

Hypothesis 2 (profiled information)**.** There is $I_j > 0$ with $I_j\,\delta_j^2 \leq \|A\delta\|^2$ for all δ.

$\|A\delta\|^2$ is the linear prediction of the increase of the cost. The hypothesis says that a displacement δ_j of the parameter of interest cannot be hidden by moving the other free parameters: whatever they do, the outputs change by at least $\sqrt{I_j}\,|\delta_j|$ in units of noise. Lemma 2 gives the best such constant.

Hypothesis 3 (bounded nonlinearity)**.** There is $\varepsilon \geq 0$ with $\|\rho(\delta)\| \leq \varepsilon$ for all $\delta \in F$.

This is the only hypothesis that involves the model away from the optimum.

Step 1: a lower bound for the increase of the cost

Lemma 1 (`dchi2_lower_grad`)**.** *Under Hypotheses 1 and 3, for every $\delta \in F$ with $\|A\delta\| \geq \varepsilon$,*

$$\Delta\chi^2(\delta) \ \geq\ (\|A\delta\| - \varepsilon)^2 - 2\gamma\|A\delta\| - 2\|r_0\|\varepsilon.$$

Proof. Since $r(\delta) = r_0 + (A\delta + \rho)$, expanding the square gives the exact identity

$$\Delta\chi^2(\delta) = 2\langle r_0, A\delta\rangle + 2\langle r_0, \rho\rangle + \|A\delta + \rho\|^2.$$

Each term is bounded from below. (i) By Hypothesis 1, $2\langle r_0, A\delta\rangle \geq -2\gamma\|A\delta\|$. (ii) By the Cauchy–Schwarz inequality and Hypothesis 3, $2\langle r_0, \rho\rangle \geq -2\|r_0\|\,\|\rho\| \geq -2\|r_0\|\varepsilon$. (iii) By the triangle inequality, $\|A\delta\| = \|(A\delta + \rho) - \rho\| \leq \|A\delta + \rho\| + \|\rho\|$, so $\|A\delta + \rho\| \geq \|A\delta\| - \varepsilon \geq 0$, and squaring preserves the inequality between non-negative numbers. Adding (i)–(iii) gives the claim. □

Reasoning. The identity separates three effects. The quadratic term $\|A\delta + \rho\|^2$ is the information: how much the outputs move. The term $2\langle r_0, A\delta\rangle$ vanishes at an exact optimum and measures how far the optimiser is from it. The cross term $2\langle r_0, \rho\rangle$ couples the misfit of the model with its nonlinearity: when the model fits the data well ($\|r_0\|$ small) the nonlinearity matters little, but when the misfit is large, as in the problems whose recorded noise does not match the data, even a moderate nonlinearity can lower the profile. This is why $\|r_0\|$ appears in the result.

Step 2: the information that remains for one parameter

Lemma 2 (`info_of_inverse`, `info_of_matrix`)**.** *If $A^\top A$ is invertible, Hypothesis 2 holds with $I_j = 1/[(A^\top A)^{-1}]_{jj}$, and this constant is the largest possible.*

Proof. Let $w = (A^\top A)^{-1} e_j$, so that $A^\top A w = e_j$ and, for every δ, $\langle Aw, A\delta\rangle = \langle A^\top A w, \delta\rangle = \delta_j$. In particular $\|Aw\|^2 = w_j = [(A^\top A)^{-1}]_{jj} > 0$. By Cauchy–Schwarz, $\delta_j^2 = \langle Aw, A\delta\rangle^2 \leq \|Aw\|^2\|A\delta\|^2 = w_j\|A\delta\|^2$, which is Hypothesis 2 with $I_j = 1/w_j$. For $\delta = w$ equality holds, so no larger constant is possible. □

Reasoning. The extremal direction w is the displacement that moves θ_j by one unit while the other free parameters compensate as well as possible in the least-squares sense; I_j is the squared output change that still remains. Equivalently, if a_j is the column of A for θ_j and P_{-j} the orthogonal projection onto the span of the other columns, then $I_j = \|(\mathrm{I} - P_{-j})a_j\|^2$, the part of the effect of θ_j that no combination of the other parameters can reproduce. This second form follows directly from $\|A\delta\| = \|\delta_j a_j + A_{-j}\delta_{-j}\| \geq |\delta_j|\,\mathrm{dist}(a_j, \mathrm{span}\, A_{-j})$, and it also holds when two *other* columns are collinear (for example two calibration parameters with the same effect), in which case $A^\top A$ is singular but the information of θ_j is still well defined. I_j was computed in this form. This is the quantity that distinguishes identifiability from collinearity: the condition number and the VIF of normalised columns measure only the angle between columns, whereas I_j measures the size of the remaining effect in units of the noise, so a parameter strongly correlated with a scaling factor can still be identifiable when the data are precise (Brannmark, Fiedler, Raia), and an uncorrelated parameter cannot be identifiable when the recorded noise is much larger than its effect (Giordano, Perelson).

Step 3: the profile crosses the threshold

Theorem 1 (`dchi2_ge_of_coord_grad`)**.** *Under Hypotheses 1–3, let*

$$t^* = \frac{\varepsilon + \gamma + \sqrt{\gamma^2 + c + 2\|r_0\|\varepsilon + 2\gamma\varepsilon}}{\sqrt{I_j}}.$$

Then $\Delta\chi^2(\delta) \geq c$ for every $\delta \in F$ with $|\delta_j| \geq t^*$.

Proof. Write $a = \|A\delta\|$, $K = \gamma^2 + c + 2\|r_0\|\varepsilon + 2\gamma\varepsilon \geq 0$ and $s = \sqrt{K}$. By Hypothesis 2, $a \geq \sqrt{I_j}\,|\delta_j| \geq \sqrt{I_j}\,t^* = \varepsilon + \gamma + s$. In particular $a \geq \varepsilon$, so Lemma 1 applies. Put $x = a - \varepsilon \geq \gamma + s \geq 0$. Then $x - \gamma \geq s \geq 0$, hence $(x-\gamma)^2 \geq s^2 = K$, that is $x^2 - 2\gamma x \geq c + 2\|r_0\|\varepsilon + 2\gamma\varepsilon$. By Lemma 1,

$$\Delta\chi^2(\delta) \;\geq\; x^2 - 2\gamma(x + \varepsilon) - 2\|r_0\|\varepsilon \;=\; (x^2 - 2\gamma x) - 2\gamma\varepsilon - 2\|r_0\|\varepsilon \;\geq\; c.$$ □

Reasoning. The lower bound of Lemma 1 is a quadratic function of a that becomes positive and grows once a exceeds the losses caused by the gradient and the nonlinearity. t^* is obtained by solving that quadratic for the value of a at which the bound reaches c, and translating it to δ_j through the information: each unit of δ_j guarantees $\sqrt{I_j}$ units of output change.

Corollary 1 (`practically_identifiable_grad`, `confidence_coord_lt`)**.** *Under Hypotheses 1–3, with $c > 0$: (a) every $\delta \in F$ with $\Delta\chi^2(\delta) < c$ satisfies $|\delta_j| < t^*$, so the 95% profile-likelihood interval of $\log|\theta_j|$ is contained in $(-t^*, t^*)$; (b) if $t^* < d_-$ and $t^* < d_+$, θ_j is practically identifiable within the bounds.*

Proof. (a) is the contrapositive of Theorem 1. For (b), $t^* > 0$ because $c > 0$ and $I_j > 0$; take $t_- = t_+ = t^*$ in the definition: any $\delta \in F$ with $\delta_j \le -t^*$ or $\delta_j \ge t^*$ has $|\delta_j| \ge t^*$, so $\Delta\chi^2(\delta) \ge c$ by Theorem 1. □

Step 4: special cases and the computable constants

Proposition 1 (`tstar_linear`, `tstarGrad_zero`)**.** *With $\gamma = 0$ the half-width reduces to $t^* = (\varepsilon + \sqrt{c + 2\|r_0\|\varepsilon})/\sqrt{I_j}$, and with also $\varepsilon = 0$ to*

$$t^* = \sqrt{c/I_j} = \sqrt{c\,[(A^\top A)^{-1}]_{jj}},$$

the half-width of the Wald (linearised) confidence interval.

Proof. Direct substitution; with $I_j = 1/w_j$, $\sqrt{c}/\sqrt{1/w_j} = \sqrt{c\,w_j}$. □

Reasoning. For a model that is exactly linear in δ and an exact optimum, $\Delta\chi^2(\delta) = \|A\delta\|^2$ and its minimum over $\delta_j = t$ is exactly $I_j t^2$ (Lemma 2, equality case), so the profile is the parabola $I_j t^2$ and the Wald interval is exact: the condition $\sqrt{c/I_j} < d_\pm$ is then necessary and sufficient. The general theorem quantifies how much the nonlinearity (ε), the misfit ($\|r_0\|$) and the optimiser (γ) can widen that interval.

Proposition 2 (`stat_of_projection`)**.** *If p is the orthogonal projection of r_0 onto the range of A ($\langle r_0 - p, A\delta\rangle = 0$ for all δ), Hypothesis 1 holds with $\gamma = \|p\|$.*

Proof. $\langle r_0, A\delta\rangle = \langle p, A\delta\rangle$, and Cauchy–Schwarz gives $|\langle p, A\delta\rangle| \le \|p\|\,\|A\delta\|$. □

p was computed by linear least squares, $p = Ac^*$ with $c^* = \arg\min_c \|r_0 - Ac\|$.

Proposition 3 (`local_injective`: local structural identifiability)**.** *Let G be the map from the free parameters to the outputs and A its derivative at 0. If $\|A\delta\| \ge \sigma\|\delta\|$ for all δ (σ the smallest singular value) and $\|G(\delta) - G(0) - A\delta\| \le M\|\delta\|^2$ for $\|\delta\| \le r$, then $G(\delta) \ne G(0)$ for every $\delta \ne 0$ with $\|\delta\| \le r$ and $M\|\delta\| < \sigma$.*

Proof. If $G(\delta) = G(0)$, the remainder equals $-A\delta$, so $\sigma\|\delta\| \le \|A\delta\| \le M\|\delta\|^2$, that is $\sigma \le M\|\delta\|$, a contradiction. □

Reasoning. This statement does not involve the data: it says that, near the optimum, different parameter values produce different outputs, so the parameters are identifiable from noise-free outputs. Practical identifiability (Corollary 1) is stronger, because it accounts for the noise and asks that the outputs change by more than the noise allows.

Relation with sufficiency

Admissibility and practical identifiability are different properties, and neither implies the other. Admissibility asks whether re-estimating $S \cup \Theta_C$ reproduces the outputs of the full model after all parameters are perturbed; it holds, for example, when θ_j and a scaling factor s enter only through their product $s\theta_j$, because the product can absorb any change, and in that case $I_j = 0$ and θ_j is not identifiable (Fujita is an instance). Conversely, a parameter can be identifiable without being sufficient to reproduce the full model. The paper establishes the two properties separately: admissibility by the re-estimation test (Theorems 1 and 2 of the main text and the other steps of Supplementary Table S4), and identifiability by the profiles, with Corollary 1 as a sufficient condition.

What is proved and what is computed

Lemmas 1 and 2, Theorem 1, Corollary 1 and Propositions 1–3 are proved in Lean for arbitrary r, A, F and constants, using only the three standard axioms (`propext`, `Classical.choice`, `Quot.sound`); Supplementary Table S7 gives their formal statements (step 32). The projection form of I_j is the elementary inequality given after Lemma 2; the formal development uses the inverse-matrix form, and the theorem itself takes Hypothesis 2 as an assumption, so either form can be used to verify it. For each problem, A is computed by central finite differences (step 0.01 in $\log|\theta|$), γ by Proposition 2, I_j by projection and $\|r_0\|$ directly. The remainder ε is the only quantity that cannot be bounded over the admissible set with the methods of this study: it was evaluated at the computed profile points. Far from the optimum, the response of an ODE model saturates and its linearisation overestimates the change of the outputs, so a single bound over the whole range of a factor of 10 is very pessimistic. The application of Corollary 1 to a particular problem is therefore a numerical check of its hypotheses, as for the other guarantees of the paper (Supplementary Table S5).

Results

Supplementary Table S20 summarises the results for the 31 problems with an admissible subset, Supplementary Figure S7 shows the profile of every selected parameter in every problem, and Supplementary Table S21 gives, for every parameter, its confidence interval and the quantities of the formal condition (I_j, the Wald half-width, γ, ε and the distances to the bounds).

Profile likelihood. In 25 of the 31 problems all selected parameters were practically identifiable, and 65 of the 74 selected parameters were identifiable (72 of them could be evaluated). The exceptions are of three kinds. In Fujita a selected parameter is confounded with a measurement scaling factor: the columns of `reaction_6_k1` and `scaling_pS6_tot` have cosine 1.00, and neither selected parameter is identifiable. In Giordano and Perelson the recorded noise levels are far larger than the residuals ($\chi^2/n < 10^{-2}$), so no parameter change raises χ^2 by 3.84; in Okuonghae and Oliveira one parameter is identifiable and the profile of the other could not be simulated. In Smith, the optimum of `k8` lies at the lower bound of the study. With the robust threshold, 23 of the 31 problems keep all their parameters identifiable (60 of the 74 parameters): Raia (χ^2/d.f. = 1,791) and Chen (168) lose one or both, and so does the second parameter of Smith (1,894), because their recorded noise levels are too small for the scatter of the data.

Local check. A VIF above 10 within $S \cup \Theta_C$ flags a correlation between a selected parameter and a calibration parameter, but it does not by itself make the parameter non-identifiable: in Brannmark, Fiedler, Isensee, Rahman, Raia, Raimundez, Schwen and Weber the VIF exceeds 10, yet all their selected parameters are identifiable in the profiles, because the data are informative enough to separate them. Only in Fujita does the confounding reach the profile.

Condition of the theorem. The linear condition ($\varepsilon = 0$, $\gamma = 0$) agrees with the profile classification for all 72 parameters whose profile could be evaluated (Supplementary Table S20, column Wald): it holds for the 65 identifiable parameters and fails for the other seven. With the nonlinear remainder evaluated at the profile points, the hypotheses of the theorem hold for 24 parameters (column Lean), and never for

a parameter whose profile is not identifiable. The quantitative certificate is therefore conservative: the profiled information I_j explains which parameters are identifiable, and the nonlinearity of the models is what prevents a certificate over the whole range of the study.

Cost of the identifiability check. The last column of Supplementary Table S20 gives the computing time needed to determine the identifiability of the selected parameters of each problem: one fit of the reduced problem and the re-estimations at the 40 profile points of its two selected parameters (more for Bruno, Lucarelli and Zhao). Because only the selected and calibration parameters are free, each re-estimation is small. The median over the 31 problems was 34 s of computing time; 20 problems needed less than a minute and 24 less than ten minutes. The check took hours only where a single simulation is slow (Chen, 44 s per simulation on the server: 15 h) or where many calibration parameters are re-estimated at every point (Raimundez, 79: 4.0 h; Isensee: 4.3 h). The profile points are independent, so on many cores the elapsed time is a fraction of these values (Chen took about 3 h on the server). These include problems in which the profile likelihood of the full model takes hours (Bachmann, 40 h of computing time) or was not run (Chen, Raimundez; Supplementary Note *Profile likelihood*), so the identifiability of the selected subset can be verified at a fraction of that cost. The times were measured on three machines (an 8-core workstation and shared servers with 128 and 256 cores) and are observed times, not benchmarks.

Subsets that are both sufficient and identifiable

Admissibility and practical identifiability are different properties (see *Relation with sufficiency* above), so a subset that can be used to reduce the model and whose parameters are also determined by the data has to satisfy both. For a problem with selected subset S and calibration parameters Θ_C four nested conditions are distinguished, each checked for every parameter of S:

1. *Reducible (admissible).* Re-estimating $S \cup \Theta_C$ with the other parameters at θ_0 reproduces the full model: the median e_{fit} over the 15 scenarios is at most $\tau_e = \sqrt{1 - c^2} = 0.436$ (Theorems 1 and 2 of the main text and steps 1–22 of Supplementary Table S4).
2. *Identifiable by the profiles.* For every θ_j, $j \in S$, the profile of the reduced problem exceeds $\Delta\chi^2 = 3.84$ on both sides within the bounds (the predicate `PracticallyIdentifiable`).
3. *Linear condition.* $\sqrt{c/I_j} < \min(d_-, d_+)$ for every $j \in S$ (Proposition 1 with $\gamma = \varepsilon = 0$): the Wald interval is inside the bounds.
4. *Hypotheses of the theorem.* $t_j^* < \min(d_-, d_+)$ for every $j \in S$, with I_j, γ (Proposition 2) and $\|r_0\|$ computed at the optimum and ε the largest nonlinear remainder at the computed profile points; by Corollary 1 every parameter of S is then practically identifiable. Because ε is evaluated at points rather than bounded over the whole admissible set, this is a numerical check of the hypotheses, not a proof for the particular model.

Of the 34 problems, 31 satisfy condition 1. Of these, 25 also satisfy condition 2 for all their selected parameters (23 with the robust threshold); the six exceptions are listed under *Profile likelihood* above. Condition 3 holds for all selected parameters in 27 problems: the 25 of condition 2 and Okuonghae and Oliveira, in which the profile of one parameter could not be simulated, so that the linear condition cannot be compared with it. Condition 4 holds for all selected parameters in seven problems, Armistead ($|S| = 2$), Blasi (2), Boehm (2), Bruno (5), Rahman (2), Sneyd (2) and Zheng (2), none of which has calibration parameters. The models of these seven problems are also among the 28 verified in Lean (Supplementary Note *Verified benchmark models*), so for them every link is in place, with different status: the hypotheses on the dynamics are proved in Lean (at the nominal parameters and in the base condition), whereas the admissibility of the subset and the hypotheses of the identifiability theorem are checked numerically at the computed points, not certified. In the other 24 admissible problems the subset is reducible and, in 18 of them, identifiable by the profiles, but condition 4 fails for at least one parameter: in Fujita, Giordano, Perelson and Smith because the information I_j is already insufficient (condition 3 fails), and in the others because of the nonlinear remainder ε (in Raimundez, for example,

both parameters satisfy condition 3 but the remainder reaches 215 and 241 noise units, Supplementary Table S21).

Scope

The result is practical identifiability of the reduced problem: local, within the bounds of the study, with the non-selected parameters at θ_0 and with the noise model of the benchmark. It is not identifiability of the full model, in which all parameters are free: in the profile likelihood of the full model only 20 of the 52 selected parameters evaluated were identifiable (Supplementary Note *Profile likelihood*). That contrast is consistent with the purpose of the method. The data do not determine all the parameters of these models, but, with the other parameters at their published values, a few selected parameters are both sufficient to reproduce the outputs and, in most problems, determined by the data. Identifiability does not imply that the estimated values are the true biological values: they depend on the values at which the other parameters are fixed.

Supplementary Table S20. Identifiability of the selected parameters in the reduced problem ($S \cup \Theta_C$ free, other parameters at θ_0, all measurements). $|S|$, $|\Theta_C|$: selected biological and calibration parameters; χ^2/n: at the optimum of the reduced problem; VIF: largest variance inflation factor of a selected parameter among the columns of $J_{S\cup\Theta_C}$ (unit-norm, uncentred) at θ_0. Profile: class of each selected parameter by profile likelihood with $\Delta\chi^2 = 3.84$ within $[|\theta_0|/10, 10|\theta_0|]$ (I, identifiable, the profile exceeds the threshold on both sides; N, practically non-identifiable, on one side; F, flat; –, the profile could not be simulated); robust: the same with the threshold scaled by $\max(1, \chi^2_{\min}/\text{d.f.})$. Wald: parameters whose linear (Wald) interval $\sqrt{3.84\,[(J^\top J)^{-1}]_{jj}}$ fits within the bounds on both sides; Lean: parameters for which the hypotheses of `practically_identifiable_grad` hold at the computed profile points up to the first crossing. Time: computing time needed to determine the identifiability of the selected parameters, that is, the fit of the reduced problem plus the re-estimations at all profile points, summed over the fits (CPU time; the points were computed in parallel, so the wall-clock time is shorter); $^{\text{S}}$, computed on a shared server (128 or 256 cores), otherwise on an 8-core workstation. For Isensee and Smith the profile was repeated after a better optimum was found, and only the last round is counted; for Chen, two profile points (one of them outside the bounds) did not finish and are not counted.

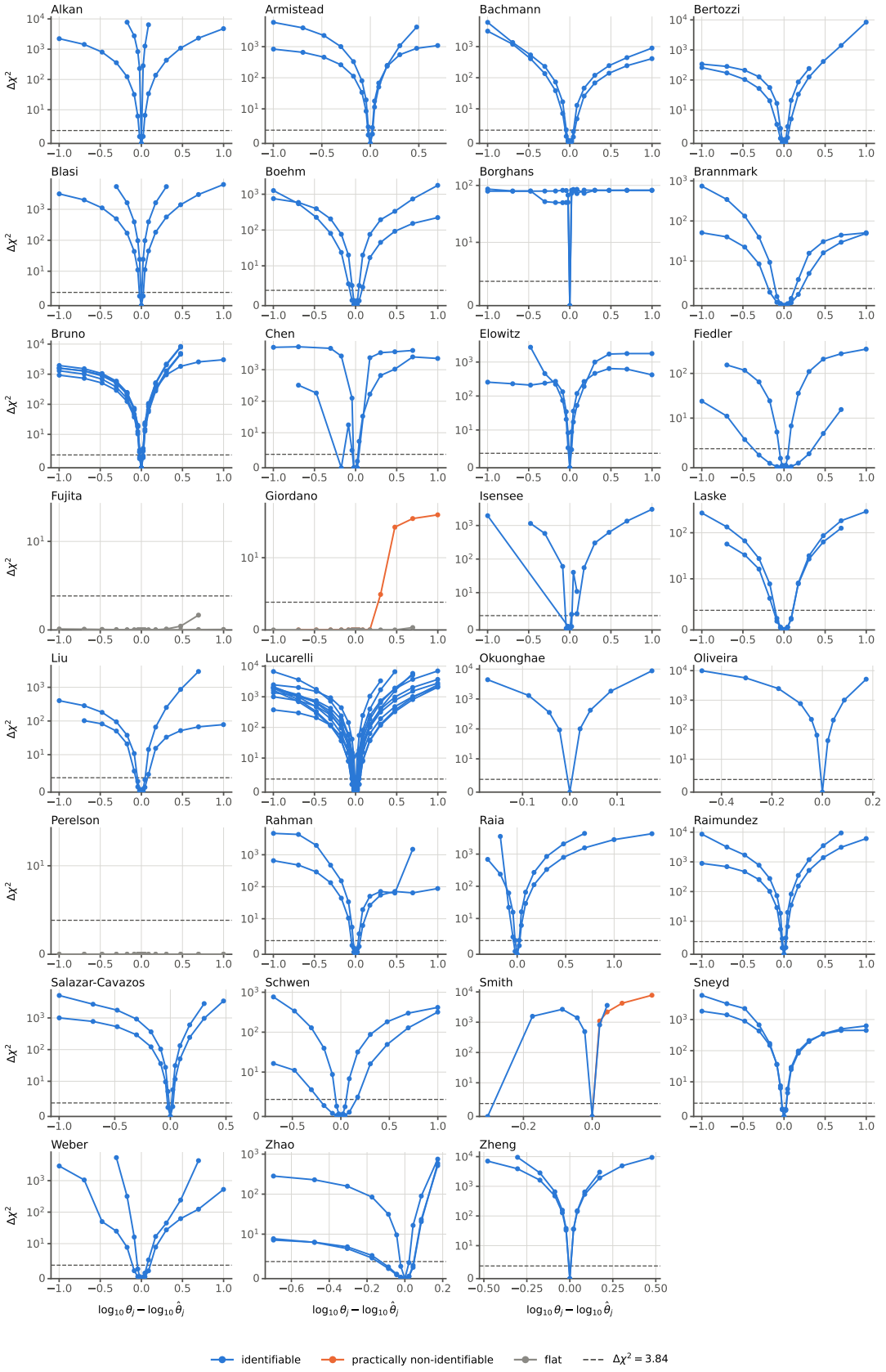
Problem	$\lvert S\rvert$	$\lvert\Theta_C\rvert$	χ^2/n	VIF	Profile	Robust	Wald	Lean	Time
Alkan	2	10	1.96	1.1	I I	I I	2/2	1/2	21.6 min
Armistead	2	0	1.04	1.1	I I	I I	2/2	2/2	<1 s
Bachmann	2	74	1.16	6.1	I I	I I	2/2	0/2	35.2 min$^{\text{S}}$
Bertozzi	2	0	1.10	5.9	I I	I I	2/2	0/2	1 s
Blasi	2	0	6.46	2.3	I I	I I	2/2	2/2	5 s
Boehm	2	0	1.04	1.4	I I	I I	2/2	2/2	4 s
Borghans	2	2	0.97	2.6	I I	I I	2/2	0/2	34 s
Brannmark	2	3	1.13	127.2	I I	I I	2/2	0/2	58 s
Bruno	5	0	0.90	1.3	I I I I I	I I I I I	5/5	5/5	19 s
Chen	2	3	168	1.1	I I	N I	2/2	0/2	14.8 h$^{\text{S}}$
Elowitz	2	2	1.07	6.0	I I	I I	2/2	0/2	22 s
Fiedler	2	8	1.04	47.2	I I	I I	2/2	0/2	35 s
Fujita	2	3	5.48	2.5e+08	F F	F F	0/2	0/2	1.1 min
Giordano	2	0	0.00	4.1	N F	N F	0/2	0/2	6 s
Isensee	2	11	1.02	41.6	I I	I I	2/2	0/2	4.3 h$^{\text{S}}$
Laske	2	2	0.18	2.7	I I	I I	2/2	1/2	1.2 min
Liu	2	0	1.02	1.3	I I	I I	2/2	0/2	7 s
Lucarelli	10	0	1.22	9.9	I I I I I I I I I I	I I I I I I I I I I	10/10	4/10	54.9 min
Okuonghae	2	0	1,413	1.3	– I	– N	2/2	0/2	4 s
Oliveira	2	0	182	1.0	– I	– I	2/2	0/2	6 s
Perelson	2	0	0.00	1.2	F F	F F	0/2	0/2	<1 s
Rahman	2	0	0.00	39.0	I I	I I	2/2	2/2	1 s

Problem	$\lvert S\rvert$	$\lvert\Theta_C\rvert$	χ^2/n	VIF	Profile	Robust	Wald	Lean	Time
Raia	2	5	1,791	2.3e+04	I I	F F	2/2	0/2	1.4 min
Raimundez	2	79	4.85	75.5	I I	I I	2/2	0/2	4.0 h^{S}
Salazar-Cavazos	2	0	44.96	5.7	I I	I I	2/2	0/2	40 s
Schwen	2	15	0.96	39.1	I I	I I	2/2	0/2	6.2 min
Smith	2	9	1,894	29.3	N I	N F	1/2	0/2	48.1 min^{S}
Sneyd	2	0	1.01	1.0	I I	I I	2/2	2/2	20 s
Weber	2	5	0.94	631.0	I I	I I	2/2	0/2	37 s
Zhao	3	0	0.27	1.0	I I I	I I I	3/3	1/3	9 s
Zheng	2	0	1.03	1.1	I I	I I	2/2	2/2	9 s

Supplementary Table S21. Identifiability of every selected parameter in the reduced problem. Class: profile-likelihood class with the threshold 3.84 / with the robust threshold (I, identifiable; N, practically non-identifiable; F, flat; –, the profile could not be simulated). CI: 95% profile-likelihood interval of $\log_{10}\theta_j - \log_{10}\hat{\theta}_j$, interpolated between profile points (∞, the profile does not cross the threshold on that side within the bounds). I_j: profiled information $1/[(A^\top A)^{-1}]_{jj}$ (computed by projection); Wald: half-width $\sqrt{3.84/I_j}$ in $\log_{10}$ units; $\gamma = \|Pr_0\|$; ε: largest nonlinear remainder at the profile points; d_-, d_+: distances from $\log_{10}\hat{\theta}_j$ to $\log_{10}(|\theta_0|/10)$ and $\log_{10}(10|\theta_0|)$; Lean: hypotheses of `practically_identifiable_grad` satisfied at the profile points up to the first crossing.

Problem	Parameter	Class	CI ($\log_{10}$)	I_j	Wald	γ	ε	d_-	d_+	Lean
Alkan	kt	I/I	$[-2.8\times10^{-3}, 2.5\times10^{-3}]$	1.0×10^{5}	2.7×10^{-3}	4.7×10^{-4}	41.9	1.00	1.00	✓
	k_rep_nhej	I/I	[−0.03, 0.03]	903.6	0.03	4.7×10^{-4}	50.7	1.00	1.00	
Armistead	k_d	I/I	[−0.02, 0.02]	1870.3	0.02	3.3×10^{-6}	46.5	1.00	1.00	✓
	k00	I/I	[−0.03, 0.03]	1031.7	0.03	3.3×10^{-6}	46.3	1.00	1.00	✓
Bachmann	CISRNADelay	I/I	[−0.04, 0.05]	364.7	0.04	0.59	158.9	1.00	1.00	
	JAK2ActEpo	I/I	[−0.06, 0.06]	194.0	0.06	0.59	30.1	1.00	1.00	
Bertozzi	R0_NY	I/I	[−0.04, 0.04]	442.3	0.04	1.5×10^{-4}	82.5	1.00	1.00	
	I0_NY	I/I	[−0.07, 0.06]	150.8	0.07	1.5×10^{-4}	103.1	1.00	1.00	
Blasi	a_basal	I/I	$[-8.7\times10^{-3}, 8.6\times10^{-3}]$	9745.2	8.6×10^{-3}	3.3×10^{-4}	52.8	1.00	1.00	✓
	a_k16_k12k16	I/I	[−0.03, 0.03]	1136.9	0.03	3.3×10^{-4}	41.3	1.00	1.00	✓
Boehm	k_imp_hetero	I/I	[−0.04, 0.04]	489.3	0.04	1.1×10^{-3}	42.5	1.00	1.00	✓
	k_phos	I/I	[−0.07, 0.08]	121.0	0.08	1.1×10^{-3}	18.2	1.00	1.00	✓
Borghans	Kz	I/I	$[-5.2\times10^{-3}, 5.0\times10^{-3}]$	1.3×10^{5}	2.4×10^{-3}	1.53	2393.2	1.00	1.00	
	K_par	I/I	$[-6.0\times10^{-3}, 4.7\times10^{-3}]$	1.1×10^{5}	2.5×10^{-3}	1.53	2417.3	1.00	1.00	
Brannmark	k1f	I/I	[−0.20, 0.22]	19.8	0.19	0.02	17.3	1.00	1.00	
	km2	I/I	[−0.11, 0.14]	56.3	0.11	0.02	38.2	1.00	1.00	
Bruno	kb1	I/I	[−0.02, 0.02]	2009.9	0.02	1.1×10^{-5}	104.7	1.00	1.00	✓
	init_bcar2	I/I	[−0.02, 0.02]	2215.7	0.02	1.1×10^{-5}	68.0	1.00	1.00	✓
	init_zea_1	I/I	[−0.02, 0.02]	1977.5	0.02	1.1×10^{-5}	62.3	1.00	1.00	✓
	init_bcar1	I/I	[−0.02, 0.02]	1358.7	0.02	1.1×10^{-5}	54.7	1.00	1.00	✓
	init_bcry_1	I/I	[−0.03, 0.02]	1144.5	0.03	1.1×10^{-5}	47.4	1.00	1.00	✓
Chen	k106	I/N	[−0.04, 0.03]	402.9	0.04	13.01	58.4	0.78	1.22	
	k44	I/I	[−0.03, 0.03]	2.6×10^{4}	5.3×10^{-3}	13.01	418.4	1.05	0.95	
Elowitz	n_Hill	I/I	[−0.02, 0.02]	1951.7	0.02	0.05	170.4	1.00	1.00	
	tau_mRNA	I/I	[−0.01, 0.01]	3646.0	0.01	0.05	242.0	1.00	1.00	
Fiedler	tau2	I/I	[−0.06, 0.06]	186.8	0.06	9.6×10^{-4}	53.6	0.97	1.03	
	k6	I/I	[−0.38, 0.36]	5.6	0.36	9.6×10^{-4}	6.7	1.22	0.78	
Fujita	init_AKT	F/F	[$-\infty$, ∞]	3.2×10^{-4}	47.81	0.34	0.1	1.00	1.00	
	reaction_6_k1	F/F	[$-\infty$, ∞]	0.1	2.56	0.34	1.0	1.00	1.00	
Giordano	alpha_4	N/N	[$-\infty$, 0.29]	7.3×10^{-4}	31.58	3.3×10^{-8}	6.7	1.00	1.00	
	alpha_28	F/F	[$-\infty$, ∞]	1.2×10^{-5}	250.77	3.3×10^{-8}	0.6	1.02	0.98	
Isensee	kf_RIIp_C_2__RII_C_2	I/I	[−0.04, 0.03]	271.7	0.05	3.0×10^{-3}	110.0	1.01	0.99	
	xi_kf_RII_C_2__RII_2	I/I	[−0.05, 0.02]	36.2	0.14	3.0×10^{-3}	21.3	0.97	1.03	
Laske	k_imp	I/I	[−0.13, 0.11]	48.8	0.12	5.0×10^{-4}	65.6	0.99	1.01	
	k_syn_R_M	I/I	[−0.11, 0.11]	57.3	0.11	5.0×10^{-4}	8.5	1.04	0.96	✓

Problem	Parameter	Class	CI ($\log_{10}$)	I_j	Wald	γ	ε	d_-	d_+	Lean
Liu	kappa	I/I	$[-0.05, 0.05]$	307.1	0.05	3.3×10^{-4}	34.6	1.00	1.00	
	phi	I/I	$[-0.07, 0.08]$	132.7	0.07	3.3×10^{-4}	23.4	1.00	1.00	
Lucarelli	init_Rec	I/I	$[-0.01, 0.01]$	4870.0	0.01	1.5×10^{-3}	158.7	1.00	1.00	
	S2tot	I/I	$[-0.02, 0.02]$	1654.9	0.02	1.5×10^{-3}	56.8	1.00	1.00	
	S3tot	I/I	$[-0.03, 0.03]$	951.5	0.03	1.5×10^{-3}	42.1	1.00	1.00	✓
	S_dephosphos	I/I	$[-0.06, 0.06]$	230.7	0.06	1.5×10^{-3}	39.3	1.00	1.00	
	S_phos	I/I	$[-0.06, 0.06]$	229.8	0.06	1.5×10^{-3}	31.0	1.00	1.00	
	geneF_turn	I/I	$[-0.02, 0.02]$	1510.9	0.02	1.5×10^{-3}	47.4	1.00	1.00	✓
	k_234	I/I	$[-0.03, 0.03]$	628.1	0.03	1.5×10^{-3}	26.3	1.00	1.00	✓
	geneK_turn	I/I	$[-0.03, 0.03]$	720.8	0.03	1.5×10^{-3}	49.2	1.00	1.00	
	S4tot	I/I	$[-0.04, 0.04]$	453.3	0.04	1.5×10^{-3}	44.7	1.00	1.00	
	geneH_turn	I/I	$[-0.04, 0.04]$	390.9	0.04	1.5×10^{-3}	3.7	1.00	1.00	✓
Okuonghae	transmission_rate_effective	–/–	–	1.1×10^{8}	8.1×10^{-5}	0.04	0	1.01	0.99	
	gamma_i	I/N	$[-4.4 \times 10^{-3}, 4.2 \times 10^{-3}]$	3.9×10^{4}	4.3×10^{-3}	0.04	18.3	0.28	1.72	
Oliveira	beta_0	–/–	–	4.9×10^{8}	3.9×10^{-5}	1.17	0	0.94	1.06	
	h_hosp_rate	I/I	$[-5.3 \times 10^{-3}, 6.5 \times 10^{-3}]$	2.2×10^{4}	5.8×10^{-3}	1.17	64.3	1.22	0.78	
Perelson	c	F/F	$[-\infty, \infty]$	3.7×10^{-10}	4.4×10^{4}	2.0×10^{-6}	1.6×10^{-5}	1.00	1.00	
	delta	F/F	$[-\infty, \infty]$	1.9×10^{-10}	6.1×10^{4}	2.0×10^{-6}	3.2×10^{-5}	1.00	1.00	
Rahman	infected_moderate_transmission_rate	I/I	$[-0.03, 0.04]$	671.3	0.03	6.0×10^{-5}	104.5	1.00	1.00	✓
	infected_normal_worsen_rate	I/I	$[-0.05, 0.06]$	235.8	0.06	6.0×10^{-5}	97.2	1.00	1.00	✓
Raia	SOCS3mRNA_production	I/F	$[-0.04, 0.03]$	340.4	0.05	0.42	124.5	0.23	1.77	
	init_Rec_i	I/F	$[-0.02, 0.02]$	227.7	0.06	0.42	161.8	0.34	1.66	
Raimundez	kdeg_membran__MKN1	I/I	$[-0.02, 0.02]$	1813.6	0.02	5.6×10^{-3}	241.0	1.00	1.00	
	ka_MAPK__MKN1	I/I	$[-0.03, 0.03]$	749.7	0.03	5.6×10^{-3}	215.2	1.00	1.00	
Salazar-Cavazos	ratio_kpkd_Y1068__FREE	I/I	$[-0.02, 0.02]$	2936.9	0.02	1.5×10^{-4}	181.1	1.00	1.00	
	ratio_kpkd_YN__FREE	I/I	$[-0.03, 0.03]$	1070.4	0.03	1.5×10^{-4}	133.4	1.00	1.00	
Schwen	ini_R1	I/I	$[-0.23, 0.16]$	21.8	0.18	11.07	14.8	0.95	1.05	
	ini_R2fold	I/I	$[-0.06, 0.06]$	225.1	0.06	11.07	22.7	1.00	1.00	
Smith	k8	N/N	$[-\infty, 1.3 \times 10^{-3}]$	1.8×10^{4}	6.4×10^{-3}	82.20	12.8	1.4×10^{-5}	2.00	
	kminus7b	I/F	$[-1.9 \times 10^{-3}, 1.5 \times 10^{-3}]$	3.5×10^{5}	1.4×10^{-3}	82.20	336.7	0.43	1.57	
Sneyd	l4	I/I	$[-0.03, 0.03]$	841.9	0.03	4.6×10^{-3}	57.5	1.00	1.00	✓
	l2	I/I	$[-0.03, 0.03]$	762.7	0.03	4.6×10^{-3}	99.5	1.00	1.00	✓
Weber	s12	I/I	$[-0.11, 0.11]$	54.4	0.12	8.9×10^{-4}	185.0	1.00	1.00	
	a21	I/I	$[-0.05, 0.07]$	195.1	0.06	8.9×10^{-4}	137.1	1.00	1.00	
Zhao	R_Stage_I_China	I/I	$[-0.15, 0.05]$	152.1	0.07	6.2×10^{-4}	19.7	0.85	1.15	
	R_Stage_II_Wuhan	I/I	$[-0.03, 0.02]$	1257.7	0.02	6.2×10^{-4}	40.0	0.99	1.01	✓
	gamma_1_Stage_I_Hubei	I/I	$[-0.13, 0.05]$	171.5	0.06	6.2×10^{-4}	17.9	0.95	1.05	
Zheng	k01_02	I/I	$[-7.4 \times 10^{-3}, 7.1 \times 10^{-3}]$	1.4×10^{4}	7.2×10^{-3}	2.8×10^{-5}	50.1	1.00	1.00	✓
	k13_12	I/I	$[-6.9 \times 10^{-3}, 7.1 \times 10^{-3}]$	1.5×10^{4}	7.0×10^{-3}	2.8×10^{-5}	42.2	1.00	1.00	✓



Supplementary Figure S7. Profile likelihood of the selected parameters in the reduced problem, for the 31 problems with an admissible subset. Each line is one selected parameter: $\Delta\chi^2$ (symmetric logarithmic scale) against the displacement of $\log_{10}\theta_j$ from the optimum of the reduced problem, with $S \cup \Theta_C$ re-estimated at each point and the other parameters at θ_0. Only points within the bounds $[|\theta_0|/10, 10|\theta_0|]$ and whose simulation succeeded are drawn. Blue, identifiable (the profile exceeds the threshold on both sides); orange, practically non-identifiable (one side); grey, flat. Dashed line, $\Delta\chi^2 = 3.84$. In Okuonghae and Oliveira the profile of one parameter could not be simulated and is not shown.

Supplementary Note: Nested validation without the θ_0 anchor

Motivation

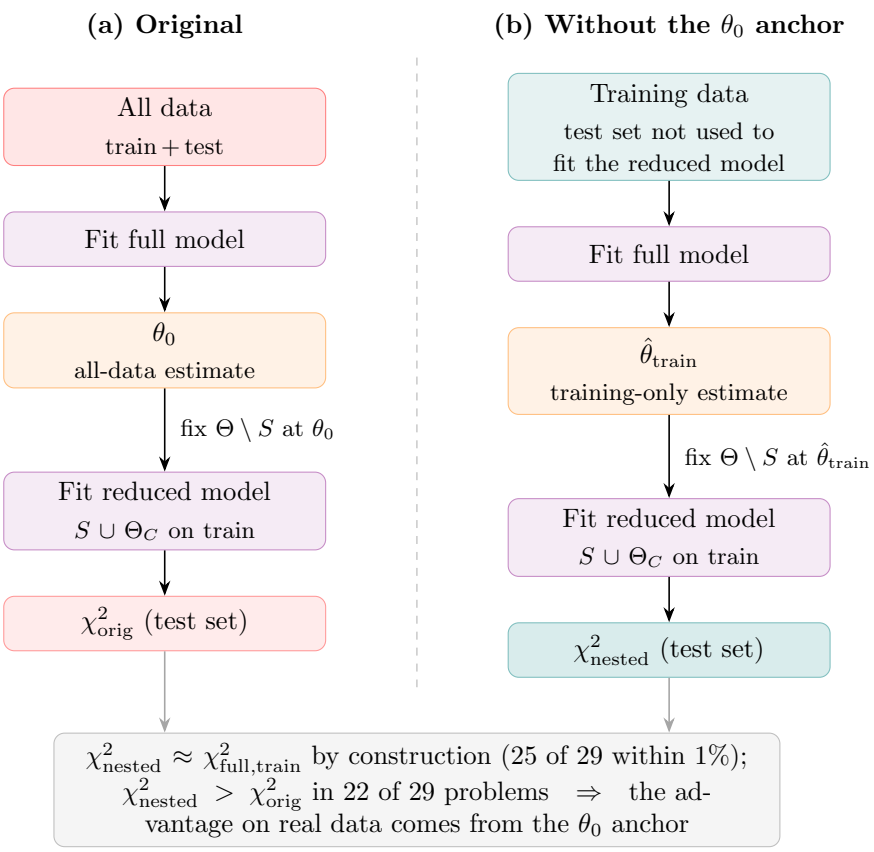
In the external validation the reduced model keeps the non-selected parameters at θ_0, the published estimate, which was obtained from *all* experimental data, including the test experiments. The reduced model may therefore inherit information about the test data through the parameters it keeps fixed. This note quantifies that effect.

Design

For each problem, a *nested* validation was run in which no value used by the reduced model has seen the test data:

1. the full model was re-estimated on the training data only, giving $\hat{\theta}_{\text{train}}$ (the full-model baseline of the external validation);
2. the reduced model was then re-estimated on the training data with the non-selected parameters fixed at $\hat{\theta}_{\text{train}}$ instead of θ_0, and with search bounds derived from $\hat{\theta}_{\text{train}}$;
3. the test χ^2/n of this nested reduced model was compared with that of the full training fit and with that of the original, θ_0-anchored reduced model.

Supplementary Figure S8 illustrates both protocols. The protocol removes the leakage through the values of the fixed parameters; it does not repeat the whole procedure on the training data: the subset S and the sensitivity matrix were not recomputed on the training data, and the full training fit starts from θ_0. It is therefore an analysis of the effect of the θ_0 anchor, not a fully leakage-free validation of the selection.



Supplementary Figure S8. The two external-validation protocols. (a) In the original protocol the non-selected parameters of the reduced model are fixed at θ_0, the published estimate obtained from *all* data including the test set. **(b)** In the protocol without the θ_0 anchor the full model is first re-estimated on the training data only, giving $\hat{\theta}_{\text{train}}$, which then serves both as the value of the fixed parameters and as the reference for the optimisation bounds; the test set is never used to construct the reduced model. The box states the outcome: the nested reduced model reproduces the full training fit, as it must because its re-estimation starts at the training optimum, and it predicts worse than the original reduced model in 22 of 29 problems.

Twenty-nine problems were evaluated. Bertozzi (selected parameters active only in the test conditions) and Blasi (single time point, no split) were excluded. For Chen, whose full re-estimation with 155 free parameters did not finish in serial form, the columns of the finite-difference Jacobian were evaluated in parallel and the integrator tolerances were $10^{-6}/10^{-8}$; its θ_0 and original values agree with those of the main analysis.

Results

Supplementary Table S22 gives the results and Supplementary Figure S9 compares them graphically.

The nested reduced model coincides with the full training fit. In 19 of the 29 problems the two test values agree to better than 0.1% and in 25 to better than 1%. This agreement is expected by construction and is not evidence about leakage: the re-estimation of the reduced model starts at $\hat{\theta}_{\text{train}}$, which already minimises the training error, so the optimiser has nothing to improve and returns essentially the full training fit. The exceptions are Schwen (+248%, a worse local minimum of the two re-estimated parameters), Raia (−5.1%), Chen (−3.7%) and Brannmark (+2.4%).

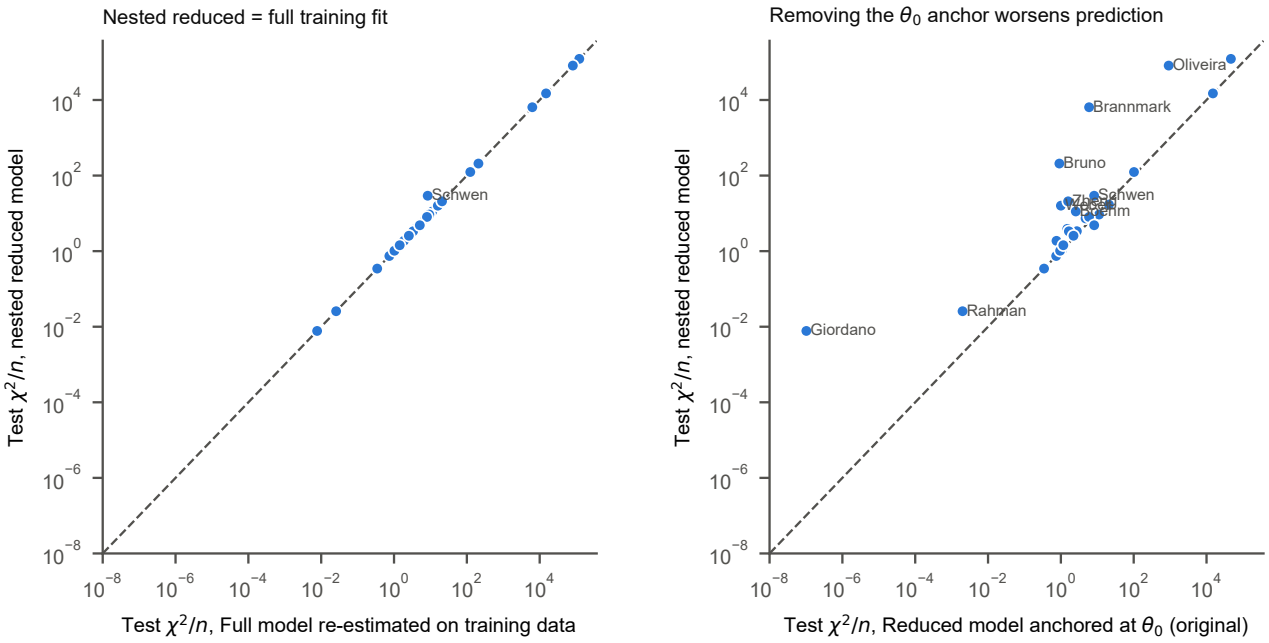
Removing the θ_0 anchor removes the advantage of the reduced model. Compared with the original reduced model, the nested reduced model predicted worse in 22 of the 29 problems (21 by more than 5%, the criterion of the main text) and more than four times worse in eight (for example Brannmark 5.96 versus 6.4×10^3, Bruno 0.91 versus 208, Weber 1.00 versus 16.0, Zheng 1.58 versus 20.8). The

better prediction of the reduced model on real data (main text, *Prediction of held-out experiments*) therefore reflects the information in the published values at which the other parameters are fixed. This does not affect the admissibility of the subsets, which is a property of the model, not of the data, and it is why the comparison of prediction without leakage is made on synthetic data, where θ_0 carries no information about the truth (Supplementary Note *Synthetic prediction experiment*).

Supplementary Table S22. Nested validation without the θ_0 anchor (29 problems). χ^2_{orig}: reduced model with the non-selected parameters at θ_0; $\chi^2_{\text{full,train}}$: full model re-estimated on the training data; χ^2_{nested}: reduced model with the non-selected parameters at $\hat{\theta}_{\text{train}}$; θ_0: published estimate. $\Delta\%$: nested relative to full; last column: nested over original (above 1, the nested reduced model predicts worse).

Problem	χ^2_{orig}	$\chi^2_{\text{full,train}}$	χ^2_{nested}	θ_0	$\Delta\%$	nested/orig
Alkan	2.75	3.37	3.37	2.47	+0.0	1.23
Armistead	1.22	1.37	1.37	1.16	-0.0	1.13
Bachmann	4.77	7.35	7.36	1.05	+0.1	1.54
Boehm	2.57	11.2	11.2	0.341	+0.0	4.37
Borghans	0.743	0.744	0.744	0.74	+0.0	1.00
Brannmark	5.96	6,277	6,431	0.938	+2.4	1079.76
Bruno	0.909	208	208	0.909	+0.0	228.69
Chen†	20.9	17.7	17.1	15.2	-3.7	0.82
Elowitz	1.5	3.89	3.89	1.38	-0.0	2.59
Fiedler	11.2	9.32	9.3	1.22	-0.2	0.83
Fujita	5.94	8.03	8.03	5.63	+0.0	1.35
Giordano	1.0×10^{-7}	0.00773	0.00773	7.4×10^{-9}	+0.0	75542.33
Isensee	1.08	1.08	1.08	1.04	-0.0	0.99
Laske	0.758	1.87	1.87	0.445	+0.0	2.47
Liu	0.954	1.02	1.02	0.942	-0.0	1.07
Lucarelli	1.66	3.34	3.35	1.02	+0.1	2.02
Okuonghae	4.7×10^4	1.2×10^5	1.2×10^5	2.9×10^4	-0.0	2.62
Oliveira	912	8.1×10^4	8.1×10^4	4.6×10^5	+0.0	89.16
Perelson	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}	+0.0	1.00
Rahman	0.00199	0.0257	0.0257	0.00148	+0.0	12.90
Raia	8.22	5.1	4.84	0.417	-5.1	0.59
Raimundez	1.17	1.43	1.43	0.98	+0.1	1.23
Salazar-Cavazos	102	124	124	96.9	+0.0	1.22
Schwen	8.2	8.4	29.3	0.768	+248.4	3.57
Smith	1.5×10^4	1.5×10^4	1.5×10^4	8.3×10^4	-0.3	1.00
Sneyd	2.22	2.56	2.56	1.37	-0.0	1.15
Weber	1	16	16	0.993	+0.1	15.96
Zhao	0.345	0.345	0.345	0.345	+0.0	1.00
Zheng	1.58	20.8	20.8	1.51	+0.0	13.16

†Chen: full-model re-estimation with the finite-difference Jacobian evaluated in parallel and integrator tolerances $10^{-6}/10^{-8}$.



Supplementary Figure S9. Nested validation without the θ_0 anchor (partial leakage control). Test χ^2/n of the nested reduced model (non-selected parameters at $\hat{\theta}_{\text{train}}$) against **a**, the full model re-estimated on the training data, with which it coincides by construction, and **b**, the original reduced model anchored at θ_0; points above the diagonal predict worse without the anchor. Dashed line, $y = x$; problems that differ by more than a factor of 3 are labelled; Perelson ($\chi^2/n < 10^{-12}$) lies outside the axes.

Supplementary Note: Pivoted QR selection

Motivation

This note examines whether a simpler, purely sensitivity-based selection criterion achieves an external validation accuracy comparable to that of the proposed method, which additionally tests admissibility and applies backward elimination. The proposed method was compared with column-pivoted QR selection (Vélez-Reyes and Verghese, 1995), a well-established subset-selection method that selects parameters by their linear independence in the sensitivity matrix, without re-estimation or test-set evaluation. The collinearity index of Brun *et al.* (Brun et al., 2001) is a different criterion and is compared separately in Supplementary Note *Brun and Morris selections*.

Method: QR-pivoted selection

Sensitivity matrix. For each problem the normalised sensitivity matrix $\widetilde{J} \in \mathbb{R}^{n_{\text{obs}} \times n_{\text{bio}}}$ is computed at θ_0 on the training data using forward finite differences. Column j is

$$\widetilde{J}_{\cdot j} = \frac{\partial z/\partial \theta_j}{\|\partial z/\partial \theta_j\|_2}, \qquad \frac{\partial z_i}{\partial \theta_j} \approx \frac{z_i(\theta_0 + \varepsilon_j e_j) - z_i(\theta_0)}{\varepsilon_j},$$

where $\varepsilon_j = \max(10^{-4}|\theta_{0,j}|, 10^{-8})$ and n_{obs} counts training-data points with finite measurements. Columns with $\|\partial z/\partial \theta_j\|_2 \le 10^{-10}$ (zero sensitivity) are discarded before pivoting.

Column-pivoted QR. Pivoted QR decomposition of $\widetilde{J}$ is

$$\widetilde{J}\,P = Q\,R, \qquad |R_{11}| \ge |R_{22}| \ge \cdots,$$

where P is a column-permutation matrix. The leading pivot columns of $\widetilde{J}P$—equivalently, the first entries of P—correspond to the parameters that contribute most independently to the measured outputs.

Subset selection. For each problem, the selection takes the top $k = |S|$ biological parameters from the QR ranking and appends the same calibration parameters Θ_C used in the original analysis.

Validation. The QR-selected subset is fitted on the training data using the same optimiser (trust-region reflective least squares) and the same number of function evaluations as the original reduced model. Test-set χ^2/n is then computed identically to the main experiment. The test set is never consulted at any stage of the QR selection or fitting.

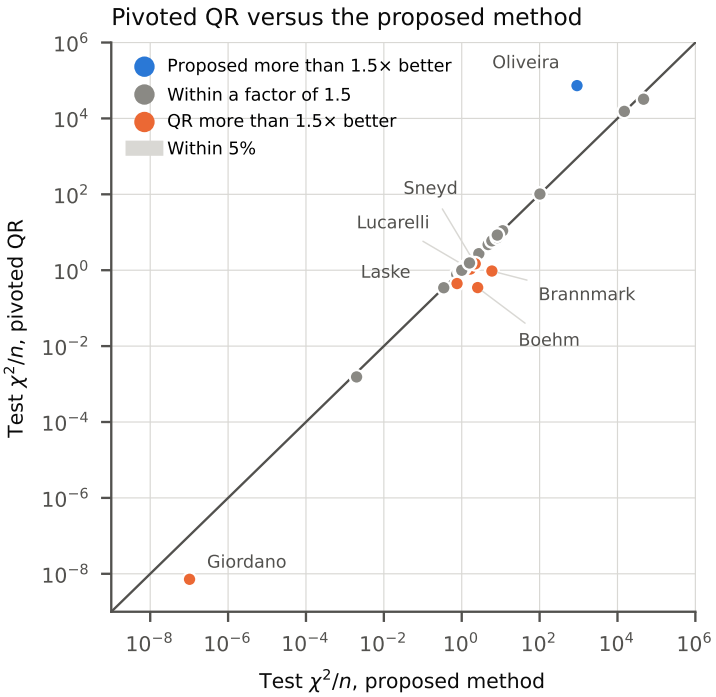
Results

Figure S10 compares χ^2_{prop} and χ^2_{QR} for all 28 evaluated problems. The grey band marks ±5% equivalence. Points on or near the diagonal correspond to problems where both methods achieve essentially the same predictive accuracy despite—in most cases—choosing entirely different parameters.

Among 26 non-trivial comparisons (both $\chi^2 > 10^{-3}$) the median ratio $\chi^2_{\text{QR}}/\chi^2_{\text{prop}} = 0.99$ and 15 problems fall within the ±5% band. QR predicted better by more than 5% in 9 problems (most notably Boehm and Brannmark) and the proposed method predicted better in 2 (most notably Oliveira, where $\chi^2_{\text{QR}} = 80 \times \chi^2_{\text{prop}}$; the QR-selected parameters failed to capture the epidemic dynamics on the test partition). In two additional problems (Giordano, Perelson) both methods fit the test data to within machine precision, making the ratio uninformative.

The most striking result is that the two methods selected *completely different* biological parameters in 19 of 28 problems (Table S23); the median overlap was 0 shared parameters. In 7 problems they shared exactly one parameter, in Lucarelli three of its ten, and only in Perelson ($k = 2$) did they agree completely. Despite this divergence, 23 of 26 comparisons produced ratios between 0.5 and 2.0. This is consistent with the sloppy-model picture (Gutenkunst et al., 2007): many distinct parameter combinations describe the training data approximately equally well, so the test-set χ^2/n is relatively insensitive to which specific k-parameter subset is selected.

The comparable performance of the two methods shows that the additional computational cost of the admissibility test and backward elimination does not translate into a systematic improvement in external validation accuracy on this benchmark. Conversely, it also does not lead to a systematic loss of accuracy: the two dramatic failures of QR on Oliveira and the large improvements of QR on Boehm and Brannmark are problem-specific and cancel on aggregate. The proposed method's value relative to pivoted-QR selection lies in the admissibility test (the reduced model is verified numerically to reproduce the full model's outputs under parameter perturbations; Supplementary Note *Admissibility of the subsets of other criteria* shows that pivoted-QR subsets of the earlier variant pass this test in only 14 of 28 problems) and in the systematic backward elimination that delivers a subset minimal with respect to the implemented test, including $|S| \geq 2$, not in a predictive accuracy advantage.



Supplementary Figure S10. Scatter plot of test-set χ^2/n for the proposed method (x-axis) versus pivoted-QR selection (y-axis), on a log-log scale. Each point is one of the 28 evaluated problems except Perelson, whose values ($\chi^2/n < 10^{-12}$) lie below the plot range. The solid diagonal is the 1:1 line; the grey band marks ±5%. Blue: proposed more than 1.5*imes* better; orange: QR more than 1.5× better; grey: within a factor of 1.5. Problems outside that factor are labelled.

Selected parameters

Supplementary Table S23 lists the biological parameters selected by each method for every problem. Parameter names follow the model's own PEtab identifiers. The overlap column gives the number of shared selections.

Supplementary Table S23. Biological parameters selected by the proposed method (S_{prop}) and by column-pivoted QR (S_{QR}) for each evaluated problem. *Overlap*: number of parameters shared by both subsets. Parameter names are PEtab identifiers; long names are line-broken.

Problem	k	S_{prop}	S_{QR}	Ovlp
Alkan	2	`kt`, `k_rep_nhej`	`p_chk1_dea_wip1`, `k_death_reox`	0
Armistead	2	`k_d`, `k00`	`k00`, `k2`	1
Bachmann	2	`CISRNADelay`, `JAK2ActEpo`	`SOCS3Inh`, `CISEqcOE`	0
Boehm	2	`k_imp_hetero`, `k_phos`	`k_phos`, `k_imp_homo`	1
Borghans	2	`Kz`, `K_par`	`Kd`, `init_A_state`	0
Brannmark	2	`k1f`, `km2`	`k1a`, `k_IRSiP_DosR`	0
Bruno	5	`kb1`, `init_bcar2`, `init_zea_1`, `init_bcar1`, `init_bcry_1`	`init_b10_1`, `init_bcar1`, `kc4`, `kc1`, `kc2`	1
Elowitz	2	`n_Hill`, `tau_mRNA`	`eff_GFP`, `tau_prot_GFP`	0
Fiedler	2	`tau2`, `k6`	`k11`, `k5`	0
Fujita	2	`init_AKT`, `reaction_6_k1`	`reaction_1_k1`, `reaction_6_k1`	1
Giordano	2	`alpha_4`, `alpha_28`	`delta_0`, `eta_0`	0
Isensee	2	`kf_RIIp_C_2__RII_C_2`, `xi_kf_RII_C_2__RII_2`	`kf_Fsk`, `KD_H89`	0
Laske	2	`k_imp`, `k_syn_R_M`	`k_imp`, `k_rel`	1
Liu	2	`kappa`, `phi`	`kappa`, `rho`	1
Lucarelli	10	`init_Rec`, `S2tot`, `S3tot`, `S_dephosphos`, `S_phos`, `geneF_turn`, `k_234`, `geneK_turn`, `S4tot`, `geneH_turn`	`S4tot`, `geneK_turn`, `geneJ_turn`, `geneB_act2`, `geneG_act1`, `geneH_turn`, `geneL_turn`, `geneE_act2`, `geneI_act1`, `geneC_act1`	3

Problem	k	S_{prop}	S_{QR}	Ovlp
Okuonghae	2	transmission_rate_effective, gamma_i	asymptomatic_start, d_D	0
Oliveira	2	beta_0, h_hosp_rate	beta_1, symptomatic_init_concentration	0
Perelson	2	c, delta	c, delta	2
Rahman	2	infected_moderate_transmission_rate, infected_normal_worsen_rate	infected_moderate_transmission_rate, treated_moderate_improve_rate	1
Raia	2	SOCS3mRNA_production, init_Rec_i	Rec_recycle, DecoyR_binding	0
Raimundez	2	kdeg_membran__MKN1, ka_MAPK__MKN1	d_MAPKtotal__MKN1_2_HS746T, d_MPI3Ktotal__fm_2_hm	0
Salazar-Cavazos	2	ratio_kpkd_Y1068__FREE, ratio_kpkd_YN__FREE	GRB2_total__FREE, kdephosYN__FREE	0
Schwen	2	ini_R1, ini_R2fold	kd2fold, kout	0
Smith	2	k8, kminus7b	kminus7a, k43r	0
Sneyd	2	l4, l2	k_2, k_3	0
Weber	2	s12, a21	a22, pu5	0
Zhao	3	R_Stage_I_China, R_Stage_II_Wuhan, gamma_1_Stage_I_Hubei	R_Stage_I_Wuhan, R_Stage_III_Wuhan, R_Stage_II_Hubei	0
Zheng	2	k01_02, k13_12	k01_11, k10_00	0

Supplementary Note: Profile likelihood

Purpose

Profile likelihood (Raue et al., 2009) is the standard method to assess which parameters the data determine, and profile-likelihood-based reduction (Maiwald et al., 2016) uses it to decide which parameters to fix. The profile of every biological parameter was computed for every problem in which it was run, for two purposes: as an independent estimate of how many parameters the data determine, and as one of the four criteria with which the proposed selection is compared (main text, *Comparison with existing selection criteria*).

Design

1. The full model (all free parameters) was re-estimated on the training data of the external validation, from θ_0, giving $\hat{\theta}_{\text{train}}$ and the minimum χ^2_{min}. The finite-difference Jacobian of this fit was evaluated in parallel.
2. For each biological parameter j, $\log|\theta_j|$ was fixed at $\log|\hat{\theta}_{\text{train},j}| \pm \{0.1, 0.3, 0.6, 1.2, 2.3\}$ (up to a factor of 10), within the same bounds $[|\theta_0|/10, 10|\theta_0|]$ as all other fits, and all other free parameters were re-estimated from $\hat{\theta}_{\text{train}}$; $\Delta\chi^2 = \chi^2 - \chi^2_{\text{min}}$. All points of a problem were computed in parallel. Points whose fit failed (simulation failure, $\Delta\chi^2 > 10^4$) were excluded.
3. If a profile point improved the minimum by more than one unit of χ^2, the full model had not converged: it was re-estimated from that point and all profiles were recomputed (at most three times; ten for Raia, which converged after four).
4. A parameter is *identifiable* if its profile exceeds $\Delta\chi^2 = 3.84$ (95%, one degree of freedom) on both sides, *bounded on one side* if only on one side, and *flat* otherwise. For the comparison, the $k = |S|$ parameters with the narrowest profile confidence interval were selected.

The computation requires repeated re-estimation of the full model. It was run for the 26 problems with at most 60 free parameters and, in a separate run with the profile points distributed over 60 cores, for Bachmann (90 free parameters; 70 min, 40 h of computing time) and Lucarelli (72; 11 min, 9 h). It was not run for Raimundez (126) and Chen (155) because of its cost, nor for Blasi, which has no training/test split. Extrapolated from Bachmann with the simulation time and the numbers of free parameters and profile points, Raimundez (about 570 points, 0.3 s per simulation) would need of the

order of 10^3 h of computing time and Chen (about 1,520 points, 35 s per simulation) of the order of 10^5 h.

Results

Supplementary Table S24 gives the classification of every problem. The median number of identifiable biological parameters was 2.5 (17% of the biological parameters); six problems had none and 14 of the 28 had two or fewer. This agrees with the size of the subsets selected by the proposed method, obtained by a different principle. The selected parameters, however, rarely coincide (Supplementary Table S25): a flat profile means that other free parameters can compensate for changes in the profiled parameter, whereas the proposed method asks which parameters suffice when all others are fixed at θ_0. Supplementary Figure S11 shows the profiles of three problems. Three problems needed re-estimations of the optimum (Borghans, Okuonghae, Raia), which shows that the full-model fit from θ_0 on the training data had not converged; Isensee kept a residual improvement of 1.16 after its three re-estimations, slightly above the criterion of one unit.

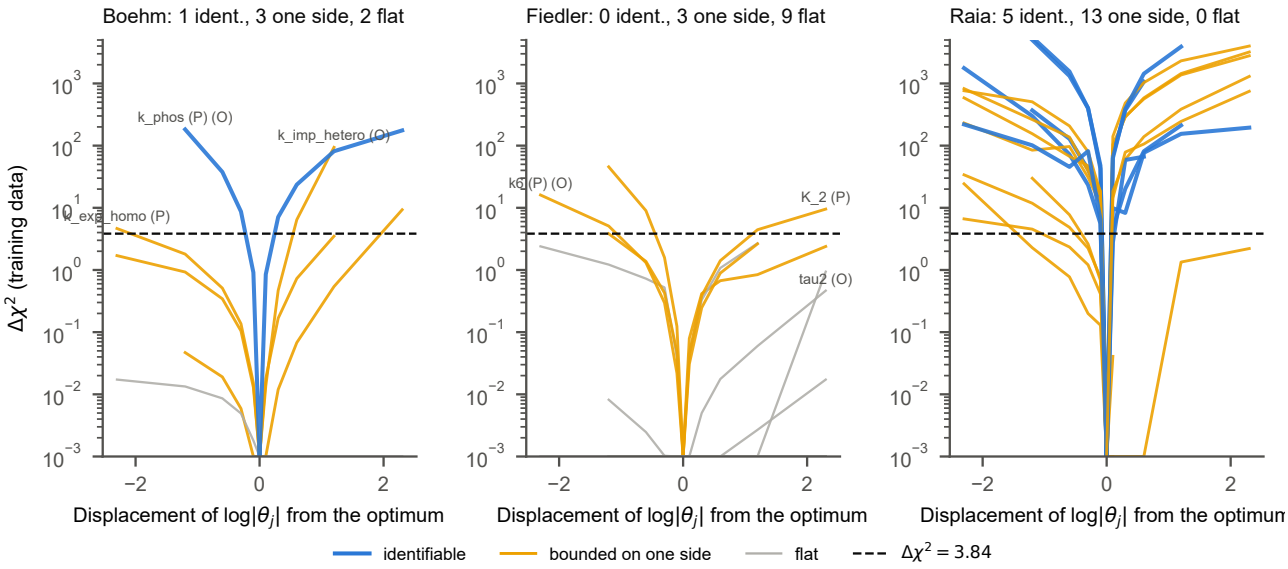
The two large problems. Bachmann and Lucarelli were profiled in a separate run, with the same protocol and the profile points distributed over 60 cores. In Bachmann (90 free parameters, of which 27 biological and the rest calibration parameters), the 221 profile points took 70 min of wall-clock time and 40 h of computing time (median 5.4 min per point, at most 54 min), against 10 s for the complete proposed method; 7 of the 27 biological parameters were identifiable, 5 bounded on one side and 15 flat. Profile likelihood selected `STAT5Imp` and `CISRNADelay`, which shares `CISRNADelay` with the proposed subset (`CISRNADelay`, `JAK2ActEpo`); both subsets are admissible (median e_{fit} 0.30 against 0.34 for the proposed subset at ±5%, 0.30 at ±20%) and predict the real test data equally well (test χ^2/n 4.77 for both). In Lucarelli (72 free parameters, all biological), the 592 points took 11 min and 9 h of computing time (median 46 s per point), against 35 s for the proposed method; 14 parameters were identifiable, 15 bounded on one side and 43 flat. The ten parameters selected by profile likelihood share four with the proposed subset (`init_Rec`, `S4tot`, `S_phos`, `geneK_turn`); they predict the real test data better (test χ^2/n 1.08 against 1.66) but do not reproduce the full model (median e_{fit} 0.74 at ±5% and 0.75 at ±20%, against 0.435 for the proposed subset). Neither problem needed a re-estimation of the optimum. Profile likelihood was not run for Chen and Raimundez because of the cost given above.

Supplementary Table S24. Profile likelihood of the biological parameters (training data). Ident.: profile above $\Delta\chi^2 = 3.84$ on both sides; one side: only on one side; flat: on neither side. Selected: the k parameters with the narrowest profile confidence interval. Common: parameters shared with the subset of the proposed method. Re-fits: times the full model was re-estimated because a profile point improved the optimum; residual: remaining improvement (in units of χ^2) after the last re-fit; a residual above one unit means that the optimum of the full model did not converge, so the classification of that problem is provisional. Failed pts: profile points excluded because the fit failed (simulation failure or $\Delta\chi^2 > 10^4$). A side is classified from its remaining points, so a failure can make a parameter look less identifiable than it is; this affects three parameters not classified as identifiable (Okuonghae gamma_i and Weber m31, one side; Weber p31, flat), whose class is therefore an upper bound on non-identifiability.

Problem	k	Bio	Ident.	One side	Flat	Selected by profile likelihood	Common	Note
Alkan	2	34	14	9	11	kt, k_death_reox	1	6 failed pts
Armistead	2	10	1	0	9	k_d, k2	1	–
Bachmann	2	27	7	5	15	STAT5Imp, CISRNADelay	1	–
Bertozzi	2	3	1	1	1	I0_CA, R0_CA	0	–
Blasi	*not evaluated: no training/test split*							
Boehm	2	6	1	3	2	k_phos, k_exp_homo	1	–

(continued)

Problem	k	Bio	Ident.	One side	Flat	Selected by profile likelihood	Common	Note
Borghans	2	20	17	2	1	Kz, v0	1	3 re-fits; res. 1.56, **not converged**
Brannmark	2	15	0	2	13	k_IRSiP_DosR, km2	1	–
Bruno	5	10	9	0	1	init_bcar1, kb1, init_bcry_1, init_b10_1, kc2	3	2 failed pts
Chen	*not evaluated: 155 free parameters, above the limit of 60*							
Elowitz	2	18	3	1	14	n_Hill, tau_prot_GFP	1	–
Fiedler	2	12	0	3	9	K_2, k6	1	–
Fujita	2	16	7	2	7	reaction_8_k1, EGFR_turnover	0	–
Giordano	2	43	0	0	43	nu_0, rho_22	0	–
Isensee	2	32	23	3	6	kf_RIIp_2__RII_2, kf_RII_C_2__RIIp_C_2	0	3 re-fits; res. 1.16, **not converged**; 6 failed pts
Laske	2	6	0	2	4	k_bind_M1, k_syn_R_M	1	–
Liu	2	7	2	0	5	kappa, rho	1	–
Lucarelli	10	72	14	15	43	init_Rec, Rec_act, geneL_turn, S4tot, geneK_turn, geneJ_turn, S_phos, geneC_turn, geneA_turn, geneB_turn	4	2 failed pts
Okuonghae	2	14	3	9	2	sigma, nu	0	2 re-fits; 4 failed pts
Oliveira	2	12	2	6	4	t_1, beta_1	0	1 failed pt
Perelson	2	2	0	0	2	c, delta	2	–
Rahman	2	9	0	1	8	behavioural_change_rate, infected_normal_worsen_rate	1	–
Raia	2	18	5	13	0	pSTAT5_dephosphorylation, pJAK2_dephosphorylation	0	4 re-fits; 2 failed pts
Raimundez	*not evaluated: 126 free parameters, above the limit of 60*							
Salazar-Cavazos	2	6	4	0	2	ratio_kpkd_YN__FREE, kdephosYN__FREE	1	–
Schwen	2	13	1	4	8	ini_R2fold, ini_R1	2	–
Smith	2	16	6	9	1	kminus7b, k30r	1	–
Sneyd	2	14	2	3	9	l4, k_3	1	–
Weber	2	26	1	11	14	pu5, m22	0	2 failed pts
Zhao	3	21	3	5	13	R_Stage_II_Hubei, R_Stage_II_China, gamma_2_Stage_III_Wuhan	0	–
Zheng	2	45	4	0	41	k02_01, k01_02	1	–



Supplementary Figure S11. Profile likelihoods of three problems. $\Delta\chi^2$ on the training data as a function of the displacement of $\log|\theta_j|$ from the optimum, for every biological parameter, coloured by class. Dashed line, 95% threshold. Labels mark the parameters selected by profile likelihood (P) and by the proposed method (O); in Raia they are omitted for legibility.

Supplementary Note: Brun and Morris selections

Purpose

The main text cites two further families of selection criteria: indices based on the collinearity of the sensitivity matrix, whose reference method is that of Brun *et al.* (Brun et al., 2001), and global sensitivity screening, of which Morris's method (Morris, 1991) is the most widely used. Both were implemented with the same protocol as the other criteria.

Design

Brun. Dimensionless sensitivities $s_{ij} = \theta_j\,\partial y_i/\partial\theta_j$ on the training outputs (forward differences with relative step 10^{-4}), importance $\delta_j^{\mathrm{msqr}} = \|s_{:,j}\|/\sqrt{n}$, collinearity index $\gamma_K = 1/\sqrt{\lambda_{\min}(\tilde{S}_K^{\top}\tilde{S}_K)}$ of the column-normalised subset and determinant measure $\rho_K = \det(S_K^{\top}S_K)^{1/(2k)}$. Among the 15 most important parameters, every k-subset was evaluated and the one with $\gamma_K \le 10$ and the largest ρ_K was selected (the one with the smallest γ_K if none satisfies the bound). The importance δ_j^{msqr} orders the parameters as the energy E_j of the proposed method; the difference lies in the selection rule and in the absence of a re-estimation test.

Morris. Ten trajectories (seed 42) on a four-level grid of $\log|\theta_j|$ within $[\theta_0/10, 10\theta_0]$ for the biological parameters (calibration parameters at θ_0), step $\Delta = 2/3$ of the range; elementary effect $\mathrm{EE}_j = \|y(x + \Delta e_j) - y(x)\|/\Delta$ on the training outputs; the k parameters with the largest $\mu_j^* = \mathrm{mean}|\mathrm{EE}_j|$ were selected.

Chen was not included, because the three selections (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. Supplementary Table S25 lists the parameters selected by all five criteria.

Implementation differences between the criteria

All criteria chose the same number k of biological parameters and were evaluated with the same tests, but they were not implemented with identical numerical settings. The differences, which a strictly controlled comparison would remove, are:

- *Data used to select.* The proposed method selects on all measured outputs of the problem; pivoted QR, Brun, Morris and profile likelihood select on the training outputs of the external validation. This is why Bertozzi, whose selected parameters act only in the test conditions, was excluded from the admissibility comparison (main text). In the other problems the training set is a subset of the outputs used by the proposed method.
- *Derivatives.* The proposed method uses central differences in $\log\theta$ with relative step 10^{-2}; pivoted QR and Brun use forward differences with relative step 10^{-4}; Morris uses finite elementary effects of 2/3 of the logarithmic range (a factor of about 21 in the parameter), so it measures global rather than local sensitivity, whereas the admissibility test is local ($\pm5\%$ and $\pm20\%$).
- *Candidates.* Brun evaluates all k-subsets of the 15 most important parameters only; the other criteria consider all biological parameters. Columns with zero sensitivity are discarded by pivoted QR and Brun (norm below 10^{-10}) and by the proposed method (zero energy).
- *Normalisation and ties in pivoted QR.* The columns are normalised to unit length, as is standard, so the magnitude of the sensitivity is discarded. As a consequence, all columns have the same norm at the first pivoting step, and the first parameter is chosen among them by rounding differences and column order; only the subsequent pivots are determined by the residual norm, that is, by collinearity with the parameters already chosen. This is a property of normalised pivoted QR, not of this implementation, and it contributes to its low admissibility.
- *Computational budget.* The criteria do not use the same number of simulations: pivoted QR and Brun need $p + 1$ simulations, Morris $10(p + 1)$, profile likelihood ten re-estimations of the full model per parameter, and the proposed method $3p$ simulations for J and the stage-2 ordering plus the re-estimations of the admissibility test for every candidate subset.

These differences do not by themselves invalidate the comparison, and some of them disadvantage the proposed method (for example, profile likelihood uses far more computation), but the comparison should be read as a comparison of the criteria as usually implemented, not of criteria that differ only in their selection rule. The comparison was not repeated with harmonised settings. The profile-likelihood classification has two further caveats, marked in Supplementary Table S24: in Borghans and Isensee a profile point still improved the optimum by more than one unit of χ^2 after the last re-estimation (1.56 and 1.16), so their full-model optimum did not converge and their classification is provisional; and failed profile points (simulation failure or $\Delta\chi^2 > 10^4$) were excluded, which can make a parameter look less identifiable than it is (three parameters in Okuonghae and Weber).

Supplementary Table S25. Biological parameters selected by each criterion (same number k for all). –, criterion not computed for the problem.

Problem	Proposed	Pivoted QR	Brun	Morris	Profile likelihood
Alkan	kt, k_rep_nhej	p_chk1_dea_wip1, k_death_reox	kt, p_atm_act_dsb	kt, k_damage_sn38_ssb	kt, k_death_reox
Armistead	k_d, k00	k00, k2	k_d, k00	k00, k_d	k_d, k2
Bachmann	CISRNADelay, JAK2ActEpo	SOCS3Inh, CISEqcOE	STAT5Imp, JAK2ActEpo	CISRNADelay, JAK2ActEpo	STAT5Imp, CISRNADelay
Bertozzi	R0_NY, I0_NY	R0_CA, I0_CA	R0_CA, I0_CA	gamma_CA, R0_CA	I0_CA, R0_CA
Blasi	a_basal, a_k16_k12k16	–	–	–	–
Boehm	k_imp_hetero, k_phos	k_phos, k_imp_homo	k_phos, k_imp_hetero	k_phos, k_imp_hetero	k_phos, k_exp_homo
Borghans	Kz, K_par	Kd, init_A_state	K_par, Vm2	K_par, beta_par	Kz, v0
Brannmark	k1f, km2	k1a, k_IRSiP_DosR	k1f, k1a	k22, k3	k_IRSiP_DosR, km2

(continued)

Problem	Proposed	Pivoted QR	Brun	Morris	Profile likelihood
Bruno	kb1, init_bcar2, init_zea_1, init_bcar1, init_bcry_1	init_b10_1, init_bcar1, kc4, kc1, kc2	init_bcar1, kb1, init_bcry_1, init_b10_1, kc2	init_bcry_1, init_b10_1, init_bcar1, kb2, kc2	init_bcar1, kb1, init_bcry_1, init_b10_1, kc2
Chen	k106, k44	–	–	–	–
Elowitz	n_Hill, tau_mRNA	eff_GFP, tau_prot_GFP	n_Hill, tau_mRNA	n_Hill, tau_prot_GFP	n_Hill, tau_prot_GFP
Fiedler	tau2, k6	k11, k5	tau2, k5	K_1, tau2	K_2, k6
Fujita	init_AKT, reaction_6_k1	reaction_1_k1, reaction_6_k1	init_AKT, reaction_6_k1	reaction_6_k1, reaction_8_k1	reaction_8_k1, EGFR_turnover
Giordano	alpha_4, alpha_28	delta_0, eta_0	alpha_4, alpha_22	alpha_4, alpha_0	nu_0, rho_22
Isensee	kf_RIIp_C_2__RII_C_2, xi_kf_RII_C_2__RII_2	kf_Fsk, KD_H89	kf_RII_C_2__RIIp_C_2, xi_kf_RII_C_2__RII_2	kf_RII_C_2__RIIp_C_2, kf_RIIp_2__RII_2	kf_RIIp_2__RII_2, kf_RII_C_2__RIIp_C_2
Laske	k_imp, k_syn_R_M	k_imp, k_rel	k_syn_R_M, k_bind_M1	k_syn_R_M, k_bind_M1	k_bind_M1, k_syn_R_M
Liu	kappa, phi	kappa, rho	kappa, rho	kappa, rho	kappa, rho
Lucarelli	init_Rec, S2tot, S3tot, S_dephosphos, S_phos, geneF_turn, k_234, geneK_turn, S4tot, geneH_turn	S4tot, geneK_turn, geneJ_turn, geneB_act2, geneG_act1, geneH_turn, geneL_turn, geneE_act2, geneI_act1, geneC_act1	init_Rec, S2tot, S_dephosphos, S_phos, Rec_act, geneF_turn, k_234, geneK_turn, geneH_turn, geneG_turn	init_Rec, S3tot, S2tot, pRec_degind, S_dephosphos, S_phos, geneF_turn, geneF_act1, Rec_act, geneF_inh1	init_Rec, Rec_act, geneL_turn, S4tot, geneK_turn, geneJ_turn, S_phos, geneC_turn, geneA_turn, geneB_turn
Okuonghae	transmission_rate_effective, gamma_i	asymptomatic_start, d_D	transmission_rate_effective, psi	sigma, nu	sigma, nu
Oliveira	beta_0, h_hosp_rate	beta_1, symptomatic_init_concentration	beta_0, t_1	beta_0, t_1	t_1, beta_1
Perelson	c, delta	c, delta	c, delta	c, delta	c, delta
Rahman	infected_moderate_transmission_rate, infected_normal_worsen_rate	infected_moderate_transmission_rate, treated_moderate_improve_rate	infected_moderate_transmission_rate, behavioural_change_rate	infected_normal_transmission_rate_relative, infected_moderate_transmission_rate	behavioural_change_rate, infected_normal_worsen_rate
Raia	SOCS3mRNA_production, init_Rec_i	Rec_recycle, DecoyR_binding	SOCS3mRNA_production, init_Rec_i	SOCS3mRNA_production, Rec_recycle	pSTAT5_dephosphorylation, pJAK2_dephosphorylation
Raimundez	kdeg_membran__MKN1, ka_MAPK__MKN1	d_MAPKtotal__MKN1_2_HS746T, d_MPI3Ktotal__fm_2_hm	kdeg_membran__MKN1, ka_MAPK__MKN1	d_AKTtotal__fm_2_hm, ki_AKT__MKN1	–
Salazar-Cavazos	ratio_kpkd_Y1068__FREE, ratio_kpkd_YN__FREE	GRB2_total__FREE, kdephosYN__FREE	ratio_kpkd_Y1068__FREE, ratio_kpkd_YN__FREE	ratio_kpkd_Y1068__FREE, ratio_kpkd_YN__FREE	ratio_kpkd_YN__FREE, kdephosYN__FREE
Schwen	ini_R1, ini_R2fold	kd2fold, kout	ini_R1, kon_unspec	ini_R2fold, ini_R1	ini_R2fold, ini_R1
Smith	k8, kminus7b	kminus7a, k43r	k8, k30f	kminus8, kminus7b	kminus7b, k30r
Sneyd	l4, l2	k_2, k_3	l4, l2	l2, k4	l4, k_3
Weber	s12, a21	a22, pu5	a21, s12	pu3, a21	pu5, m22
Zhao	R_Stage_I_China, R_Stage_II_Wuhan, gamma_1_Stage_I_Hubei	R_Stage_I_Wuhan, R_Stage_III_Wuhan, R_Stage_II_Hubei	R_Stage_I_China, gamma_1_Stage_I_Wuhan, R_Stage_II_Hubei	R_Stage_I_China, gamma_2_Stage_I_China, R_Stage_II_China	R_Stage_II_Hubei, R_Stage_II_China, gamma_2_Stage_III_Wuhan
Zheng	k01_02, k13_12	k01_11, k10_00	k01_11, k12_13	k01_02, k12_13	k02_01, k01_02

Supplementary Note: Comparison on real data

Supplementary Table S26 gives the test χ^2/n on the held-out experiments of every criterion and baseline, together with the full model and the published estimate, and Supplementary Table S27 the paired comparison with the proposed method over problems. When the non-selected parameters are kept at θ_0, all selection criteria and the random and calibration-only baselines predict similarly: the median ratios to the proposed method are 0.99 or 1.00, and only the earlier pivoted-QR variant (ratio 0.99, $p = 0.026$, unadjusted) and calibration-only re-estimation (0.99, $p = 0.003$) differ significantly, slightly in favour of the alternative; standard pivoted QR does not ($P = 0.16$, Table 2 of the main text). Oliveira is the only problem in which the specific selection matters strongly. As discussed in the main text, this test does not discriminate between criteria because the fixed parameters carry the information

of θ_0, which was estimated with all data.

Supplementary Table S26. Test χ^2/n on the held-out experiments (real data) of every selection criterion and baseline. All variants keep the non-selected parameters at θ_0 except the full model, which re-estimates all parameters on the training data. Bold: lowest value among the seven selection variants. θ_0, published estimate, obtained using all data including the test experiments.

Problem	Proposed	Calib. only	Random	QR (earlier)	Brun	Morris	Profile	Full	θ_0
Alkan	2.75	**2.72**	2.72	2.73	2.74	2.86	2.74	3.37	2.47
Armistead	1.22	**1.16**	1.16	1.21	1.22	1.22	1.37	1.37	1.16
Bachmann	4.77	4.76	4.84	**4.75**	4.75	4.77	4.77	7.35	1.05
Boehm	2.57	**0.341**	1.09	0.349	2.57	2.57	0.425	11.2	0.341
Borghans	0.743	0.778	0.768	0.776	**0.742**	0.743	0.744	0.744	0.74
Brannmark	5.96	**0.938**	2.95	0.951	1.17	147	1.82	6,277	0.938
Bruno	0.909	**0.909**	0.911	0.911	0.909	0.911	0.909	208	0.909
Chen	**20.9**	–	–	–	–	–	–	17.7	15.2
Elowitz	1.5	**1.39**	1.44	1.41	1.5	1.43	1.43	3.89	1.38
Fiedler	11.2	11.7	11.2	10.9	11	11.1	**9.56**	9.32	1.22
Fujita	5.94	**5.94**	6.02	5.94	5.94	5.94	5.98	8.03	5.63
Giordano	1.0×10^{-7}	7.4×10^{-9}	8.7×10^{-9}	7.2×10^{-9}	2.7×10^{-8}	1.3×10^{-8}	1.4×10^{-8}	0.00773	7.4×10^{-9}
Isensee	1.08	1.08	1.08	1.08	1.08	**1.08**	1.08	1.08	1.04
Laske	0.758	0.445	0.445	0.445	**0.445**	**0.445**	**0.445**	1.87	0.445
Liu	0.954	**0.942**	0.942	1.03	1.03	1.03	1.03	1.02	0.942
Lucarelli	1.66	**1.02**	1.06	1.07	1.02	1.03	1.08	3.34	1.02
Okuonghae	4.7×10^4	2.9×10^4	5.6×10^4	3.2×10^4	8.0×10^4	5.1×10^4	5.1×10^4	1.2×10^5	2.9×10^4
Oliveira	912	4.6×10^5	2.0×10^4	7.3×10^4	**553**	**553**	3.0×10^5	8.1×10^4	4.6×10^5
Perelson	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}	4.0×10^{-13}
Rahman	0.00199	**0.00148**	0.00171	0.00154	0.00247	0.00163	0.00272	0.0257	0.00148
Raia	8.22	**0.439**	2.21	7.24	8.23	8.15	0.799	5.1	0.417
Raimundez	1.17	**1.17**	1.18	1.17	1.17	1.17	–	1.43	0.98
Salazar-Cavazos	102	**96.9**	103	102	102	102	119	124	96.9
Schwen	8.2	**8.06**	8.35	8.43	8.3	8.2	8.2	8.4	0.768
Smith	1.5×10^4	1.6×10^4	1.6×10^4	1.5×10^4	1.4×10^4	1.5×10^4	1.5×10^4	1.5×10^4	8.3×10^4
Sneyd	2.22	**1.37**	1.89	1.48	2.22	2.18	1.52	2.56	1.37
Weber	1	0.997	1.01	**0.997**	1	45.6	0.997	16	0.993
Zhao	**0.345**	0.345	0.345	0.345	0.345	0.345	0.345	0.345	0.345
Zheng	1.58	**1.51**	1.52	1.55	1.59	1.58	3.16	20.8	1.51

Supplementary Table S27. Real data: paired comparison of the test χ^2/n of the proposed method with each alternative over problems. Better/worse: the proposed method lower/higher by more than 5%; equal: within 5%. Median ratio: alternative over the proposed method (below 1, the alternative predicts better). p: two-sided Wilcoxon signed-rank test on the log ratios.

The proposed method compared with	Problems	Proposed better	Proposed worse	Equal	Median ratio	
Pivoted QR (earlier variant)	28	2	10	16	0.99	0.02
Brun	28	3	5	20	1.00	0.2
Morris	28	4	6	18	1.00	0.4
Profile likelihood	27	7	9	11	1.00	0.4
Random subsets (median of 20)	28	3	9	16	1.00	0.2
Calibration only ($k = 0$)	28	2	13	13	0.99	0.002
Full model re-estimated on training data	29	20	3	6	1.23	8.2e-0
Reduced model, nested without θ_0 anchor	29	21	3	5	1.35	7.2e-0
Published estimate θ_0	29	2	20	7	0.90	0.0002

Supplementary Table S28 classifies every problem according to how the proposed method, pivoted QR and the two baselines compare.

Supplementary Table S28. External validation against three baseline selectors, with a classification of each problem. Test χ^2/n of the proposed method, of calibration-only re-estimation ($|S| = 0$), of pivoted QR and of random biological subsets of the same size k (median over 20 draws). Category: **S**, sloppy—all four values within 10% of each other; **C**, calibration dominates (bold, lowest value); **O**, the proposed method clearly best; **P**, all methods near zero (<0.01); **F**, all methods fail ($\chi^2/n \gg 1$).

Problem	k	χ^2_{prop}	χ^2_{vac}	χ^2_{QR}	χ^2_{rand}	Cat.
Alkan	2	2.748	2.720	2.727	2.722	S
Armistead	2	1.218	1.157	1.207	1.159	S
Bachmann	2	4.768	4.759	4.746	4.845	S
Boehm	2	2.567	**0.341**	**0.349**	1.095	C
Borghans	2	0.743	0.778	0.776	0.768	S
Brannmark	2	5.956	**0.938**	**0.951**	2.952	C
Bruno	5	0.909	0.909	0.911	0.911	S
Elowitz	2	1.498	1.391	1.407	1.439	S
Fiedler	2	11.21	11.73	10.95	11.22	S
Fujita	2	5.939	5.937	5.941	6.022	S
Giordano	2	<0.01	<0.01	<0.01	<0.01	P
Isensee	2	1.082	1.081	1.081	1.081	S
Laske	2	0.758	**0.445**	**0.445**	**0.445**	C
Liu	2	0.954	0.942	1.026	0.942	S
Lucarelli	10	1.659	**1.015**	1.073	1.059	C
Okuonghae	2	4.7×10^4	2.9×10^4	3.2×10^4	5.6×10^4	F
Oliveira	2	**912**	4.6×10^5	7.3×10^4	2.0×10^4	O
Perelson	2	<0.01	<0.01	<0.01	<0.01	P
Rahman	2	<0.01	<0.01	<0.01	<0.01	P
Raia	2	8.224	**0.439**	7.236	2.213	C
Raimundez	2	1.167	1.167	1.168	1.182	S
Salazar-C.	2	101.8	96.93	101.7	103.2	F
Schwen	2	8.196	8.063	8.432	8.353	S
Smith	2	1.5×10^4	1.6×10^4	1.5×10^4	1.6×10^4	F
Sneyd	2	2.222	**1.373**	1.480	1.890	C
Weber	2	1.001	0.997	0.997	1.014	S
Zhao	3	0.345	0.345	0.345	0.345	S
Zheng	2	1.582	1.505	1.553	1.521	S

Supplementary Note: Admissibility of the subsets of other criteria

Purpose

The proposed method returns a subset only if re-estimating it reproduces the full model under random parameter changes (admissibility). This note applies exactly the same test to the subsets of the other criteria, to determine whether they have this property without being selected for it.

Design

The test replicates the code of the method: 15 scenarios $\theta^{(k)} = \theta_0 \odot (1 + \varepsilon u^{(k)})$ with the same seed (42), re-estimation of the subset together with the calibration parameters from two starting points, scenarios with non-finite outputs discarded, and the subset is admissible if the smaller of the median e_{fit} and the median e_{rel} is at most $\tau_e = 0.436$ (with $|S| \geq 2$), exactly as for the proposed method (Methods of the main text; with nonlinear re-estimation the median e_{fit} is never larger than the median e_{rel}, so the rule reduces to the median e_{fit}). The eight problems that the method treats to first order were treated in the same way for every criterion. As a check, the test reproduces the published e_{fit} of the proposed

subsets (for example Armistead 0.330, Boehm 0.094, Fiedler 0.147). Because the proposed subsets are selected by this test at $\varepsilon = 0.05$, it was repeated at $\varepsilon = 0.20$, a level used by none of the criteria.

Results

Supplementary Table S29 and Supplementary Figure S12 give the result of every criterion in every problem, and Supplementary Table S30 the paired comparison. At ±5% the proposed subsets were admissible in all 28 problems, against 26 (Brun), 23 (Morris), 18 of 27 (profile likelihood) and 14 (pivoted QR); at ±20%, 27 of 28 against 25, 23, 18 of 27 and 15. Bertozzi is not included: the proposed method selected its parameters on all measured outputs, which in this problem include the test conditions, whereas the other criteria selected on the training outputs, so the criteria did not use the same information. There is no problem in which the subset of another criterion is admissible and the proposed one is not. The only problem in which the proposed subset loses admissibility at ±20% is Elowitz, one of the problems that are admissible with a small margin at ±5% (main text, *Robustness*).

Supplementary Table S29. Admissibility of the subsets of every criterion: smaller of the median e_{fit} and the median e_{rel} over 15 scenarios (the quantity compared with $\tau_e = 0.436$ by the method), for parameter changes of ±5% (left) and ±20% (right). ✓, admissible. The eight problems that the method treats to first order are treated in the same way for every criterion.

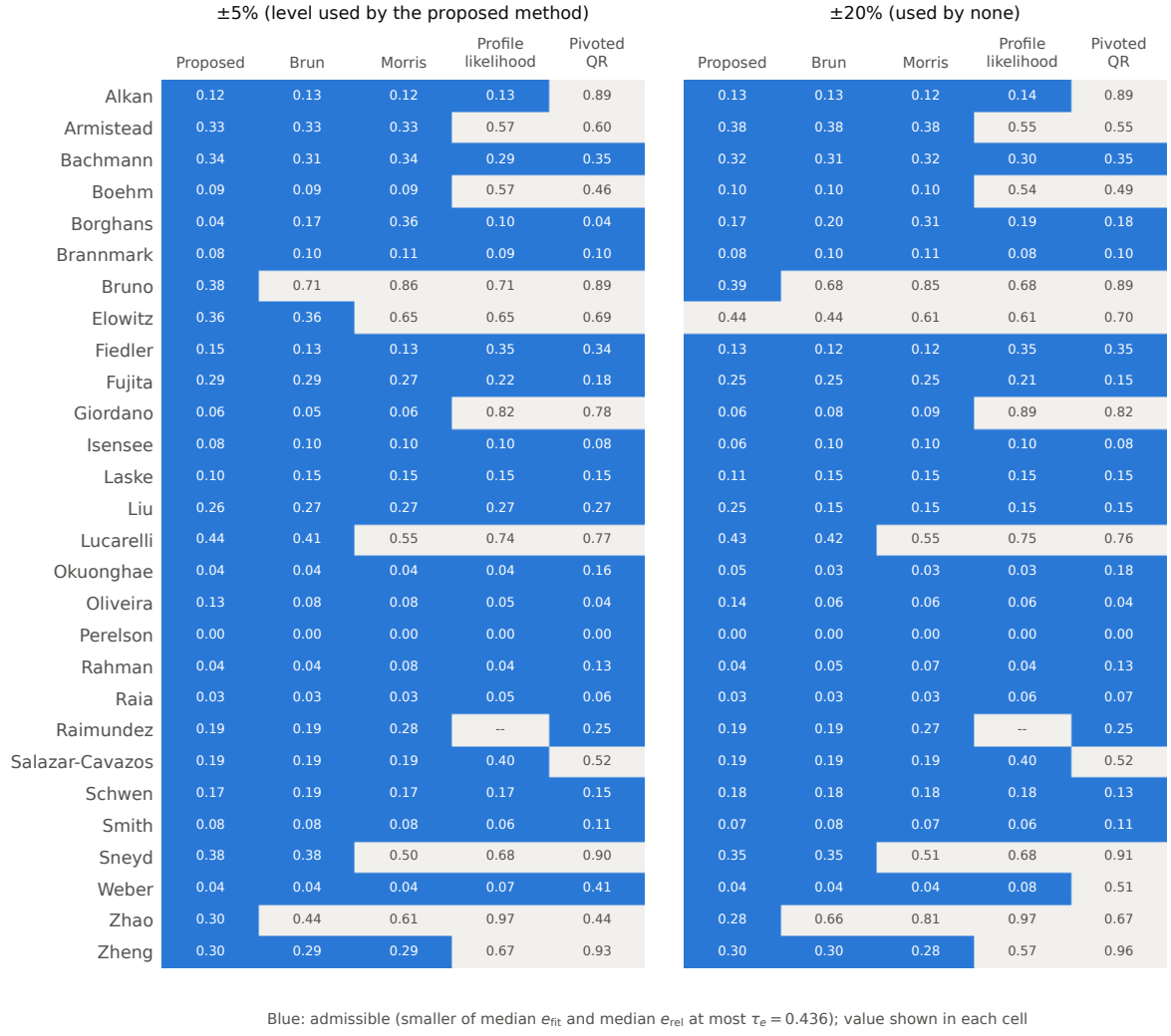
	±5%					±20%				
Problem	Proposed	QR	Brun	Morris	Profile	Proposed	QR	Brun	Morris	Profile
Alkan	0.121✓	0.945	0.126✓	0.116✓	0.131✓	0.133✓	0.940	0.131✓	0.122✓	0.136✓
Armistead	0.330✓	0.741	0.330✓	0.330✓	0.567	0.385✓	0.708	0.385✓	0.385✓	0.555
Bachmann	0.337✓	0.503	0.307✓	0.337✓	0.295✓	0.324✓	0.499	0.312✓	0.324✓	0.303✓
Boehm	0.094✓	0.670	0.094✓	0.094✓	0.565	0.095✓	0.648	0.095✓	0.095✓	0.535
Borghans	0.043✓	0.533	0.166✓	0.355✓	0.097✓	0.169✓	0.345✓	0.198✓	0.313✓	0.190✓
Brannmark	0.078✓	0.116✓	0.098✓	0.105✓	0.088✓	0.080✓	0.111✓	0.103✓	0.110✓	0.084✓
Bruno	0.378✓	0.894	0.705	0.858	0.705	0.388✓	0.893	0.680	0.847	0.680
Elowitz	0.358✓	0.719	0.358✓	0.650	0.650	0.436	0.715	0.436	0.610	0.610
Fiedler	0.147✓	0.133✓	0.133✓	0.134✓	0.347✓	0.133✓	0.115✓	0.118✓	0.118✓	0.346✓
Fujita	0.294✓	0.177✓	0.294✓	0.268✓	0.217✓	0.252✓	0.154✓	0.252✓	0.251✓	0.211✓
Giordano	0.058✓	0.909	0.054✓	0.064✓	0.816	0.059✓	0.964	0.080✓	0.086✓	0.893
Isensee	0.076✓	0.102✓	0.098✓	0.103✓	0.103✓	0.058✓	0.096✓	0.098✓	0.103✓	0.103✓
Laske	0.101✓	0.125✓	0.149✓	0.149✓	0.149✓	0.106✓	0.131✓	0.146✓	0.146✓	0.146✓
Liu	0.261✓	0.271✓	0.271✓	0.271✓	0.271✓	0.252✓	0.146✓	0.146✓	0.146✓	0.146✓
Lucarelli	0.435✓	0.911	0.413✓	0.549	0.740	0.428✓	0.908	0.422✓	0.549	0.746
Okuonghae	0.041✓	0.112✓	0.041✓	0.042✓	0.042✓	0.047✓	0.169✓	0.034✓	0.035✓	0.035✓
Oliveira	0.129✓	0.039✓	0.078✓	0.078✓	0.048✓	0.141✓	0.039✓	0.062✓	0.062✓	0.060✓
Perelson	0.000✓	0.000✓	0.000✓	0.000✓	0.000✓	0.000✓	0.000✓	0.000✓	0.000✓	0.000✓
Rahman	0.037✓	0.112✓	0.035✓	0.085✓	0.035✓	0.040✓	0.124✓	0.047✓	0.069✓	0.039✓
Raia	0.033✓	0.042✓	0.033✓	0.034✓	0.048✓	0.032✓	0.040✓	0.032✓	0.032✓	0.061✓
Raimundez	0.187✓	0.282✓	0.187✓	0.277✓	–	0.194✓	0.279✓	0.194✓	0.275✓	–
Salazar-Cavazos	0.187✓	0.520	0.187✓	0.187✓	0.398✓	0.188✓	0.518	0.188✓	0.188✓	0.396✓
Schwen	0.171✓	0.189✓	0.193✓	0.171✓	0.171✓	0.179✓	0.173✓	0.184✓	0.179✓	0.179✓
Smith	0.078✓	0.085✓	0.078✓	0.077✓	0.060✓	0.069✓	0.088✓	0.075✓	0.069✓	0.065✓
Sneyd	0.381✓	0.736	0.381✓	0.497	0.682	0.349✓	0.731	0.349✓	0.512	0.685
Weber	0.041✓	0.482	0.041✓	0.044✓	0.069✓	0.039✓	0.543	0.039✓	0.044✓	0.081✓
Zhao	0.296✓	0.970	0.438	0.608	0.975	0.282✓	0.972	0.662	0.807	0.975
Zheng	0.300✓	0.697	0.294✓	0.292✓	0.671	0.296✓	0.580	0.296✓	0.285✓	0.571
Admissible	28/28	14/28	26/28	23/28	18/27	27/28	15/28	25/28	23/28	18/27

Supplementary Table S30. Admissibility of the proposed subsets compared with those of each criterion: exact McNemar test over problems (problems in which only one of the two subsets is admissible).

Level	Criterion	Problems	Only proposed admissible	Only the other	p
±5%	Pivoted QR	28	14	0	0.00012
±5%	Brun	28	2	0	0.5
±5%	Morris	28	5	0	0.062

(continued)

Level	Criterion	Problems	Only proposed admissible	Only the other	p
±5%	Profile likelihood	27	9	0	0.0039
±20%	Pivoted QR	28	12	0	0.00049
±20%	Brun	28	2	0	0.5
±20%	Morris	28	4	0	0.12
±20%	Profile likelihood	27	8	0	0.0078



Supplementary Figure S12. Admissibility of the subsets of every criterion in every problem. Cell value: smaller of the median e_{fit} and the median e_{rel} over 15 scenarios; blue, admissible ($\leq \tau_e = 0.436$). Left, ±5%; right, ±20%. –, criterion not computed.

Supplementary Note: Comparison of selection criteria on synthetic data

Design

The synthetic prediction experiment (Supplementary Note *Synthetic prediction experiment*) was repeated with the subsets of every criterion, with the same 20 seeds (42–61), so that the five criteria are evaluated on exactly the same truths and noise realisations. Each subset, with the calibration parameters, was re-estimated on the synthetic training data from θ_0 and e_{pred} was computed on the

test data. The comparison uses the cases common to the five criteria in which the parameter change exceeded the noise (281 cases in 24 problems, as in Table 2 of the main text; pivoted QR in the standard Golub–Klema–Stewart form). Because the 20 seeds of a problem are not independent, the main statistic resamples whole problems (Methods, *Statistics*).

Results

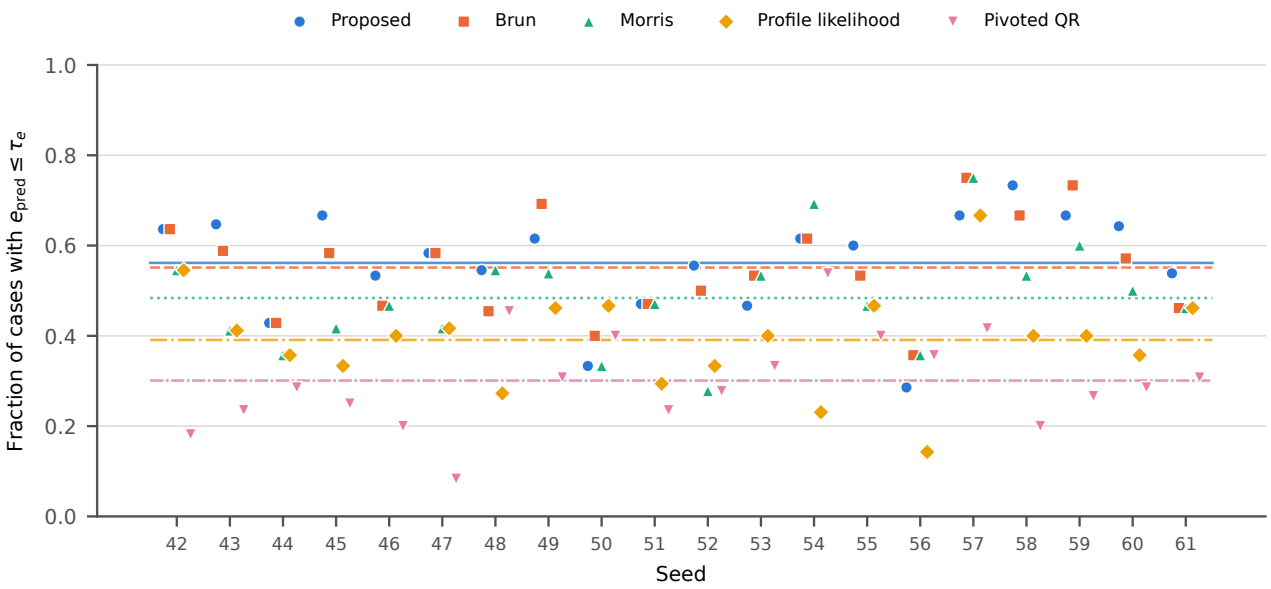
Supplementary Table S31 gives the results per problem and Supplementary Table S32 the statistics. The proposed subsets reached $e_{\mathrm{pred}} \leq \tau_e$ in 56% of the cases, those of Brun in 55%, Morris 48%, profile likelihood 39% and pivoted QR 30%; the advantage of the proposed method over pivoted QR (26 percentage points, 95% confidence interval 12–40), profile likelihood (17, 7–28) and Morris (8, 2–16) is significant, whereas the difference from Brun (1, −1 to 3) is not, which does not establish equivalence. The per-problem median prediction error differed significantly only from that of pivoted QR (Wilcoxon $p = 0.02$; $p \geq 0.23$ for the other criteria). Supplementary Figure S13 shows that the ordering of the criteria is similar across seeds.

Supplementary Table S31. Synthetic prediction experiment for the five criteria (20 seeds, cases common to all criteria in which the parameter change exceeded the noise; the same cases as in Table 2 of the main text; pivoted QR in the standard Golub–Klema–Stewart form). Each entry: number of cases with $e_{\mathrm{pred}} \leq \tau_e$ and, in parentheses, the median e_{pred}.

Problem	Cases	Proposed	Pivoted QR	Brun	Morris	Profile
Alkan	3	3 (0.21)	0 (0.76)	3 (0.28)	3 (0.16)	3 (0.28)
Armistead	12	7 (0.42)	6 (0.40)	7 (0.42)	7 (0.42)	4 (0.65)
Bachmann	2	0 (1.63)	0 (1.42)	0 (1.81)	0 (1.36)	0 (1.75)
Boehm	4	1 (0.53)	2 (0.43)	1 (0.53)	1 (0.53)	0 (0.94)
Borghans	11	2 (0.81)	1 (0.91)	1 (0.83)	1 (0.83)	1 (0.85)
Brannmark	18	1 (1.84)	1 (1.71)	0 (1.23)	0 (7.56)	0 (0.99)
Bruno	6	4 (0.34)	0 (1.00)	4 (0.34)	0 (0.99)	4 (0.34)
Elowitz	17	11 (0.27)	3 (0.63)	11 (0.27)	7 (0.54)	7 (0.54)
Fiedler	6	1 (1.00)	1 (0.94)	1 (1.00)	1 (0.99)	0 (1.68)
Fujita	18	14 (0.31)	17 (0.23)	14 (0.31)	14 (0.30)	16 (0.24)
Isensee	11	8 (0.15)	11 (0.12)	7 (0.15)	6 (0.24)	3 (0.94)
Liu	1	0 (0.64)	0 (0.50)	0 (0.50)	0 (0.50)	0 (0.50)
Okuonghae	20	17 (0.16)	12 (0.41)	18 (0.06)	18 (0.09)	18 (0.09)
Oliveira	20	12 (0.39)	6 (1.36)	13 (0.35)	13 (0.35)	7 (0.63)
Perelson	1	0 (20.95)	0 (20.95)	0 (20.95)	0 (20.95)	0 (20.95)
Rahman	19	15 (0.12)	16 (0.25)	13 (0.13)	13 (0.39)	14 (0.23)
Raia	8	7 (0.24)	0 (1.01)	7 (0.24)	7 (0.23)	0 (1.01)
Salazar-Cavazos	18	12 (0.33)	0 (0.92)	12 (0.33)	12 (0.33)	2 (0.83)
Schwen	9	0 (1.62)	0 (1.62)	0 (1.42)	0 (1.62)	0 (1.62)
Smith	20	18 (0.10)	8 (0.67)	19 (0.11)	18 (0.09)	19 (0.11)
Sneyd	1	1 (0.27)	0 (0.98)	1 (0.27)	0 (0.88)	1 (0.40)
Weber	20	11 (0.36)	0 (0.86)	11 (0.36)	2 (0.86)	6 (0.59)
Zhao	16	0 (1.00)	0 (1.00)	0 (1.00)	0 (1.00)	0 (1.00)
Zheng	20	12 (0.41)	0 (0.95)	11 (0.42)	11 (0.41)	4 (0.72)
Total	281	157 (56%)	84 (30%)	154 (55%)	134 (48%)	109 (39%)

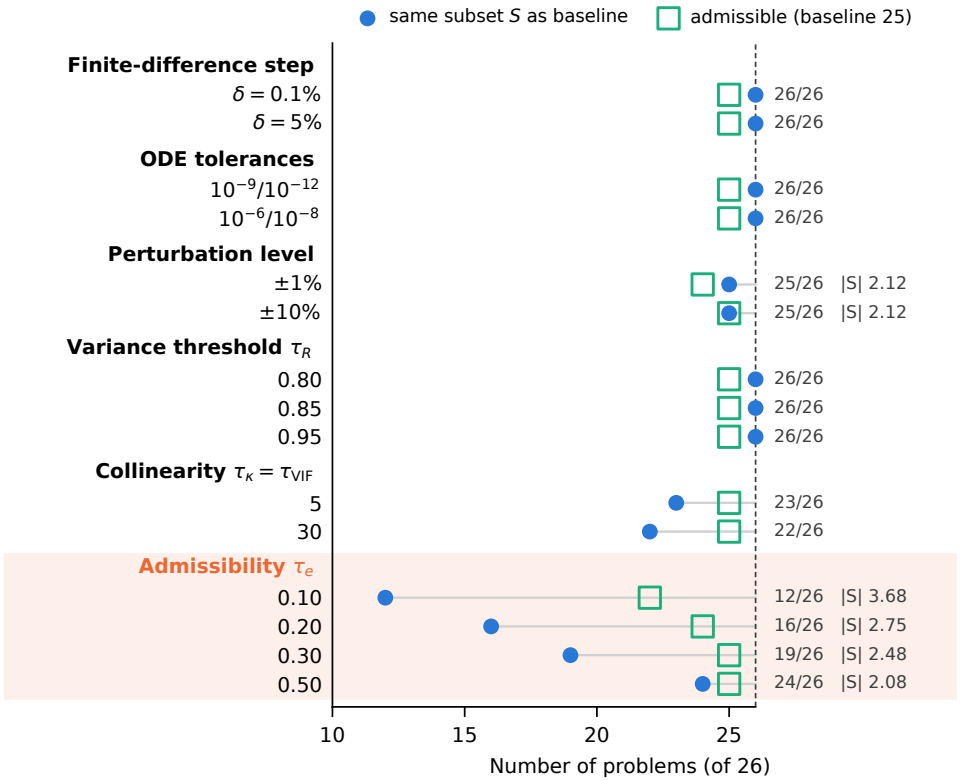
Supplementary Table S32. Synthetic experiment, statistics over problems. Left: difference in the fraction of cases with $e_{\text{pred}} \leq \tau_e$ (proposed minus criterion, percentage points) with a 95% confidence interval and p-value from resampling whole problems with all their seeds (10,000 replicates). Right: per-problem median over seeds of the ratio $e_{\text{pred}}(\text{criterion})/e_{\text{pred}}(\text{proposed})$ (problems with at least 10 seeds), its median and the two-sided Wilcoxon p-value.

Criterion	Difference	95% CI	p	Problems	Median ratio	p
Pivoted QR	+26.0	+11.5 to +39.6	0.0001	28	1.08	0.02
Brun	+1.1	-1.1 to +3.4	0.43	28	1.00	0.37
Morris	+8.2	+2.1 to +16.1	0.0052	28	1.00	0.29
Profile likelihood	+17.1	+7.4 to +27.6	0.0008	27	1.04	0.23



Supplementary Figure S13. Synthetic experiment: fraction of cases with $e_{\text{pred}} \leq \tau_e$ for each criterion and seed. Horizontal lines, mean over seeds. Cases common to the five criteria with signal.

Supplementary Note: Sensitivity to the method settings



Supplementary Figure S14. Sensitivity of the selection to the method settings. The complete method was re-run on 26 problems changing one setting at a time with respect to the baseline ($\delta = 1\%$, tolerances $10^{-7}/10^{-10}$, perturbation $\pm 5\%$, $\tau_R = 0.89$, $\tau_\kappa = \tau_{\mathrm{VIF}} = 10$, $\tau_e = 0.436$; 25 admissible problems with a mean of 2.16 selected biological parameters). Filled circles, number of problems whose selected subset is identical to the baseline; open squares, number of problems with an admissible subset; dashed line, all 26 problems. $|S|$, mean number of selected biological parameters among admissible problems, where it differs from the baseline. The admissibility threshold τ_e (shaded) is the only setting with a large effect. Results for every problem are given in Supplementary Table S33.

The method depends on a finite-difference step δ, the integrator tolerances, the perturbation size used for admissibility and four thresholds (τ_R, $\tau_\kappa = \tau_{\mathrm{VIF}}$, τ_e). The complete method (selection, admissibility and backward elimination, with the calibration parameters always re-estimated) was re-run changing one setting at a time, with all others at their baseline values ($\delta = 1\%$, relative/absolute tolerances $10^{-7}/10^{-10}$, perturbation $\pm 5\%$, $\tau_R = 0.89$, $\tau_\kappa = \tau_{\mathrm{VIF}} = 10$, $\tau_e = 0.436$). The 26 problems whose complete method could be re-run fifteen times in reasonable time were used; the other 8 (Alkan, Bachmann, Beer, Chen, Isensee, Lang, Lucarelli, Raimundez) need hours per run. For each setting and problem, the table records whether the selected subset is identical to the baseline and whether the problem is admissible.

Supplementary Table S33. Effect of each setting on every problem: =, same subset and admissibility as the baseline; ≠, different subset (still admissible); ×, not admissible; ∘, not admissible in the baseline either. Columns in the order of the rows of Supplementary Fig. S14.

Problem	$\delta = 0.1\%$	$\delta = 5\%$	tolerances $10^{-9}/10^{-12}$	tolerances $10^{-6}/10^{-8}$	perturbation ±1%	perturbation ±10%	$\tau_R = 0.80$	$\tau_R = 0.85$	$\tau_R = 0.95$	$\tau_\kappa = \tau_{\mathrm{VIF}} = 5$	$\tau_\kappa = \tau_{\mathrm{VIF}} = 30$	$\tau_e = 0.10$	$\tau_e = 0.20$	$\tau_e = 0.30$	$\tau_e = 0.50$
Armistead	=	=	=	=	=	=	=	=	=	=	=	≠	≠	≠	=
Bertozzi	=	=	=	=	=	=	=	=	=	=	=	≠	≠	≠	=
Blasi	=	=	=	=	=	=	=	=	=	=	=	≠	≠	≠	=
Boehm	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
Borghans	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
Brannmark	=	=	=	=	=	=	=	=	=	=	≠	=	=	=	=
Bruno	=	=	=	=	=	=	=	=	=	=	=	≠	≠	≠	≠
Crauste	∘	∘	∘	∘	∘	∘	∘	∘	∘	∘	∘	∘	∘	∘	∘
Elowitz	=	=	=	=	=	=	=	=	=	=	=	≠	≠	≠	=
Fiedler	=	=	=	=	=	=	=	=	=	=	=	≠	=	=	=
Fujita	=	=	=	=	=	=	=	=	=	=	=	≠	≠	=	=
Giordano	=	=	=	=	=	=	=	=	=	=	≠	=	=	=	=
Laske	=	=	=	=	=	=	=	=	=	=	=	≠	=	=	=
Liu	=	=	=	=	=	=	=	=	=	=	=	×	×	=	=
Okuonghae	=	=	=	=	=	=	=	=	=	=	≠	=	=	=	=
Oliveira	=	=	=	=	=	=	=	=	=	=	≠	×	=	=	=
Perelson	=	=	=	=	×	=	=	=	=	=	=	=	=	=	=
Rahman	=	=	=	=	=	=	=	=	=	≠	=	=	=	=	=
Raia	=	=	=	=	=	=	=	=	=	≠	=	=	=	=	=
Salazar-Cavazos	=	=	=	=	=	=	=	=	=	≠	=	×	=	=	=
Schwen	=	=	=	=	=	=	=	=	=	=	=	≠	=	=	=
Smith	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
Sneyd	=	=	=	=	=	=	=	=	=	=	=	≠	≠	≠	=
Weber	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
Zhao	=	=	=	=	≠	≠	=	=	=	=	=	≠	≠	=	≠
Zheng	=	=	=	=	=	=	=	=	=	=	=	≠	≠	≠	=
Admissible	25	25	25	25	24	25	25	25	25	25	25	22	24	25	25
Same S	26	26	26	26	25	25	26	26	26	23	22	12	16	19	24

Reproduction error in units of the noise

Purpose. The admissibility threshold bounds the relative error $e_{\mathrm{fit}} = \|y(\hat{\theta}_{S\cup\Theta_C}) - y(\theta^{(k)})\|/\|\Delta y_{\mathrm{full}}\|$, the part of the change of the model response that the reduced model cannot reproduce after re-estimation. A relative error does not say whether the remaining discrepancy matters for the data: 40% of a change that is itself much smaller than the measurement noise is irrelevant, whereas 5% of a change of a hundred noise units is not. The threshold $\tau_e = \sqrt{1 - c^2} = 0.436$ follows from the cosine requirement $c = 0.9$ (Theorem 1 of the main text), which is a design choice; to judge it against the precision of the measurements the same error was also expressed in absolute terms.

Derivation. The outputs of the method are scaled by the recorded noise, $y_i = h(g_i)/\sigma_i$, so a Euclidean norm of a change of y is measured in noise units. For scenario k at ±5%, with n measurements,

$$\mathrm{change}_k = \frac{\|\Delta y^{(k)}_{\mathrm{full}}\|}{\sqrt{n}}, \qquad \mathrm{error}_k = e^{(k)}_{\mathrm{fit}}\,\frac{\|\Delta y^{(k)}_{\mathrm{full}}\|}{\sqrt{n}} = \frac{\|y(\hat{\theta}^{(k)}_{S\cup\Theta_C}) - y(\theta^{(k)})\|}{\sqrt{n}}$$

are the root mean square per measurement of the change of the full model and of the error that remains after re-estimating the selected and calibration parameters, both in units of σ. An error below

one means that, on average over the measurements, the reduced model reproduces the perturbed full model to within the noise of the data. The values of $e_{\mathrm{fit}}^{(k)}$ are those of the robustness analysis at ±5% (the scenarios of the selection), and $\|\Delta y_{\mathrm{full}}^{(k)}\|$ was recomputed with the same directions (`absolute_error.py` in the reproducibility package; results in `results/absolute_error/`).

Results. Supplementary Table S34 gives the median change, the median error and the largest error over the scenarios for every admissible problem. The median error is below 0.1σ in 17 of the 31 problems and below σ in 29, and the largest error over the scenarios is below σ in 27. The relative error and the absolute error answer different questions: Lucarelli, with the largest relative error (43.5%, on the threshold), has an absolute error of only 0.05σ, because a ±5% perturbation changes its outputs by 0.13σ; in Giordano and Perelson the perturbations change the outputs by far less than the noise ($10^{-4}\sigma$ or less), so their relative errors, although computed, have no consequence for the data. Conversely, in Okuonghae and Oliveira a ±5% perturbation changes the outputs by about 30 and 120 noise units, so a small relative error (4% and 13%) still leaves an error of about one and fourteen noise units; these are also the problems in which even the published θ_0 fits the data far worse than the recorded noise (main text, *Prediction of held-out experiments*), which indicates that their recorded noise levels are much smaller than the actual scatter. An application that requires the reduced model to stay within the noise of its own data should therefore check this absolute error in addition to e_{fit}, or choose c accordingly.

Supplementary Table S34. Reproduction error of the selected subsets in units of the recorded noise, over the scenarios at ±5%. n: number of measurements; Valid: valid scenarios. Change: root mean square of Δy_{full} per measurement, in units of σ (median over the scenarios); Error: root mean square of the remaining error after re-estimation, $e_{\mathrm{fit}}\|\Delta y_{\mathrm{full}}\|/\sqrt{n}$ (median and maximum over the scenarios). A value below one means that the reduced model reproduces the perturbed full model to within the noise of the measurements; a change below one means that the perturbation itself is smaller than the noise (25 of the 31 problems). For the six problems decided to first order (Alkan, Bachmann, Chen, Isensee, Lucarelli, Raimundez) the median e_{fit} is the first-order value of this run; Table 1 of the main text gives the nonlinear re-estimation of the final subset.

Problem	n	Valid	Median e_{fit} (%)	Change (RMS, σ)	Error (RMS, σ)	Max error
Alkan	1733	15	12.1	0.30	0.03	0.04
Armistead	58	15	33.0	0.23	0.11	0.20
Bachmann	541	15	33.7	0.14	0.04	0.06
Bertozzi	22	15	32.9	0.33	0.16	0.26
Blasi	252	15	40.2	0.29	0.11	0.18
Boehm	48	15	9.4	0.13	9.4×10^{-3}	0.02
Borghans	111	15	4.3	0.87	0.03	1.01
Brannmark	43	15	7.8	0.81	0.06	0.14
Bruno	77	15	37.8	0.38	0.16	0.23
Chen	120	8	8.2	1.95	0.15	0.21
Elowitz	58	15	35.8	0.57	0.16	0.26
Fiedler	72	15	14.7	0.17	0.02	0.04
Fujita	144	15	29.4	0.72	0.22	0.33
Giordano	313	15	5.8	5.5×10^{-5}	3.6×10^{-6}	7.5×10^{-6}
Isensee	687	4	9.8	0.43	0.05	0.06
Laske	42	15	10.1	0.21	0.02	0.03
Liu	100	15	26.1	0.07	0.01	0.02
Lucarelli	1755	15	43.5	0.13	0.05	0.06
Okuonghae	92	15	4.1	31.56	1.18	3.68
Oliveira	120	15	12.9	1.2×10^{2}	13.74	46.98
Perelson	16	2	0.0	2.9×10^{-7}	0	0
Rahman	23	15	3.7	1.21	0.05	0.09
Raia	205	15	3.3	4.24	0.10	0.19
Raimundez	627	15	18.7	0.81	0.17	0.26
Salazar-Cavazos	18	15	18.7	0.74	0.11	0.27

Problem	n	Valid	Median e_{fit} (%)	Change (RMS, σ)	Error (RMS, σ)	Max error
Schwen	286	15	17.1	0.14	0.02	0.04
Smith	62	15	7.8	9.31	0.66	1.29
Sneyd	135	15	38.1	0.15	0.05	0.14
Weber	135	15	4.1	0.90	0.04	0.07
Zhao	82	15	29.6	0.42	0.15	0.23
Zheng	60	15	30.0	0.98	0.28	0.57

Supplementary Note: Computational cost

The dominant cost of the selection is the number of model simulations: the relative sensitivity matrix J by central differences needs $2p$ simulations and the stage-2 ordering by finite effects p more, $3p$ in total; all of them are independent and can run in parallel. The greedy selection, the conditioning tests, backward elimination and the first-order e_{fit} use only J (linear algebra, negligible in comparison). The admissibility test adds, for each candidate subset, 15 scenario simulations and the re-estimations, whose cost depends on the number of retained parameters rather than on p. Supplementary Table S35 gives the measured time of one simulation of all measured outputs (all experimental conditions) at θ_0 and the resulting time for the $3p$ simulations, in series and on four cores, for the problems timed.

Supplementary Table S35. Computational cost of the sensitivity computations: estimated parameters p, outputs used, time of one simulation, number of simulations for J and the stage-2 ordering ($3p$), and total time in series and on four cores.

Problem	p	Outputs	t_{sim} (s)	$3p$	Series	4 cores
Boehm	9	measured	0.00168	27	0 s	0 s
Elowitz	21	measured	0.00243	63	0 s	0 s
Fiedler	22	measured	0.00257	66	0 s	0 s
Smith	25	measured	0.346	75	26 s	6 s
Weber	36	measured	0.00374	108	0 s	0 s
Raia	39	measured	0.0106	117	1 s	0 s
Alkan	44	measured	0.13	132	17 s	4 s
Zheng	46	measured	0.00449	138	1 s	0 s
Isensee	46	measured	0.64	138	88 s	22 s
Giordano	50	measured	0.00613	150	1 s	0 s
Beer	72	measured	0.173	216	37 s	9 s
Lucarelli	84	measured	0.04	252	10 s	3 s
Bachmann	113	measured	0.04	339	14 s	3 s
Raimundez	136	measured	0.37	408	3 min	38 s
Chen	155	measured	35	465	4.5 h	68 min
Lang	294	measured	0.65	882	10 min	2 min
Froehlich	4,231	states	2.93	12,693	10.3 h	2.6 h

The time of Smith was measured later on a different machine and scaled by the ratio of the times of Boehm and Weber measured on both machines (factor 1.55). For Froehlich (4,231 parameters, 1,228 species) the analysis was carried out on the state trajectories; on its measured outputs each evaluation requires about 9,000 simulations (one per experimental condition), which was infeasible with the resources available.

Time of the complete method and comparison with profile likelihood

The table above covers only the sensitivity computations. To measure the cost of the complete method, including the admissibility test (15 scenarios with re-estimation for every candidate subset) and backward elimination, it was run from scratch, without any cached result, for all 34 problems, once with the

sensitivity computations distributed over 16–64 cores and once on a single core per problem (Supplementary Table S36). The complete method took between 2 s and 33 min per problem (median 19 s) in parallel. For the six problems selected with the first-order e_{fit}, these times do not include the nonlinear re-estimation of the final subset, which took from 4 min (Lucarelli) to 12.8 h (Chen) of processor time (Supplementary Table S42). 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. Its cost is not proportional to p: it depends mostly on the simulation time and on the number of parameters re-estimated in the admissibility test, so that Smith (25 parameters, slow simulations, nine calibration parameters) took 19 min and Lang (294 parameters) about 3 min. Profile likelihood, with all the profile points of a problem computed simultaneously on 120 cores, was much slower whenever simulations are expensive (Alkan 27 min against 25 s; Isensee more than 1 h for the first round, and about 5 h including the re-estimations of the optimum (110 h of processor time), against 10 min; Zheng 5 min against 14 s), took 70 min for Bachmann and 11 min for Lucarelli, the two admissible problems with more than 60 free parameters in which it was run (on 60 cores; 10 and 35 s for the proposed method), was not run for Chen and Raimundez, and was faster than the proposed method on small problems, for which the fixed cost of the admissibility test dominates; over the 28 problems with both times the proposed method was faster than the first round of profile likelihood in 12 (in 13 when all rounds are counted). All runs shared a 128-core server with other computations, so the times may be inflated by resource contention; main-text Fig. 4d plots the two times against each other.

Supplementary Table S36. Computation time (seconds of wall-clock time) of the complete proposed method (selection, admissibility test and backward elimination, run from scratch without cached results) with the sensitivity computations in parallel and on a single core, and of profile likelihood with all profile points of a problem computed in parallel (120 cores; 60 cores for Bachmann and Lucarelli, whose more than 60 free parameters were profiled in a separate run) (first round, and including the re-estimations of the optimum). Profile likelihood was not run for three reasons, given in the last column: not run, because of its computational cost (Chen, Raimundez; Supplementary Note *Profile likelihood*); no split, no training/test split on which to select (Blasi); not admissible, the proposed method found no admissible subset, so there was nothing to compare. The proposed method was faster in 12 of the 28 problems with both times. All runs shared a 128-core server with other computations, so the times may be inflated by resource contention. For the eight problems treated to first order the parallel time is that of the run that reproduces the reported results (Supplementary Note *Reproducibility package*).

Problem	p	Proposed, parallel	Proposed, 1 core	Profile, 1st round	Profile, total	
Alkan	44	25.4	46.0	1,645	1,645	
Armistead	14	6.8	5.9	2.7	2.7	
Bachmann	113	9.7	27.2	4,225	4,225	
Beer	72	197	309	–	–	not admissible
Bertozzi	8	3.2	4.3	0.5	0.5	
Blasi	9	4.5	4.5	–	–	no split
Boehm	9	4.5	3.4	0.2	0.2	
Borghans	23	17.9	37.1	30.7	124	
Brannmark	22	27.1	37.1	13.2	13.2	
Bruno	13	23.2	33.8	2.3	2.3	
Chen	155	1,959	–	–	–	not run
Crauste	12	36.4	51.1	–	–	not admissible
Elowitz	21	11.5	8.7	9.6	9.6	
Fiedler	22	17.0	12.1	0.6	0.6	
Fujita	19	77.8	78.8	25.9	25.9	
Giordano	50	12.5	16.9	123	123	
Isensee	46	584	1,415	4,453	18,861	
Lang	294	199	–	–	–	not admissible
Laske	13	26.9	28.1	5.3	5.3	
Liu	9	4.2	6.3	1.1	1.1	

(continued)

Problem	p	Proposed, parallel	Proposed, 1 core	Profile, 1st round	Profile, total	
Lucarelli	84	34.7	75.1	651	651	
Okuonghae	16	3.4	5.7	6.8	23.0	
Oliveira	12	5.0	10.9	5.1	5.1	
Perelson	3	2.1	2.7	0.4	0.4	
Rahman	9	2.9	4.2	2.1	2.1	
Raia	39	30.5	49.2	14.1	211	
Raimundez	136	120	306	–	–	not run
Salazar-Cavazos	6	67.5	37.6	13.9	13.9	
Schwen	30	92.9	147	110	110	
Smith	25	1,144	1,698	854	854	
Sneyd	15	20.7	17.6	4.0	4.0	
Weber	36	17.6	23.1	120	120	
Zhao	28	7.5	21.8	20.5	20.5	
Zheng	46	13.9	17.0	287	287	

Supplementary Note: Computing time of the analyses

This note reports the computing time of every analysis of the paper, problem by problem, with the purpose of showing what it costs to select and verify a subset with the method, how that cost compares with the standard identifiability analysis, profile likelihood of the full model, and where the computing time of the whole study was spent.

How the times were measured

Each program records, for every problem, the elapsed time of its computation (field `segundos`, i.e. seconds, of the result files of the reproducibility package). Several analyses use parallelism within a problem: the selection distributes the columns of J and the stage-2 ordering over 16 to 64 processes; the nested validation distributes the finite-difference Jacobian of each fit of the full model; profile likelihood of the full model computes all points of a problem in parallel on up to 120 cores. Other analyses consist of many independent single-core tasks (the seeds of the synthetic experiment, the scenarios and seeds of the admissibility test of the other criteria, the points of the identifiability profiles); for them the tables give the sum over the tasks, which is processor time and is larger than the elapsed time. The problems with expensive simulations were computed on shared servers with 128 and 256 cores and the others on an 8-core workstation; the times on shared servers may be inflated by other computations that ran at the same time. The complete method was timed in a run from scratch, without cached sensitivities (Supplementary Table S36 compares it with profile likelihood). Profile likelihood of the full model is given both as elapsed time, with the points of a problem in parallel, and as processor time, the sum over its points; only the processor time is added to the totals. The per-problem tables give each time in seconds, minutes or hours, whichever is most readable; the totals are given in hours. For the six problems selected with the first-order e_{fit}, the nonlinear re-estimation of the final subset is listed separately (Supplementary Table S42), because it is not part of the timed selection.

What the times show

Selecting a subset is affordable. The complete method took a median of 19 s per problem and less than ten minutes in 32 of the 34 problems; the slowest were Chen (33 min, 35 s per simulation), Smith (19 min) and Isensee (10 min). On one core the same computation took at most about three times longer (Supplementary Table S39). Where the selection used the first-order e_{fit}, verifying the final subset by nonlinear re-estimation was considerably more expensive than the selection itself: 12.8 h of processor time for Chen, 4.1 h for Raimundez and 3.9 h for Isensee, against 33, 2 and 10 min for the

selection.

Verifying it is affordable too. Determining the identifiability of the selected parameters (one fit of the reduced problem and the re-estimations at the profile points) took a median of 34 s per problem, and it was feasible in all 31 admissible problems, including those in which profile likelihood of the full model took hours or was not run (Supplementary Table S40). Profile likelihood of the full model, the standard identifiability analysis, took a median of 17 s of elapsed time in the 28 problems in which it was run (Supplementary Table S41), so on small problems it is as fast as the proposed method, which was faster than its first round in 12 of these 28 problems; it took 70 min for Bachmann and 11 min for Lucarelli (40 and 9 h of processor time) and 5.2 h for Isensee (110 h of processor time), and it was not run for Chen and Raimundez, for Blasi, which has no training/test split, or for the three problems without an admissible subset (Supplementary Table S36), in which the method and the identifiability check of the reduced problem still ran in minutes to hours.

Where the computing time went. All analyses reported in the paper together took about 697 hours of processor time (Supplementary Table S37, in hours and grouped by purpose; Supplementary Table S38, per analysis). The method itself cost about 23 h for the 34 problems: 1.3 h for the selection and 21.3 h for the nonlinear verification of the final subset, almost all of it in Chen, Raimundez and Isensee. The remaining ∼674 h were spent on the evaluation of the method (248 h, including 5.7 h for the test of the minimum subset size) and on the comparison with other criteria (426 h, including 41 h for the identifiability profiles of the subsets of the other criteria and 17 h for the admissibility test with the calibration parameters fixed), which were carried out for this paper and are not needed to apply the method. This is the sum over all cores of the times recorded by each task in the final runs, whose results are those of the reproducibility package; 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 to 256 cores and on an 8-core workstation, the elapsed time was much shorter than the processor time. Most of the time was spent in analyses that re-estimate models many times and that are not part of the method: profile likelihood of the full model (193 h, 110 h of it for Isensee), the identifiability profiles (26 h, and 102 h more for the check with 20 points per side), the synthetic experiments (44 h over the three amplitudes and 86 h for the log-uniform and sparse truths of all criteria), the tests of the subsets of all criteria (86 h), the variants of the method settings (50 h, because each variant reruns the complete method) and the nonlinear check of the first-order problems (21 h). The check of the identifiability hypotheses over the admissible set did not record times per problem and is not included; it took about 3 h of elapsed time on an 88-core server. Five problems with slow simulations or many parameters, Isensee, Chen, Raimundez, Smith and Bachmann, account for 87% of the total, whereas the median over the problems of the time of all analyses together is about one hour.

Supplementary Table S37. Processor time of all the analyses reported in the paper, in hours, grouped by purpose. Processor time is the sum over all cores of the times recorded by each task in the final runs (Supplementary Tables S39–S42); because the analyses ran in parallel on servers with 88 to 256 cores and on an 8-core workstation, the elapsed time was much shorter. Only the final runs are counted; the compilation of the Lean proofs and the check of the identifiability hypotheses over the admissible set (about 3 h of elapsed time on an 88-core server, not recorded per task) are not included.

Analysis	Hours	Share
The method itself		
Selection of the subset (34 problems, sensitivities in parallel)	1.3	0.2%
Nonlinear verification of the final subset (six problems selected to first order)	21.3	3.1%
Subtotal	*22.7*	*3.3%*
Evaluation of the method		
Selection repeated with fixed scenarios and on the training outputs	4.8	0.7%
Sensitivity to the method settings (15 variants)	50.3	7.2%
Robustness to the size of parameter changes (±1% to ±50%, paired and independent directions)	5.2	0.7%

Analysis	Hours	Share
Single-parameter subsets	3.2	0.5%
Validation on real data (external and nested)	2.3	0.3%
Synthetic experiments (±5%, ±20%, ±50%) and true signal	45.5	6.5%
Identifiability of the selected parameters: profiles (10 points per side)	26.0	3.7%
Identifiability: profiles with 20 points per side	102.4	14.7%
Identifiability: subsets selected on the training outputs	3.0	0.4%
Minimum subset size: best single parameters through robustness, admissibility at ±20%, synthetic and identifiability tests	5.7	0.8%
Subtotal	*248.3*	*35.6%*
Comparison with other criteria		
Profile likelihood of the full model (processor time)	193.0	27.7%
Admissibility and synthetic tests of the subsets of all criteria (±5% and ±20%, both scenario sets)	86.5	12.4%
Synthetic comparison with log-uniform and sparse truths (all criteria)	86.3	12.4%
Admissibility at ±5% with the calibration parameters fixed (14 problems, all subsets)	16.7	2.4%
Identifiability of the subsets of Brun, Morris, pivoted QR and profile likelihood (profiles)	40.8	5.8%
Selection by pivoted QR (both variants), Brun, Morris, random subsets and calibration only	2.8	0.4%
Subtotal	*426.1*	*61.1%*
All analyses	**697.1**	100%

Supplementary Table S38. Summary of the computing time per analysis over the problems of Supplementary Tables S39–S42: number of problems, median and maximum time per problem and total time, in processor time where an analysis consists of independent tasks; the elapsed time of profile likelihood and the selection on one core are listed but not added to the total.

Analysis	Problems	Median	Maximum (problem)	Total (h)
Selection, parallel	34	19 s	32.6 min (Chen)	1.3
Selection, 1 core	32	25 s	28.3 min (Smith)	1.3
Settings (15 variants)	28	2.6 min	25.8 h (Raimundez)	50.3
Robustness, paired	31	34 s	22.1 min (Smith)	1.4
Robustness, independent	31	27 s	1.7 h (Smith)	3.8
Single parameter	27	32 s	1.6 h (Smith)	3.2
External validation	29	7 s	34.0 min (Smith)	1.9
Nested validation	29	<1 s	11.0 min (Chen)	0.4
Synthetic, ±20%, 20 seeds	29	2.8 min	7.3 h (Chen)	22.3
Synthetic, ±5%, 10 seeds	29	1.6 min	2.8 h (Chen)	8.6
Synthetic, ±50%, 10 seeds	29	2.6 min	4.1 h (Chen)	13.1
True signal, 3 amplitudes	29	31 s	54.8 min (Chen)	1.5
Identifiability of S	31	34 s	14.8 h (Chen)	26.0
Pivoted QR	28	<1 s	1.9 min (Isensee)	0.0
Brun and Morris	29	10 s	29.6 min (Isensee)	0.8
Random subsets (20)	29	4 s	1.2 h (Isensee)	1.6
Calibration only	28	<1 s	4.6 min (Isensee)	0.1
Criteria test, ±5%	29	1.8 min	13.8 h (Isensee)	23.2
Criteria test, ±20%	29	33 s	1.4 h (Smith)	2.3
Profile likelihood, full model (elapsed)	28	17 s	5.2 h (Isensee)	7.6
Profile likelihood, full model (processor)	27	5.5 min	110.1 h (Isensee)	193.0
Nonlinear check of S	6	2.2 h	12.7 h (Chen)	21.3
Selection, fixed scen.	34	18 s	1.2 h (Chen)	2.4
Selection, training	34	16 s	1.2 h (Chen)	2.4
Pivoted QR (GKS)	29	<1 s	9.2 min (Isensee)	0.3
Criteria, ±5%, fixed	29	5.3 min	16.7 h (Smith)	39.6
Criteria, ±20%, fixed	29	3.4 min	13.7 h (Smith)	21.3
Synthetic, log-uniform	29	2.2 min	22.1 h (Raimundez)	48.5
Synthetic, sparse	29	1.8 min	17.3 h (Raimundez)	37.8
Profiles, 20 per side	31	47 s	64.0 h (Chen)	102.4
Profiles, training S	17	36 s	1.6 h (Bachmann)	3.0
Profiles, other criteria	28	55 s	25.2 h (Isensee)	40.8
Test of $k_{\min} = 2$	15	3.3 min	4.7 h (Smith)	5.7

Analysis	Problems	Median	Maximum (problem)	Total (h)
Calibration fixed	14	8.7 min	10.6 h (Smith)	16.7
All analyses				697

Supplementary Table S39. Computing time of the method and of the analyses of its behaviour, for every problem. Selection: complete method (sensitivities, stages 1 and 2, admissibility test and backward elimination) run from scratch, with the sensitivities in parallel or on one core; Settings: the 15 variants of Supplementary Note *Sensitivity to the method settings* together; Robustness: seven perturbation sizes with paired or independent directions; Single parameter: Supplementary Note *Single-parameter subsets*. Times recorded by each program for each problem: elapsed time of the computation of that problem with the parallelism described in the text; where an analysis consists of independent single-core tasks (seeds, scenarios, profile points) the time is their sum, i.e. processor time. –, not computed (problem not admissible, not applicable or not run). Large problems were computed on shared servers with 128 or 256 cores, the others on an 8-core workstation; times on shared servers may be inflated by resource contention.

Problem	*Selection, parallel*	*Selection, 1 core*	*Settings (15 variants)*	*Robustness, paired*	*Robustness, independent*	*Single parameter*
Alkan	25 s	46 s	–	47 s	32 s	1.0 h
Armistead	7 s	6 s	56 s	10 s	6 s	4 s
Bachmann	10 s	27 s	–	34 s	16 s	–
Beer	3.3 min	5.1 min	–	–	–	–
Bertozzi	3 s	4 s	12 s	4 s	3 s	2 s
Blasi	4 s	4 s	1.3 min	11 s	9 s	8 s
Boehm	4 s	3 s	20 s	11 s	7 s	6 s
Borghans	18 s	37 s	2.6 min	1.7 min	1.1 min	3.4 min
Brannmark	27 s	37 s	3.5 min	2.1 min	1.6 min	2.2 min
Bruno	23 s	34 s	4.9 min	22 s	16 s	8 s
Chen	32.6 min	–	–	7.0 min	1.4 h	–
Crauste	36 s	51 s	3.3 min	–	–	–
Elowitz	12 s	9 s	2.2 min	39 s	28 s	32 s
Fiedler	17 s	12 s	1.9 min	54 s	38 s	1.1 min
Fujita	1.3 min	1.3 min	26.5 min	9.6 min	3.2 min	3.0 min
Giordano	12 s	17 s	1.0 min	30 s	20 s	44 s
Isensee	9.7 min	23.6 min	–	13.5 min	11.5 min	–
Lang	3.3 min	–	18.8 h	–	–	–
Laske	27 s	28 s	3.8 min	4.6 min	1.4 min	30 s
Liu	4 s	6 s	39 s	46 s	12 s	12 s
Lucarelli	35 s	1.3 min	–	46 s	11 s	3.9 min
Okuonghae	3 s	6 s	16 s	23 s	6 s	15 s
Oliveira	5 s	11 s	36 s	49 s	12 s	18 s
Perelson	2 s	3 s	2 s	5 s	1 s	<1 s
Rahman	3 s	4 s	11 s	14 s	5 s	5 s
Raia	30 s	49 s	4.9 min	<1 s	2.2 min	2.3 min
Raimundez	2.0 min	5.1 min	25.8 h	2.6 min	2.7 min	–
Salazar-Cavazos	1.1 min	38 s	5.0 min	3 s	1.5 min	1.8 min
Schwen	1.5 min	2.5 min	24.6 min	10.7 min	8.1 min	14.3 min
Smith	19.1 min	28.3 min	4.0 h	22.1 min	1.7 h	1.6 h
Sneyd	21 s	18 s	2.0 min	<1 s	27 s	36 s
Weber	18 s	23 s	2.8 min	32 s	1.2 min	2.8 min
Zhao	8 s	22 s	2.5 min	21 s	17 s	12 s
Zheng	14 s	17 s	5.5 min	22 s	16 s	21 s

Supplementary Table S40. Computing time of the validation and identifiability analyses, for every problem: external and nested validation on real data, synthetic experiment (20 seeds at ±20%, 10 seeds at ±5% and ±50%), true signal $y(\theta^*) - y(\theta_0)$ at the three amplitudes, and the identifiability of the selected parameters (fit of the reduced problem and all profile points). Times recorded by each program for each problem: elapsed time of the computation of that problem with the parallelism described in the text; where an analysis consists of independent single-core tasks (seeds, scenarios, profile points) the time is their sum, i.e. processor time. –, not computed (problem not admissible, not applicable or not run). Large problems were computed on shared servers with 128 or 256 cores, the others on an 8-core workstation; times on shared servers may be inflated by resource contention.

Problem	*External validation*	*Nested validation*	*Synthetic, ±20%, 20 seeds*	*Synthetic, ±5%, 10 seeds*	*Synthetic, ±50%, 10 seeds*	*True signal, 3 amplitudes*	*Identifiability of S*
Alkan	5.3 min	6 s	59.2 min	23.6 min	33.7 min	1.5 min	21.6 min
Armistead	<1 s	<1 s	58 s	10 s	11 s	26 s	<1 s
Bachmann	4.8 min	1.6 min	1.6 h	34.3 min	53.5 min	58 s	35.2 min
Beer	–	–	–	–	–	–	–
Bertozzi	–	–	–	–	–	–	1 s
Blasi	–	–	–	–	–	–	5 s
Boehm	<1 s	<1 s	24 s	12 s	10 s	27 s	4 s
Borghans	6 s	<1 s	1.4 min	1.1 min	2.6 min	26 s	34 s
Brannmark	8 s	<1 s	2.8 min	1.6 min	1.6 min	29 s	58 s
Bruno	<1 s	<1 s	18 s	8 s	9 s	25 s	19 s
Chen	31.3 min	11.0 min	7.3 h	2.8 h	4.1 h	54.8 min	14.8 h
Crauste	–	–	–	–	–	–	–
Elowitz	6 s	<1 s	1.4 min	39 s	42 s	26 s	22 s
Fiedler	5 s	<1 s	2.7 min	1.1 min	1.3 min	43 s	35 s
Fujita	7 s	<1 s	1.9 min	1.0 min	1.5 min	30 s	1.1 min
Giordano	28 s	<1 s	8.4 min	5.0 min	3.7 min	43 s	6 s
Isensee	27.4 min	10.6 min	5.0 h	1.7 h	2.9 h	13.2 min	4.3 h
Lang	–	–	–	–	–	–	–
Laske	3 s	<1 s	1.4 min	42 s	41 s	43 s	1.2 min
Liu	<1 s	<1 s	27 s	12 s	10 s	27 s	7 s
Lucarelli	3.2 min	49 s	51.7 min	23.6 min	29.2 min	1.2 min	54.9 min
Okuonghae	7 s	<1 s	1.8 min	54 s	49 s	25 s	4 s
Oliveira	2 s	<1 s	46 s	20 s	20 s	28 s	6 s
Perelson	<1 s	<1 s	12 s	6 s	5 s	22 s	<1 s
Rahman	<1 s	<1 s	30 s	10 s	12 s	25 s	1 s
Raia	18 s	<1 s	3.2 min	1.6 min	2.6 min	31 s	1.4 min
Raimundez	3 s	1.2 min	57.9 min	37.0 min	1.3 h	2.7 min	4.0 h
Salazar-Cavazos	21 s	2 s	4.7 min	2.1 min	3.0 min	2.5 min	40 s
Schwen	39 s	2 s	12.5 min	7.4 min	9.7 min	34 s	6.2 min
Smith	34.0 min	19 s	4.0 h	1.3 h	1.7 h	3.8 min	48.1 min
Sneyd	2 s	<1 s	43 s	19 s	34 s	29 s	20 s
Weber	1.1 min	<1 s	13.4 min	5.9 min	9.2 min	31 s	37 s
Zhao	22 s	<1 s	5.7 min	3.1 min	4.0 min	27 s	9 s
Zheng	1.3 min	<1 s	27.4 min	10.7 min	21.3 min	33 s	9 s

Supplementary Table S41. Computing time of the comparison with other criteria, for every problem: selection and validation by pivoted QR, Brun and Morris, the random and calibration-only baselines, the admissibility and synthetic tests of the subsets of all criteria (±5%, including the synthetic seeds, and ±20%), the admissibility test at ±5% with the calibration parameters fixed (all 27 subsets of each of the 14 problems with calibration parameters), and profile likelihood of every biological parameter of the full model (all rounds), as elapsed time with the profile points in parallel and as processor time, the sum over the points. Times recorded by each program for each problem: elapsed time of the computation of that problem with the parallelism described in the text; where an analysis consists of independent single-core tasks (seeds, scenarios, profile points) the time is their sum, i.e. processor time. –, not computed (problem not admissible, not applicable or not run). Large problems were computed on shared servers with 128 or 256 cores, the others on an 8-core workstation; times on shared servers may be inflated by resource contention.

Problem	*Pivoted QR*	*Brun and Morris*	*Random subsets (20)*	*Calibration only*	*Criteria test, ±5%*	*Criteria test, ±20%*	*Calibration fixed*	*Profile likelihood, full model (elapsed)*	*Profile likelihood, full model (processor)*
Alkan	4 s	1.9 min	2.4 min	5 s	28.6 min	50 s	5.3 min	27.4 min	13.1 h
Armistead	<1 s	<1 s	<1 s	<1 s	28 s	14 s	–	3 s	17 s
Bachmann	1 s	1.8 min	10.1 min	19 s	1.5 h	3.8 min	5.7 min	1.2 h	40.3 h
Beer	–	–	–	–	–	–	–	–	–
Bertozzi	<1 s	2 s	<1 s	<1 s	16 s	6 s	–	<1 s	<1 s
Blasi	–	–	–	–	–	–	–	–	–
Boehm	<1 s	3 s	<1 s	<1 s	35 s	10 s	–	<1 s	8 s
Borghans	<1 s	6 s	12 s	<1 s	3.0 min	1.3 min	12.5 min	2.1 min	13.4 min
Brannmark	<1 s	10 s	5 s	<1 s	4.3 min	2.1 min	9.2 min	13 s	5.1 min
Bruno	<1 s	3 s	<1 s	<1 s	39 s	13 s	–	2 s	26 s
Chen	–	–	–	–	–	–	–	–	–
Crauste	–	–	–	–	–	–	–	–	–
Elowitz	<1 s	4 s	2 s	<1 s	1.1 min	23 s	2.1 min	10 s	5.5 min
Fiedler	<1 s	6 s	4 s	<1 s	1.8 min	34 s	3.6 min	<1 s	1.8 min
Fujita	<1 s	9 s	6 s	<1 s	10.9 min	2.9 min	15.2 min	26 s	10.3 min
Giordano	<1 s	12 s	<1 s	<1 s	1.3 min	23 s	–	2.1 min	1.1 h
Isensee	1.9 min	29.6 min	1.2 h	4.6 min	13.8 h	25.5 min	3.1 h	5.2 h	110.1 h
Lang	–	–	–	–	–	–	–	–	–
Laske	<1 s	11 s	6 s	<1 s	3.7 min	1.2 min	7.7 min	5 s	55 s
Liu	<1 s	3 s	<1 s	<1 s	35 s	11 s	–	1 s	7 s
Lucarelli	3 s	1.4 min	52 s	<1 s	11.1 min	33 s	–	10.9 min	8.7 h
Okuonghae	<1 s	10 s	<1 s	<1 s	34 s	9 s	–	23 s	1.9 min
Oliveira	<1 s	2 s	1 s	<1 s	40 s	12 s	–	5 s	1.4 min
Perelson	<1 s	2 s	<1 s	<1 s	6 s	1 s	–	<1 s	<1 s
Rahman	<1 s	3 s	<1 s	<1 s	16 s	4 s	–	2 s	26 s
Raia	<1 s	11 s	11 s	<1 s	4.0 min	1.8 min	8.2 min	3.5 min	10.9 min
Raimundez	18 s	7.5 min	10 s	2 s	2.9 h	5.4 min	1.0 h	–	–
Salazar-Cavazos	2 s	12 s	12 s	2 s	8.4 min	1.6 min	–	14 s	2.6 min
Schwen	<1 s	28 s	30 s	2 s	19.4 min	6.5 min	46.7 min	1.8 min	55.0 min
Smith	5 s	5.3 min	8.0 min	14 s	3.1 h	1.4 h	10.6 h	14.2 min	8.8 h
Sneyd	<1 s	3 s	2 s	<1 s	1.3 min	25 s	–	4 s	2.2 min
Weber	<1 s	10 s	4 s	<1 s	2.4 min	51 s	4.7 min	2.0 min	3.0 h
Zhao	<1 s	5 s	2 s	<1 s	44 s	13 s	–	20 s	12.9 min
Zheng	<1 s	14 s	2 s	<1 s	1.0 min	15 s	–	4.8 min	5.8 h

Supplementary Table S42. Computing time of the additional analyses, for every problem: nonlinear re-estimation of the final subset in the 15 scenarios for the problems selected with the first-order e_{fit}; the complete selection repeated with the fixed scenario set and on the training outputs; standard pivoted QR; the admissibility tests of all criteria with the fixed scenario set at ±5% (including the synthetic seeds) and ±20%; the synthetic comparison with log-uniform and sparse truths (all criteria, ten seeds); and the identifiability profiles with 20 points per side and of the subsets selected on the training outputs; the identifiability profiles of the subsets of Brun, Morris, pivoted QR and profile likelihood (sum over the four criteria); and the test of the minimum subset size $k_{\min} = 2$ (all its tests for the best single parameters; for Smith and Schwen the robustness was split into one process per perturbation size and its time is taken from the checkpoint files, `analysis/kmin_times.py`). The check of the identifiability hypotheses over the admissible set (512 Sobol points and adversarial search for five radii) did not record times per problem and is not included; it took about 3 h of elapsed time on an 88-core server. Times recorded by each program for each problem: elapsed time of the computation of that problem with the parallelism described in the text; where an analysis consists of independent single-core tasks (seeds, scenarios, profile points) the time is their sum, i.e. processor time. –, not computed (problem not admissible, not applicable or not run). Large problems were computed on shared servers with 128 or 256 cores, the others on an 8-core workstation; times on shared servers may be inflated by resource contention.

Problem	*Nonlinear check of S*	*Selection, fixed scen.*	*Selection, training*	*Pivoted QR (GKS)*	*Criteria, ±5%, fixed*	*Criteria, ±20%, fixed*	*Synthetic, log-uniform*	*Synthetic, sparse*	*Profiles, 20 per side*	*Profiles, training S*	*Profiles, other criteria*	*Test of $k_{\min} = 2$*
Alkan	8.1 min	27 s	29 s	14 s	26.9 min	6.2 min	1.1 h	51.8 min	1.2 h	38.9 min	45.4 min	8.8 min
Armistead	–	2 s	3 s	<1 s	45 s	41 s	13 s	10 s	6 s	–	<1 s	–
Bachmann	23.2 min	10 s	19 s	35 s	1.2 h	5.8 min	3.3 h	3.7 h	49.5 min	1.6 h	1.5 h	–
Beer	–	2.7 min	3.1 min	–	–	–	–	–	–	–	–	–
Bertozzi	–	2 s	1 s	<1 s	24 s	22 s	11 s	9 s	5 s	–	1 s	–
Blasi	–	4 s	2 s	–	–	–	–	–	12 s	–	–	–
Boehm	–	3 s	2 s	1 s	1.1 min	54 s	32 s	29 s	11 s	–	4 s	–
Borghans	–	19 s	18 s	1 s	13.9 min	10.5 min	3.4 min	2.5 min	1.8 min	1.0 min	1.4 min	3.3 min
Brannmark	–	18 s	1.4 min	1 s	12.3 min	11.0 min	2.5 min	2.1 min	2.5 min	1.3 min	2.0 min	4.3 min
Bruno	–	18 s	14 s	<1 s	1.9 min	1.7 min	46 s	36 s	43 s	24 s	49 s	–
Chen	12.7 h	1.2 h	1.2 h	–	–	–	–	–	64.0 h	–	–	–
Crauste	–	25 s	9 s	–	–	–	–	–	–	–	–	–
Elowitz	–	5 s	9 s	<1 s	3.3 min	3.1 min	2.2 min	1.6 min	47 s	–	24 s	–
Fiedler	–	7 s	8 s	<1 s	5.3 min	5.0 min	2.9 min	2.6 min	1.1 min	36 s	1.1 min	1.9 min
Fujita	–	1.0 min	1.1 min	<1 s	25.5 min	13.3 min	3.3 min	4.0 min	2.6 min	–	1.2 min	2.5 min
Giordano	–	5 s	5 s	1 s	3.1 min	2.6 min	1.9 min	1.8 min	14 s	8 s	10 s	1.0 min
Isensee	3.9 h	12.8 min	11.8 min	9.2 min	11.4 h	3.7 h	18.5 h	12.8 h	23.9 h	–	25.2 h	–
Lang	–	10.7 min	8.2 min	–	–	–	–	–	–	–	–	–
Laske	–	21 s	2.4 min	<1 s	10.1 min	9.0 min	4.2 min	3.6 min	51 s	1.3 min	2.7 min	4.1 min
Liu	–	6 s	3 s	<1 s	1.8 min	1.9 min	31 s	32 s	5 s	10 s	19 s	–
Lucarelli	3.8 min	35 s	40 s	7 s	7.3 min	2.0 min	28.2 min	22.9 min	43.8 min	37.4 min	1.6 h	–
Okuonghae	–	3 s	3 s	<1 s	53 s	54 s	51 s	45 s	<1 s	2 s	2 s	15 s
Oliveira	–	6 s	6 s	<1 s	1.5 min	1.3 min	55 s	47 s	3 s	4 s	6 s	26 s
Perelson	–	1 s	<1 s	<1 s	5 s	12 s	17 s	15 s	<1 s	–	–	–
Rahman	–	2 s	1 s	<1 s	34 s	31 s	28 s	28 s	<1 s	<1 s	<1 s	15 s
Raia	–	38 s	58 s	2 s	12.8 min	12.1 min	5.8 min	4.1 min	54 s	44 s	1.4 min	3.7 min
Raimundez	4.1 h	2.5 min	2.2 min	5.1 min	6.5 h	1.0 h	22.1 h	17.3 h	9.2 h	–	9.0 h	–
Salazar-Cavazos	–	39 s	29 s	6 s	17.9 min	13.4 min	9.9 min	10.6 min	33 s	–	1.0 min	4.2 min
Schwen	–	2.7 min	3.4 min	5 s	1.1 h	59.9 min	19.4 min	15.7 min	4.1 min	–	6.6 min	24.6 min
Smith	–	31.1 min	35.5 min	1.4 min	16.7 h	13.7 h	1.9 h	1.8 h	2.4 h	–	2.4 h	4.7 h
Sneyd	–	12 s	13 s	<1 s	3.9 min	3.4 min	56 s	51 s	29 s	–	20 s	–
Weber	–	25 s	35 s	<1 s	6.1 min	6.3 min	1.9 min	1.8 min	1.1 min	36 s	40 s	2.5 min
Zhao	–	18 s	6 s	<1 s	1.9 min	1.8 min	48 s	48 s	11 s	3 s	13 s	–
Zheng	–	5 s	10 s	<1 s	2.4 min	1.8 min	1.3 min	1.3 min	15 s	8 s	16 s	–

Supplementary Note: System-by-system summary

This note summarises, for each of the 34 benchmark problems, what the model describes, which parameters the method selected and what they mean in the model, how well the reduced model reproduces the full one, and how it predicted held-out data. The meaning of each selected parameter was taken from the reactions and rules of the SBML model in which it appears (following assignment rules and condition-specific parameters to the reactions they affect). The numbers are those of Table 1 of the main text and Supplementary Table S1: p is the number of estimated parameters, κ and VIF are the condition number and largest variance inflation factor of the selected sensitivity columns, R_{var} is the fraction of sensitivity energy kept, $\cos\Delta$ and e_{fit} are medians over 15 random scenarios at $\pm 5\%$ (threshold $\tau_e = 0.436$), the level of guarantee is defined in the Results of the main text, and the robustness patterns A–D are those of Fig. 5 of the main text. In sloppy models the selected subset is often one of several admissible ones (with the training outputs, 20 of the 29 problems with an admissible subset chose a different one), so the descriptions below give the role of the selected parameters in the model, not a unique biological finding. In the external validation, both the reduced and the full model are refitted to the training data, and the test χ^2/n (mean squared residual divided by the recorded noise variance; about 1 for a model that predicts within the noise) is computed on the held-out data. The published estimate
$theta_0$ was obtained with all the data, including the test data, so its test
chi^2/n is a reference rather than a competing prediction.

Alkan_SciSignal2018

Cell-cycle progression, DNA double-strand-break damage and repair under chemotherapy (36 species). Source: `doi:10.1126/scisignal.aat0229`. The full model has 44 parameters, 10 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `kt` and `k_rep_nhej` (stage 1, ranking by sensitivity energy); the reduced problem has 73% fewer free parameters than the full one. `kt` is the cell-cycle transition rate, shared by the transitions G1→S, S→G2 and G2→division; `k_rep_nhej` is the rate of repair by non-homologous end joining, which returns damaged S and G2 cells to the cycle. For this subset $\kappa = 1.3$ and VIF 1.1 (no collinearity), $R_{\mathrm{var}} = 94\%$, median $\cos\Delta = 0.98$ and median $e_{\mathrm{fit}} = 12.0\%$ at $\pm 5\%$ after nonlinear re-estimation of the final subset (12.1% to first order during the selection): the reduced model reproduces the full one, with guarantee level 3 ($\cos\Delta \geq 0.9$ and $R_{\mathrm{var}} \geq 0.89$); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$ (sizes above $\pm 5\%$ evaluated with the first-order e_{fit}).

External validation on held-out experimental conditions (555 test and 1178 training measurements): test χ^2/n 2.75 for the reduced model, 3.37 for the full model refitted to the same data and 2.47 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The model was not verified in Lean because the automatic checks of the domain and quasi-positivity conditions fail.

Armistead_CellDeathDis2024

Ceramide and sphingolipid metabolism (4 species). Source: `doi:10.1038/s41419-024-07134-2`. The full model has 14 parameters. The method selected two biological parameters, `k_d` and `k00` (stage 2, ranking by finite-perturbation effects); the reduced problem has 86% fewer free parameters than the full one. `k_d` is the degradation rate of ceramide; `k00` is the basal rate of de novo production of sphinganine. For this subset $\kappa = 1.0$ and VIF 1.0 (no collinearity), $R_{\mathrm{var}} = 58\%$, median $\cos\Delta = 0.81$ and median $e_{\mathrm{fit}} = 33.0\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\mathrm{fit}} \leq \tau_e$ only); robustness pattern B: admissible up to at least $\pm 20\%$ but not for larger changes.

External validation on held-out experimental conditions (29 test and 29 training measurements): test χ^2/n 1.22 for the reduced model, 1.37 for the full model refitted to the same data and 1.16 for the

published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Bachmann_MSB2011

JAK2/STAT5 signalling induced by erythropoietin (Epo) with negative feedback by CIS and SOCS3 (25 species). Source: `doi:10.1038/msb.2011.50`. The full model has 113 parameters, 74 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `CISRNADelay` and `JAK2ActEpo` (stage 2, ranking by finite-perturbation effects); the reduced problem has 33% fewer free parameters than the full one. `JAK2ActEpo` is the rate of Epo-induced activation of JAK2 at the receptor (EpoR:JAK2 → EpoR:pJAK2); `CISRNADelay` is the transit rate through the chain of delay compartments of CIS mRNA, i.e. it sets the delay of the CIS feedback. For this subset $\kappa = 2.8$ and VIF 2.5 (no collinearity), $R_{\text{var}} = 36\%$, median $\cos\Delta = 0.85$ and median $e_{\text{fit}} = 33.8\%$ at $\pm 5\%$ after nonlinear re-estimation of the final subset (33.7% to first order during the selection): the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$ (sizes above $\pm 5\%$ evaluated with the first-order e_{fit}).

External validation on held-out experimental conditions (202 test and 339 training measurements): test χ^2/n 4.77 for the reduced model, 7.35 for the full model refitted to the same data and 1.05 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Beer_MolBioSystems2014

Growth of bacterial cultures and production of the pigment indigoidine by engineered synthetase variants (4 species, 19 conditions). Source: `doi:10.1039/c3mb70594c`. The full model has 72 parameters. **Not admissible.** The selection stopped at 18 parameters (synthesis, dimerisation and degradation rates of particular strains, growth rates β of particular experiments and the initial amount of bacteria), because any further parameter violated the conditioning test ($\kappa = 7.2$, VIF 9.998, just below the limit of 10). Re-estimating these 18 parameters gave a median e_{fit} of 50%, above the threshold of 43.6%; the outputs depend on many strain- and experiment-specific parameters, and no small subset can compensate for all of them. The Lean verification of this model is only conditional, because the growth check fails.

Bertozzi_PNAS2020

SIR model of the COVID-19 epidemic, fitted to two regions (3 species). Source: `doi:10.1073/pnas.2006520117`. The full model has 8 parameters. The method selected two biological parameters, `R0_NY` and `I0_NY` (stage 2, ranking by finite-perturbation effects); the reduced problem has 75% fewer free parameters than the full one. `R0_NY` is the basic reproduction number in New York, which sets the infection rate; `I0_NY` is the initial number of infected individuals in New York. For this subset $\kappa = 3.9$ and VIF 4.3 (no collinearity), $R_{\text{var}} = 78\%$, median $\cos\Delta = 0.92$ and median $e_{\text{fit}} = 32.9\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern B: admissible up to at least $\pm 20\%$ but not for larger changes.

External validation was not possible: both selected parameters belong only to the held-out condition, so they cannot be estimated from the training data. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Blasi_CellSystems2016

Acetylation of the four lysines of histone H4 (16 acetylation states). Source: `doi:10.1016/j.cels.2016.01.002`. The full model has 9 parameters. The method selected two biological parameters, `a_basal` and `a_k16_k12k16` (stage 1, ranking by sensitivity energy); backward elimination had removed `a_k12k16_k8k12k16`, `a_k5k12_k5k8k12` and `a_k8k12k16_4ac` from the initial selection; the reduced problem has 78% fewer free parameters than the full one. `a_basal` is the basal acetylation rate, from which most of the transition rates are derived by assignment rules; `a_k16_k12k16` is the rate of acetylation of K12 in H4 already acetylated at K16. For this subset $\kappa = 1.8$ and VIF 1.4 (no collinearity), $R_{\text{var}} = 79\%$, median $\cos\Delta = 0.91$ and median $e_{\text{fit}} = 40.2\%$ at $\pm 5\%$: the reduced model reproduces the full one, although only just within the threshold, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern B: admissible up to at least $\pm 20\%$ but not for larger changes.

External validation was not possible: the problem has a single condition and a single measurement time, so the data cannot be split. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Boehm_JProteomeRes2014

Phosphorylation, dimerisation and nuclear import of STAT5A and STAT5B (8 species). Source: `doi:10.1021/pr5006923`. The full model has 9 parameters. The method selected two biological parameters, `k_imp_hetero` and `k_phos` (stage 2, ranking by finite-perturbation effects); the reduced problem has 78% fewer free parameters than the full one. `k_phos` is the rate of phosphorylation and dimerisation of STAT5 (the three reactions forming pApA, pApB and pBpB); `k_imp_hetero` is the rate of nuclear import of the heterodimer pApB. For this subset $\kappa = 1.0$ and VIF 1.0 (no collinearity), $R_{\text{var}} = 68\%$, median $\cos\Delta = 0.93$ and median $e_{\text{fit}} = 9.4\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on later time points ($t \geq 80$; 18 test and 30 training measurements): test χ^2/n 2.57 for the reduced model, 11.22 for the full model refitted to the same data and 0.34 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data much better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Borghans_BiophysChem1997

Intracellular Ca^{2+} oscillations by Ca^{2+}-induced Ca^{2+} release (3 species). Source: `doi:10.1016/s0301-4622(97)00010-0`. The full model has 23 parameters, 2 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `Kz` and `K_par` (stage 2, ranking by finite-perturbation effects); the reduced problem has 83% fewer free parameters than the full one. `Kz` is the threshold constant of the release of Ca^{2+} from the Ca^{2+}-sensitive pool into the cytosol; `K_par` is the constant of the removal of cytosolic Ca^{2+}. For this subset $\kappa = 2.7$ and VIF 2.4 (no collinearity), $R_{\text{var}} = 21\%$, median $\cos\Delta = 0.86$ and median $e_{\text{fit}} = 4.3\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on later time points ($t \geq 6.48875$; 37 test and 74 training measurements): test χ^2/n 0.74 for the reduced model, 0.74 for the full model refitted to the same data and 0.74 for the published θ_0. The two refitted models predict equally well (test χ^2/n within 5%). The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Brannmark_JBC2010

Insulin signalling in adipocytes: insulin receptor and IRS1 phosphorylation and internalisation (9 species). Source: `doi:10.1074/jbc.M110.106849`. The full model has 22 parameters, 3 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `k1f` and `km2` (stage 2, ranking by finite-perturbation effects); the reduced problem has 77% fewer free parameters than the full one. `k1f` is the rate of dephosphorylation and recycling of the internalised phosphorylated receptor (IRiP $\rightarrow$ IRi); `km2` is the rate of dephosphorylation of IRS1 (IRSiP $\rightarrow$ IRS). For this subset $\kappa = 1.9$ and VIF 1.4 (no collinearity), $R_{\text{var}} = 32\%$, median $\cos\Delta = 0.85$ and median $e_{\text{fit}} = 7.8\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (26 test and 17 training measurements): test χ^2/n 5.96 for the reduced model, 6,277 for the full model refitted to the same data and 0.94 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data much better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Bruno_JExpBot2016

Enzymatic cleavage of carotenoids (β-carotene, β-cryptoxanthin, zeaxanthin) into apocarotenoids (7 species, 6 experiments). Source: `doi:10.1093/jxb/erw356`. The full model has 13 parameters. The method selected five biological parameters, `kb1`, `init_bcar2`, `init_zea_1`, `init_bcar1` and `init_bcry_1` (stage 1, ranking by sensitivity energy); backward elimination had removed `szea` and `init_b10_1` from the initial selection; the reduced problem has 62% fewer free parameters than the full one. `kb1` is the cleavage rate of β-carotene; the other four are initial amounts of the substrates in particular experiments: β-carotene (`init_bcar1`, `init_bcar2`), zeaxanthin (`init_zea_1`) and β-cryptoxanthin (`init_bcry_1`). For this subset $\kappa = 2.3$ and VIF 1.6 (no collinearity), $R_{\text{var}} = 75\%$, median $\cos\Delta = 0.90$ and median $e_{\text{fit}} = 37.8\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (28 test and 49 training measurements): test χ^2/n 0.91 for the reduced model, 208 for the full model refitted to the same data and 0.91 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data much better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Chen_MSB2009

ErbB receptor network (ErbB1–4) with the MAPK and PI3K/AKT cascades (500 species). Source: `doi:10.1038/msb.2008.74`. The full model has 155 parameters, 3 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `k106` and `k44` (stage 2, ranking by finite-perturbation effects); the reduced problem has 97% fewer free parameters than the full one. `k106` is the binding rate of PIP2 to the receptor complexes that carry PI3K; `k44` is the binding rate of MEK to phosphorylated Raf, the first step of MEK activation. For this subset $\kappa = 1.0$ and VIF 1.0 (no collinearity), $R_{\text{var}} = 25\%$, median $\cos\Delta = 0.90$ and median $e_{\text{fit}} = 8.1\%$ at $\pm 5\%$ after nonlinear re-estimation of the final subset (8.2% to first order during the selection): the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$ (sizes above $\pm 5\%$ evaluated with the first-order e_{fit}).

External validation on held-out experimental conditions (30 test and 90 training measurements): test χ^2/n 20.88 for the reduced model, 17.71 for the full model refitted to the same data and 15.22 for the published θ_0. Here the refitted full model predicts the test data better than the reduced model.

Even θ_0 has $\chi^2/n \gg 1$, which indicates that the noise levels recorded in the benchmark do not match the scatter of the data. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Crauste_CellSystems2017

CD8 T-cell response to infection: naive, early and late effector and memory cells and pathogen (5 species). Source: `doi:10.1016/j.cels.2017.01.014`. The full model has 12 parameters. **Not admissible.** The published parameters lie at the edge of a regime in which the solution grows without bound in finite time. With the published parameters the effector populations reach a peak of 9.6×10^5 cells on day 9.7 and then decline (5976 cells on day 28); after a change of only 1% in a single parameter (for example, a decrease of the proliferation rate of early effectors ρ_E) they keep growing and the solution breaks down between days 8.8 and 9.9, well within the measured horizon of 28 days. This happened for 9 of the 24 one-at-a-time perturbations of $\pm 1\%$ and for 7 of the 15 scenarios at $\pm 5\%$, for every integrator tolerance tested (relative tolerances from 10^{-10} to 10^{-5}), so the sensitivity matrix and the admissibility test are not defined and no local reduction method is applicable. Consistently, the existence of solutions could only be proved in Lean on a short horizon ($T \leq 1$, Supplementary Note *Verified benchmark models*).

Elowitz_Nature2000

The repressilator, a synthetic oscillator of three mutually repressing genes with a GFP reporter (8 species). Source: `doi:10.1038/35002125`. The full model has 21 parameters, 2 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `n_Hill` and `tau_mRNA` (stage 1, ranking by sensitivity energy); backward elimination had removed `tau_prot`, `tau_prot_GFP` and `eff` from the initial selection; the reduced problem has 81% fewer free parameters than the full one. `n_Hill` is the Hill coefficient of transcriptional repression, shared by the four transcription reactions; `tau_mRNA` is the mRNA lifetime, which enters the degradation of the three mRNAs and the translation rates. For this subset $\kappa = 2.9$ and VIF 2.6 (no collinearity), $R_{\mathrm{var}} = 70\%$, median $\cos\Delta = 0.89$ and median $e_{\mathrm{fit}} = 35.8\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\mathrm{fit}} \leq \tau_e$ only); robustness pattern B: admissible up to at least $\pm 20\%$ but not for larger changes.

External validation on later time points ($t \geq 410$; 20 test and 38 training measurements): test χ^2/n 1.50 for the reduced model, 3.89 for the full model refitted to the same data and 1.38 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Fiedler_BMCSystBiol2016

RAF/MEK/ERK cascade with ERK feedback, under different inhibitor treatments (6 species). Source: `doi:10.1186/s12918-016-0319-7`. The full model has 22 parameters, 8 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `tau2` and `k6` (stage 2, ranking by finite-perturbation effects); the reduced problem has 55% fewer free parameters than the full one. `tau2` is a time constant of the activation input of RAF; `k6` is the rate of dephosphorylation of pERK, and it also enters the steady state from which every simulation starts. For this subset $\kappa = 1.5$ and VIF 1.2 (no collinearity), $R_{\mathrm{var}} = 42\%$, median $\cos\Delta = 0.87$ and median $e_{\mathrm{fit}} = 14.7\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\mathrm{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (20 test and 52 training measurements): test χ^2/n 11.21 for the reduced model, 9.32 for the full model refitted to the same data and 1.22 for the

published θ_0. Here the refitted full model predicts the test data better than the reduced model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Fujita_SciSignal2010

EGF-dependent AKT signalling and S6 phosphorylation (9 species). Source: `doi:10.1126/scisignal.2000810`. The full model has 19 parameters, 3 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `init_AKT` and `reaction_6_k1` (stage 2, ranking by finite-perturbation effects); the reduced problem has 74% fewer free parameters than the full one. `init_AKT` is the initial amount of AKT; `reaction_6_k1` is the rate of release of pAKT and pS6 from the pAKT:S6 complex, i.e. of S6 phosphorylation. For this subset $\kappa = 2.3$ and VIF 1.9 (no collinearity), $R_{\text{var}} = 36\%$, median $\cos\Delta = 0.77$ and median $e_{\text{fit}} = 29.4\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (48 test and 96 training measurements): test χ^2/n 5.94 for the reduced model, 8.03 for the full model refitted to the same data and 5.63 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Giordano_Nature2020

SIDARTHE model of the COVID-19 epidemic in Italy, with parameters that change at the dates of the containment measures (8 species). Source: PubMed 32322102. The full model has 50 parameters. The method selected two biological parameters, `alpha_4` and `alpha_28` (stage 2, ranking by finite-perturbation effects); the reduced problem has 96% fewer free parameters than the full one. `alpha_4` and `alpha_28` are the transmission coefficient α (infection by undetected asymptomatic individuals) after the changes of day 4 and day 28. For this subset $\kappa = 3.7$ and VIF 3.9 (no collinearity), $R_{\text{var}} = 74\%$, median $\cos\Delta = 0.97$ and median $e_{\text{fit}} = 5.8\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on later time points ($t \geq 30$; 112 test and 201 training measurements): test χ^2/n <0.01 for the reduced model, <0.01 for the full model refitted to the same data and <0.01 for the published θ_0. All fits reproduce the data almost exactly, so the comparison is not informative. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Isensee_JCB2018

Activation of protein kinase A type II by cAMP in sensory neurons (25 species). Source: `doi:10.1083/jcb.201708053`. The full model has 46 parameters, 11 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `kf_RIIp_C_2__RII_C_2` and `xi_kf_RII_C_2__RII_2` (stage 1, ranking by sensitivity energy); backward elimination had removed `ks_AC_cAMP`, `KD_cAMP`, `xi_AC_cAMP_Fsk`, `kf_RIIp_cAMP_C_2__RIIp_2` and `kf_RII_2__RII_C_2` from the initial selection; the reduced problem has 72% fewer free parameters than the full one. `kf_RIIp_C_2__RII_C_2` is the rate of dephosphorylation of the regulatory subunit within the holoenzyme ($\text{pRII}_2{:}\text{C}_2 \rightarrow \text{RII}_2{:}\text{C}_2$); `xi_kf_RII_C_2__RII_2` is the factor that scales the dissociation of the catalytic subunits from the phosphorylated holoenzyme. For this subset $\kappa = 1.0$ and VIF 1.0 (no collinearity), $R_{\text{var}} = 68\%$, median $\cos\Delta = 0.87$ and median $e_{\text{fit}} = 11.3\%$ at $\pm 5\%$ after nonlinear re-estimation of the final subset (7.6% to first order during the selection): the reduced model reproduces the full one, with guarantee level 1

($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to ±50% (sizes above ±5% evaluated with the first-order e_{fit}).

External validation on held-out experimental conditions (219 test and 468 training measurements): test χ^2/n 1.08 for the reduced model, 1.08 for the full model refitted to the same data and 1.04 for the published θ_0. The two refitted models predict equally well (test χ^2/n within 5%). The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Lang_PLOSComputBiol2024

Mammalian cell cycle (E2F, RB, cyclins and CDK inhibitors) in the cytoplasm and nucleus (124 species). Source: `doi:10.1371/journal.pcbi.1011151`. The full model has 294 parameters, 16 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. **Not admissible.** Ranking by finite-perturbation effects (stage 2), the selection stopped at 13 parameters (`kDeE2f1`, `kDeCe`, `kDeCa2`, `kPhCAEB`, `kDpCdh`, `kImImSKP2_1`, `kImImCDKN1A_CCNE_A_B_Ser130p__1`, `ic_wt_Nuc__CDKN1B_CCNE_A_Thr187p_`, `ic_wt_Cyt__CDKN1B_CCNE_A_Thr187u_`, `ic_wt_Nuc__SKP2`, `ic_wt_Cyt__CCNE_CDKN1A_B_`, `ic_wt_Cyt__SKP2` and `ic_wt_Cyt__CCNE_CDKN1A_B_1__CDKN1A_CCNE_A_B_1_Ser130u_`), because any further parameter violated the conditioning test ($\kappa = 8.7$, VIF 9.99, just below the limit of 10). Re-estimating these 13 parameters gave a median first-order e_{fit} of 56%, above the threshold of 43.6%. Admissible pairs nevertheless exist: on the same scenarios, the search with conditioning thresholds of 5 or 30 instead of 10 (Supplementary Note *Random scenarios and large problems*, Supplementary Table S14) adds a different second parameter and finds {`kDeE2f1`, `kSySkp2`} ($e_{\text{fit}} = 13\%$) and {`kDeE2f1`, `kSyCb1`} ($e_{\text{fit}} = 16\%$). With the default threshold the second parameter is `kDeCe`, whose pair is not admissible, and the greedy search then grows until the conditioning limit without becoming admissible; the failure is therefore a limitation of the greedy search in this oscillating cell-cycle model (simulated to $t = 136{,}472$), not evidence that no small subset exists. The simulation of this model also differs from the reference simulation of the benchmark more than in any other problem (22% of the measurements by more than one noise unit; Supplementary Note *Verification of the simulator*), so this result should be read with that caveat.

Laske_PLOSComputBiol2019

Influenza A virus replication: nuclear import of viral ribonucleoproteins and viral RNA synthesis. Source: `doi:10.1371/journal.pcbi.1006944`. The full model has 13 parameters, 2 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `k_imp` and `k_syn_R_M` (stage 1, ranking by sensitivity energy); backward elimination had removed `k_bind_M1` from the initial selection; the reduced problem has 69% fewer free parameters than the full one. `k_imp` is the rate of nuclear import of viral ribonucleoproteins (cytoplasm → nucleus); `k_syn_R_M` is the synthesis rate of viral mRNA, from which the rates of the eight mRNA segments are derived. For this subset $\kappa = 1.4$ and VIF 1.1 (no collinearity), $R_{\text{var}} = 84\%$, median $\cos\Delta = 0.99$ and median $e_{\text{fit}} = 10.1\%$ at ±5%: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to ±50%.

External validation on held-out experimental conditions (9 test and 33 training measurements): test χ^2/n 0.76 for the reduced model, 1.87 for the full model refitted to the same data and 0.44 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Liu_IFACPapersOnLine2025

Virus–cell model: uninfected cells with logistic growth, five stages of infected cells and free virus (7 states, written as rate rules). Source: `doi:10.1016/j.ifacol.2025.07.160`. The full model has 9 parameters. The method selected two biological parameters, `kappa` and `phi` (stage 2, ranking by finite-perturbation effects); the reduced problem has 78% fewer free parameters than the full one. `kappa` is the carrying capacity of the logistic growth of the cells; `phi` is the rate of transition between successive stages of infection. For this subset $\kappa = 1.3$ and VIF 1.1 (no collinearity), $R_{\text{var}} = 81\%$, median $\cos\Delta = 0.92$ and median $e_{\text{fit}} = 26.1\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (50 test and 50 training measurements): test χ^2/n 0.95 for the reduced model, 1.02 for the full model refitted to the same data and 0.94 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The model was not verified in Lean because the model has events that depend on the state.

Lucarelli_CellSystems2018

TGF-β/SMAD signalling and the expression of its target genes (several dozen species). Source: `doi:10.1016/j.cels.2017.11.010`. The full model has 84 parameters. The method selected ten biological parameters, `init_Rec`, `S2tot`, `S3tot`, `S_dephosphos`, `S_phos`, `geneF_turn`, `k_234`, `geneK_turn`, `S4tot` and `geneH_turn` (stage 2, ranking by finite-perturbation effects); the reduced problem has 88% fewer free parameters than the full one. `init_Rec` is the amount of receptor; `S2tot`, `S3tot` and `S4tot` are the total amounts of SMAD2, SMAD3 and SMAD4; `S_phos` and `S_dephosphos` are the rates of phosphorylation and dephosphorylation of SMAD2/3; `k_234` is the rate of formation of the ppSMAD2:ppSMAD3:SMAD4 complex; `geneF_turn`, `geneH_turn` and `geneK_turn` are the turnover rates of three target genes. For this subset $\kappa = 6.5$ and VIF 9.8 (no collinearity), $R_{\text{var}} = 64\%$, median $\cos\Delta = 0.81$ and median $e_{\text{fit}} = 43.585\%$ at $\pm 5\%$ after nonlinear re-estimation of the final subset (43.5% to first order during the selection): the reduced model reproduces the full one, although only just within the threshold, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern C: irregular, with an error that stays close to the threshold and crosses it at some sizes (sizes above $\pm 5\%$ evaluated with the first-order e_{fit}).

External validation on held-out experimental conditions (598 test and 1157 training measurements): test χ^2/n 1.66 for the reduced model, 3.34 for the full model refitted to the same data and 1.02 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Okuonghae_ChaosSolitonsFractals2020

Model of the COVID-19 epidemic in Lagos, Nigeria, with detected and undetected infected classes (8 state variables). Source: `doi:10.1016/j.chaos.2020.110032`. The full model has 16 parameters. The method selected two biological parameters, `transmission_rate_effective` and `gamma_i` (stage 2, ranking by finite-perturbation effects); the reduced problem has 88% fewer free parameters than the full one. `transmission_rate_effective` is the effective transmission rate (force of infection, susceptible $\rightarrow$ exposed); `gamma_i` is the recovery rate of detected infected individuals. For this subset $\kappa = 1.3$ and VIF 1.1 (no collinearity), $R_{\text{var}} = 53\%$, median $\cos\Delta = 1.00$ and median $e_{\text{fit}} = 4.1\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on later time points ($t \geq 32$; 32 test and 60 training measurements): test χ^2/n 46,705 for the reduced model, 122,443 for the full model refitted to the same data and 28,748 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. Even θ_0 has $\chi^2/n \gg 1$, which indicates that the noise levels recorded in the benchmark do not

match the scatter of the data. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Oliveira_NatCommun2021

Model of the COVID-19 epidemic in Bahia, Brazil, with hospital and intensive-care compartments. Source: `doi:10.1038/s41467-020-19798-3`. The full model has 12 parameters. The method selected two biological parameters, `beta_0` and `h_hosp_rate` (stage 2, ranking by finite-perturbation effects); the reduced problem has 83% fewer free parameters than the full one. `beta_0` is the transmission rate (susceptible → exposed); `h_hosp_rate` is the fraction of symptomatic individuals who are hospitalised, which splits the outflow of symptomatic cases between recovery, hospital and ICU. For this subset $\kappa = 1.2$ and VIF 1.0 (no collinearity), $R_{\text{var}} = 47\%$, median $\cos\Delta = 0.95$ and median $e_{\text{fit}} = 12.9\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on later time points ($t \geq 40$; 40 test and 80 training measurements): test χ^2/n 912 for the reduced model, 81,349 for the full model refitted to the same data and 460,396 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model, whose training error (χ^2/n 8.75) is much smaller than its test error. Even θ_0 has $\chi^2/n \gg 1$, which indicates that the noise levels recorded in the benchmark do not match the scatter of the data. The model was not verified in Lean because the model has events that depend on the parameters.

Perelson_Science1996

HIV-1 dynamics in a patient treated with a protease inhibitor (infected cells, infectious and non-infectious virus). Source: `doi:10.1126/science.271.5255.1582`. The full model has 3 parameters. The method selected two biological parameters, `c` and `delta` (stage 1, ranking by sensitivity energy); the reduced problem has 33% fewer free parameters than the full one. `c` is the clearance rate of free virions; `delta` is the death rate of productively infected cells. For this subset $\kappa = 1.9$ and VIF 1.5 (no collinearity), $R_{\text{var}} = 100\%$, median $\cos\Delta = 1.00$ and median $e_{\text{fit}} = 0.0\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 3 ($\cos\Delta \geq 0.9$ and $R_{\text{var}} \geq 0.89$); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on later time points ($t \geq 2.038$; 6 test and 10 training measurements): test χ^2/n <0.01 for the reduced model, <0.01 for the full model refitted to the same data and <0.01 for the published θ_0. All fits reproduce the data almost exactly, so the comparison is not informative. The model was not verified in Lean because the quasi-positivity check fails for two terms.

Rahman_MBS2016

Transmission model with infected classes of different severity and a treated class. Source: `doi:10.1016/j.mbs.2016.07.009`. The full model has 9 parameters. The method selected two biological parameters, `infected_moderate_transmission_rate` and `infected_normal_worsen_rate` (stage 2, ranking by finite-perturbation effects); the reduced problem has 78% fewer free parameters than the full one. `infected_moderate_transmission_rate` is the transmission rate of moderately infected individuals, to which the rates of the other classes are tied by assignment rules (it sets the force of infection); `infected_normal_worsen_rate` is the rate of progression from normal to moderate infection. For this subset $\kappa = 5.9$ and VIF 9.3 (no collinearity), $R_{\text{var}} = 60\%$, median $\cos\Delta = 0.97$ and median $e_{\text{fit}} = 3.7\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on later time points ($t \geq 15$; 8 test and 15 training measurements): test χ^2/n <0.01 for the reduced model, 0.03 for the full model refitted to the same data and <0.01 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data much better than the refitted full model. The

hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Raia_CancerResearch2011

IL-13 signalling through JAK/STAT6 with SOCS3 feedback in Hodgkin lymphoma cells (several dozen parameters, 4 conditions). Source: `doi:10.1158/0008-5472.CAN-10-2987`. The full model has 39 parameters, 5 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `SOCS3mRNA_production` and `init_Rec_i` (stage 2, ranking by finite-perturbation effects); the reduced problem has 82% fewer free parameters than the full one. `SOCS3mRNA_production` is the production rate of SOCS3 mRNA, which drives the negative feedback; `init_Rec_i` is the initial amount of internalised receptor, which is recycled to the membrane. For this subset $\kappa = 5.5$ and VIF 8.1 (no collinearity), $R_{\text{var}} = 69\%$, median $\cos\Delta = 1.00$ and median $e_{\text{fit}} = 3.3\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (72 test and 133 training measurements): test χ^2/n 8.22 for the reduced model, 5.10 for the full model refitted to the same data and 0.42 for the published θ_0. Here the refitted full model predicts the test data better than the reduced model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Raimundez_PCB2020

EGFR and MET signalling in two gastric cancer cell lines (MKN1 and HS746T) treated with EGF, HGF and cetuximab. Source: `doi:10.1371/journal.pcbi.1007147`. The full model has 136 parameters, 79 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `kdeg_membran__MKN1` and `ka_MAPK__MKN1` (stage 2, ranking by finite-perturbation effects); the reduced problem has 40% fewer free parameters than the full one. `kdeg_membran__MKN1` is the degradation rate of membrane receptors (EGFR, EGFR:EGF, EGFR:cetuximab) in MKN1 cells; `ka_MAPK__MKN1` is the rate constant of MAPK (ERK) phosphorylation in MKN1 cells; its published value lies at the upper bound of the benchmark range (10^3). For this subset $\kappa = 1.3$ and VIF 1.1 (no collinearity), $R_{\text{var}} = 42\%$, median $\cos\Delta = 0.79$ and median $e_{\text{fit}} = 18.6\%$ at $\pm 5\%$ after nonlinear re-estimation of the final subset (18.7% to first order during the selection): the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$ (sizes above $\pm 5\%$ evaluated with the first-order e_{fit}).

External validation on held-out experimental conditions (208 test and 419 training measurements): test χ^2/n 1.17 for the reduced model, 1.43 for the full model refitted to the same data and 0.98 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

SalazarCavazos_MBoC2020

Multisite phosphorylation of the epidermal growth factor receptor (EGFR). Source: `doi:10.1091/mbc.E19-09-0548`. The full model has 6 parameters. The method selected two biological parameters, `ratio_kpkd_Y1068__FREE` and `ratio_kpkd_YN__FREE` (stage 2, ranking by finite-perturbation effects); the reduced problem has 67% fewer free parameters than the full one. `ratio_kpkd_Y1068__FREE` and `ratio_kpkd_YN__FREE` are the ratios of phosphorylation to dephosphorylation rate of tyrosine Y1068 and of the other tyrosines in receptors that are not bound to adaptor proteins. For this subset $\kappa = 5.4$ and VIF 7.9 (no collinearity), $R_{\text{var}} = 84\%$, median $\cos\Delta = 0.97$ and median $e_{\text{fit}} = 18.7\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness

pattern A: admissible at all perturbation sizes up to ±50%.

External validation on held-out experimental conditions (3 test and 15 training measurements): test χ^2/n 102 for the reduced model, 124 for the full model refitted to the same data and 96.93 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. Even θ_0 has $\chi^2/n \gg 1$, which indicates that the noise levels recorded in the benchmark do not match the scatter of the data. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Schwen_PONE2014

Binding and uptake of insulin by hepatocytes through receptors of low and high affinity. Source: `doi:10.1371/journal.pone.0133653`. The full model has 30 parameters, 15 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `ini_R1` and `ini_R2fold` (stage 2, ranking by finite-perturbation effects); the reduced problem has 43% fewer free parameters than the full one. `ini_R1` is the initial amount of low-affinity receptors, and `ini_R2fold` the factor that gives the amount of high-affinity receptors from it; both set the binding and uptake of insulin. For this subset $\kappa = 3.1$ and VIF 3.0 (no collinearity), $R_{\text{var}} = 46\%$, median $\cos\Delta = 0.93$ and median $e_{\text{fit}} = 17.1\%$ at ±5%: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to ±50%.

External validation on held-out experimental conditions (19 test and 267 training measurements): test χ^2/n 8.20 for the reduced model, 8.40 for the full model refitted to the same data and 0.77 for the published θ_0. The two refitted models predict equally well (test χ^2/n within 5%). The model was not verified in Lean because the quasi-positivity check fails for two terms.

Smith_BMCSystBiol2013

Insulin signalling through IRS1, PI3K and serine phosphorylation of IRS1. Source: `doi:10.1186/1752-0509-7-41`. The full model has 25 parameters, 9 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `k8` and `kminus7b` (stage 2, ranking by finite-perturbation effects); the reduced problem has 56% fewer free parameters than the full one. `k8` is the binding rate of PI3K to tyrosine-phosphorylated IRS1; `kminus7b` is the rate of removal of the serine phosphorylation of IRS1 (reverse of the inhibitory serine phosphorylation). For this subset $\kappa = 3.2$ and VIF 3.1 (no collinearity), $R_{\text{var}} = 41\%$, median $\cos\Delta = 0.96$ and median $e_{\text{fit}} = 7.8\%$ at ±5%: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to ±50%.

External validation on held-out experimental conditions (15 test and 47 training measurements): test χ^2/n 14,977 for the reduced model, 14,976 for the full model refitted to the same data and 82,921 for the published θ_0. The two refitted models predict equally well (test χ^2/n within 5%). Even θ_0 has $\chi^2/n \gg 1$, which indicates that the noise levels recorded in the benchmark do not match the scatter of the data. The model was not verified in Lean because the rate laws contain a non-differentiable maximum.

Sneyd_PNAS2002

Gating of the IP_3 receptor Ca^{2+} channel (open, shut and inactivated states). Source: `doi:10.1073/pnas.032281999`. The full model has 15 parameters. The method selected two biological parameters, `l4` and `l2` (stage 2, ranking by finite-perturbation effects); the reduced problem has 87% fewer free parameters than the full one. `l2` and `l4` are rate constants of the Ca^{2+}-dependent transitions of the receptor from the rest state to the open and inactivated states. For this subset $\kappa = 1.1$ and VIF 1.0 (no

collinearity), $R_{\text{var}} = 44\%$, median $\cos\Delta = 0.88$ and median $e_{\text{fit}} = 38.1\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (45 test and 90 training measurements): test χ^2/n 2.22 for the reduced model, 2.56 for the full model refitted to the same data and 1.37 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Weber_BMC2015

Feedback between protein kinase D (PKD), the ceramide transfer protein CERT and PI4KIIIβ at the Golgi. Source: `doi:10.1186/s12918-015-0147-1`. The full model has 36 parameters, 5 of which are measurement calibration parameters (scales and offsets) that are always re-estimated; noise parameters stay at their published values. The method selected two biological parameters, `s12` and `a21` (stage 2, ranking by finite-perturbation effects); the reduced problem has 81% fewer free parameters than the full one. `s12` is the constant production rate of PKD; `a21` is the first-order removal rate of PI4KIIIβ. For this subset $\kappa = 3.7$ and VIF 4.0 (no collinearity), $R_{\text{var}} = 25\%$, median $\cos\Delta = 0.41$ and median $e_{\text{fit}} = 4.1\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (59 test and 76 training measurements): test χ^2/n 1.00 for the reduced model, 15.95 for the full model refitted to the same data and 0.99 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data much better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Zhao_QuantBiol2020

Model of the COVID-19 epidemic in China with unquarantined and quarantined infected individuals, in two stages and three regions. Source: PubMed 32219006. The full model has 28 parameters. The method selected three biological parameters, `R_Stage_I_China`, `R_Stage_II_Wuhan` and `gamma_1_Stage_I_Hubei` (stage 2, ranking by finite-perturbation effects); the reduced problem has 89% fewer free parameters than the full one. `R_Stage_I_China` and `R_Stage_II_Wuhan` are the reproduction numbers of stage I in China and of stage II in Wuhan, which set the infection rate; `gamma_1_Stage_I_Hubei` is the quarantine rate of infected individuals in Hubei in stage I. For this subset $\kappa = 1.2$ and VIF 1.0 (no collinearity), $R_{\text{var}} = 69\%$, median $\cos\Delta = 0.90$ and median $e_{\text{fit}} = 29.6\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 2 ($\cos\Delta \geq 0.9$ as well); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on held-out experimental conditions (25 test and 57 training measurements): test χ^2/n 0.35 for the reduced model, 0.35 for the full model refitted to the same data and 0.35 for the published θ_0. The two refitted models predict equally well (test χ^2/n within 5%). The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Zheng_PNAS2012

Methylation of histone H3 at lysines 27 and 36 (combinatorial methylation states). Source: `doi:10.1073/pnas.1205707109`. The full model has 46 parameters. The method selected two biological parameters, `k01_02` and `k13_12` (stage 2, ranking by finite-perturbation effects); the reduced problem has 96% fewer free parameters than the full one. `k01_02` is the rate of methylation of K36 from mono- to dimethyl in histones without K27 methylation; `k13_12` is the rate of demethylation of K36 from tri- to dimethyl in histones with K27me1. For this subset $\kappa = 1.4$ and VIF 1.1 (no collinearity), $R_{\text{var}} = 35\%$,

median $\cos\Delta = 0.57$ and median $e_{\text{fit}} = 30.0\%$ at $\pm 5\%$: the reduced model reproduces the full one, with guarantee level 1 ($e_{\text{fit}} \leq \tau_e$ only); robustness pattern A: admissible at all perturbation sizes up to $\pm 50\%$.

External validation on later time points ($t \geq 10$; 30 test and 30 training measurements): test χ^2/n 1.58 for the reduced model, 20.82 for the full model refitted to the same data and 1.51 for the published θ_0. The reduced model, anchored at θ_0, predicts the test data much better than the refitted full model. The hypotheses on the dynamics were verified in Lean (Supplementary Note *Verified benchmark models*).

Supplementary Note: Correspondence between claims and Lean theorems

Supplementary Table S43 lists every claim of the main text that refers to the formal development, the Lean theorem that supports it (file and name in `lean_proofs/`), the hypotheses of the theorem as stated in Lean, and the numerical quantities of this study that instantiate them. A claim holds for a result of the paper only when the hypotheses in the third column are satisfied by the values in the fourth; where they are checked numerically rather than proved, the table says so.

Supplementary Table S43. Claims of the main text and the Lean theorems that support them.

Claim (main text)	Lean theorem	Hypotheses (literal)	Instance in this study
Theorem 1: rescaling leaves a relative error $\sqrt{1-\cos^2}$	`ScaledError.min_rel_error_eq`	$a \neq 0$, $b \neq 0$ in a real inner-product space	$a = \Delta y_{\text{full}}$, $b = \Delta y_S$ of each scenario; $\tau_e = \sqrt{1-0.9^2} = 0.436$
Level 2: median $\cos\Delta \geq c$ implies the threshold for a rescaling	`ScaledError.scaled_error_cos`	as above	median $\cos\Delta \geq 0.9$ in 17 subsets (Table 1, Level)
Theorem 2, first part: the result of backward elimination is admissible and no element can be removed	`Poda.prune_spec`	any predicate Q on finite sets; $Q(S)$ for the starting set	Q = admissibility as implemented, including the rule $\lvert S\rvert \geq 2$; holds for all 31 subsets without further conditions (non-trivial only for Bruno, Lucarelli and Zhao, the subsets with more than two parameters)
Theorem 2, second part: no proper subset is admissible	`Poda.prune_inclusion_minimal`	Q monotone (every superset of an admissible set is admissible)	checked empirically: 0 violations among the 660 nested pairs evaluated during the selections (Methods)
More re-estimated parameters never increase the fitting error; $e_{\text{fit}} \leq e_{\text{rel}}$	`RefitMonotone.fitErr_anti`, `fitErr_le_erel`, `admissible_mono`	exact minimisation over the feasible set	holds for the exact optimum; the implementation uses a local optimiser from two starting points and includes the value without re-estimation, so $e_{\text{fit}} \leq e_{\text{rel}}$ holds by construction in every scenario; the computed e_{fit} bounds the exact one from above, so a subset admissible as computed is admissible exactly, but not conversely
Greedy selection keeps κ and VIF below the thresholds; a rejected candidate can be discarded permanently	`GreedySelection.greedy_ok`, `greedy_discard_permanent`, `greedy_kappa_vif_guarantee`	the filter is antitone (true for κ and VIF of any fixed column normalisation)	the implementation uses centred and scaled columns of the biological parameters; all 31 subsets $\kappa \leq 6.5$, VIF ≤ 9.8 (with the calibration parameters included the VIF exceeds 10 in ten problems)

(continued)

Claim (main text)	Lean theorem	Hypotheses (literal)	Instance in this study
κ and VIF bounds give a unique and stable linearised least-squares solution	`Identifiability Conditioning.ls_kappa_stability`, `ls_vif_coordinate_stability`	bounds on κ and VIF of the unit-length columns Z	$\kappa(Z) \leq 6.6$, $\mathrm{VIF}(Z) \leq 10$ in 30 subsets; not satisfied for Rahman ($\kappa = 12.4$, VIF= 39; Supplementary Table S44); for the problem actually solved, with Θ_C re-estimated together with θ_S, $\kappa > 10$ in ten problems (last column of Supplementary Table S44)
Floating-point κ and VIF certify the exact values	`NumericalRobustness. kappa_certified_from_approx`, `vif_certified_from_approx`	a bound on $\|Z - \tilde{Z}\|$	not instantiated: no bound on the rounding error of J was computed
Level 3: the first-order error is bounded by the discarded energy	`FirstOrderErrorBound. first_order_error_from_Rvar`	$R_{\mathrm{var}}(S) \geq \tau_R$; first-order model of a verified model	none: Alkan and Perelson satisfy $R_{\mathrm{var}} \geq \tau_R$, but their models are not verified in Lean
Bounds transfer to transformed outputs	`OutputComposition. output_sq_bound`, `log_lipschitz`, `log10_lipschitz`	outputs bounded below by $c > 0$ for log transforms	all problems with log-transformed observables
First-order e_{fit} differs from the exact one by a second-order term	`LinearizedFit.fit_upper_of_lin`, `fit_lower_of_lin`	a bound M on the second-order remainder	not instantiated (M not computed); the six first-order problems were re-estimated nonlinearly instead (Supplementary Note *Nonlinear check of the first-order problems*)
Model hypotheses: solution exists, stays non-negative and is differentiable on the horizon	final theorem of each file `Models/<problem>.lean`	checked by the Lean kernel (`decide +kernel`) at the nominal parameters and in the base condition (first segment of piecewise models)	28 models (27 of the 34 problems and Froehlich); Crauste only for $T \leq 1$; the selected initial conditions of Bertozzi and Bruno are not parameters of the Lean files
Practical identifiability of a selected parameter	`ProfileIdentifiability. practically_identifiable_grad`	$\|R\delta - R0 - A\delta\| \leq \varepsilon$ for all δ in the admissible set F; $\|\langle r_0, A\delta\rangle\| \leq \gamma\|A\delta\|$; $t^* <$ distance to both bounds	ε estimated at the computed profile points (not bounded over F): hypotheses satisfied at all evaluated points for 24 parameters, all of S in 7 problems; a numerical check, not a certificate
Finite-perturbation certificates	`CertifiedFiniteODE. certified_cos_ode`, `ExactLinearCertificate. certified_cos_ode_exact`	explicit Lipschitz and remainder constants	constants 10^2–10^{24}: no scenario at $\pm 5\%$ is certified in any problem (Supplementary Table S5)

Supplementary Note: Conditioning with unit-length columns

The selection tests κ and VIF on columns of J_S that are centred and scaled to unit variance (the correlation matrix of the sensitivity columns), whereas the Lean statements on the linearised least-squares problem use columns scaled to unit length without centring. Both are antitone in the sense required by the greedy selection, so the selection itself is not affected; the statements on the linearised problem, however, apply only when the unit-length values satisfy the bounds. Supplementary Table S44 gives both for every final subset (`analysis/conditioning_check.py`; central differences with step 1%, $2|S| + 2|\Theta_C|$ simulations per problem). They agree closely except in Rahman, where the unit-length VIF is 39. With the calibration parameters included, the unit-length matrix is far worse conditioned in several problems (for example Fujita and Isensee), because scaling factors of different observables can have nearly proportional sensitivities; this does not affect the selection, which conditions only the biological parameters.

Supplementary Table S44. Conditioning of the final subsets. κ and VIF of the selected columns, centred and scaled (as in the selection) and scaled to unit length without centring (as in the Lean statements); last column, κ of the unit-length columns of the selected and calibration parameters together.

		Centred, scaled		Unit length		With Θ_C
Problem	$\lvert S\rvert$	κ	VIF	κ	VIF	κ
Alkan	2	1.33	1.08	1.31	1.08	1.44
Armistead	2	1.00	1.00	1.29	1.07	1.29
Bachmann	2	2.83	2.53	2.43	2.02	5.29
Bertozzi	2	3.91	4.33	4.65	5.92	4.65
Blasi	2	1.76	1.36	2.64	2.27	2.64
Boehm	2	1.02	1.00	1.87	1.45	1.87
Borghans	2	2.74	2.41	2.79	2.48	4.15
Brannmark	2	1.86	1.44	1.27	1.06	25.43
Bruno	5	2.26	1.57	1.64	1.27	1.64
Chen	2	1.01	1.00	1.03	1.00	1.45
Elowitz	2	2.86	2.58	2.52	2.13	5.55
Fiedler	2	1.52	1.18	1.61	1.25	14.47
Fujita	2	2.34	1.91	2.62	2.25	3.7e+04
Giordano	2	3.67	3.88	3.81	4.15	3.81
Isensee	2	1.02	1.00	2.11	1.67	3.7e+14
Laske	2	1.45	1.14	1.21	1.04	2.94
Liu	2	1.29	1.07	1.73	1.33	1.73
Lucarelli	10	6.54	9.85	6.58	9.95	6.58
Okuonghae	2	1.34	1.09	1.72	1.33	1.72
Oliveira	2	1.23	1.04	1.02	1.00	1.02
Perelson	2	1.90	1.48	1.50	1.17	1.50
Rahman	2	5.93	9.28	12.41	39.02	12.41
Raia	2	5.50	8.07	5.02	6.81	364.64
Raimundez	2	1.34	1.09	1.41	1.12	21.89
SalazarCavazos	2	5.43	7.87	4.54	5.67	4.54
Schwen	2	3.15	3.00	3.14	3.00	21.01
Smith	2	3.20	3.08	3.46	3.52	14.02
Sneyd	2	1.07	1.00	1.22	1.04	1.22
Weber	2	3.74	4.02	3.79	4.11	51.62
Zhao	3	1.17	1.02	1.00	1.00	1.00
Zheng	2	1.40	1.12	1.40	1.12	1.40

Supplementary Note: Nonlinear check of the first-order problems

In Alkan, Bachmann, Chen, Isensee, Lucarelli and Raimundez, admissibility was decided during the selection with the first-order $e_{\mathrm{fit,lin}}$. To check this decision, the final subset of each problem was re-estimated nonlinearly in the same 15 scenarios at $\pm 5\%$, with exactly the settings used for the other problems: trust-region reflective least squares in $\log|\theta|$ within $[|\theta_0|/10,\ 10|\theta_0|]$, two starting points ($\theta_0$ and the scenario value), at most $100(|S| + |\Theta_C| + 1)$ function evaluations, and the value without re-estimation included (`analysis/efit_nonlinear.py`; one process per scenario).

Results. The nonlinear re-estimation confirmed admissibility in all six problems (Supplementary Table S45). In Alkan, Bachmann, Chen and Raimundez the nonlinear and first-order medians agree to within 0.002; in Isensee the nonlinear error is larger (0.113 against 0.076) but well below the threshold, with six valid scenarios. In Lucarelli the nonlinear median, 0.43585, lies 4×10^{-5} below $\tau_e = 0.43589$, with 8 of the 15 scenarios passing: its admissibility is confirmed but borderline, as stated in the main text. The computation took 21 h of computing time over the six problems, 13 h of them for Chen (35 s per simulation).

Supplementary Note: Selection on the training outputs, null controls and standard pivoted QR

This note reports three further analyses of the comparison of selection criteria. (i) The main proposed subsets were selected on all measured outputs, whereas the other criteria selected on the training

Supplementary Table S45. Nonlinear check of the problems decided to first order. Median e_{fit} over the 15 scenarios at ±5% to first order (as in the selection) and after nonlinear re-estimation of the final subset; valid scenarios, fraction of them with $e_{\text{fit}} \leq \tau_e$, and computing time.

Problem	e_{fit} first order	e_{fit} nonlinear	Valid scenarios	Pass rate	CPU time (h)
Alkan	0.12126	0.12011	15	100%	0.1
Bachmann	0.33710	0.33802	15	100%	0.4
Chen	0.08201	0.08080	10	100%	12.7
Isensee	0.07589	0.11320	6	100%	3.9
Lucarelli	0.43545	0.43585	15	53%	0.1
Raimundez	0.18709	0.18570	15	100%	4.1

outputs, which may favour the proposed method in the tests that use held-out conditions; the selection was therefore repeated on the training outputs. (ii) Null baselines show how selective the admissibility test is in each problem. (iii) Pivoted QR is reported in the standard Golub–Klema–Stewart form; an earlier, direct pivoting variant (column-pivoted QR of the normalised sensitivity matrix itself) is given for completeness.

Design. (i) The complete selection was repeated with the rows of the training set only (`selection.py` with `REPRO_OUTPUTS=train`), with the fixed scenario set of the Methods; the resulting subsets were evaluated with the same admissibility test (on all outputs, ±5% and ±20%), the same real-data test and the same synthetic experiment (20 seeds) as the other criteria. (ii) The admissibility of re-estimating only the calibration parameters ($k = 0$, without the rule $|S| \geq 2$) and of 20 random k-subsets per problem (the subsets of the real-data random baseline, seeds 0–19) was evaluated at ±5% and ±20%. To keep the comparison consistent, the five criteria were re-evaluated with the same rule: the fixed scenario set, nonlinear re-estimation, and, only in the eight problems that the method treats to first order (Alkan, Bachmann, Beer, Chen, Isensee, Lang, Lucarelli, Raimundez), the first-order $e_{\text{fit,lin}} = \min\{e_{\text{rel}}, \text{projection}\}$ evaluated scenario by scenario. These are the values of Table 2 of the main text and of Supplementary Table S46. (iii) Standard pivoted QR (Golub, Klema and Stewart): SVD of the normalised training sensitivity matrix and column-pivoted QR of $V_k^\top$ (`qr_selection.py` with `QR_VARIANT=gks`). (iv) The synthetic comparison was repeated with log-uniform and sparse truths (Methods), ten seeds.

Results. (i) *Selection on the training outputs.* With the training outputs alone, 29 of the 34 problems have an admissible subset and 20 select a different subset from the main one, mostly another pair of parameters: in sloppy models many pairs suffice. Bertozzi has no admissible pair on the training outputs, although a single parameter (`RO_CA`) reaches $e_{\text{fit}} = 0.08$, and Perelson cannot be evaluated because its training outputs change by less than the numerical noise floor. Evaluated on all outputs, these subsets were admissible in 24 of 27 comparable problems at ±5% and 23 of 27 at ±20%, against 28 and 27 of 28 for the main subsets: the main subsets owe part of their perfect score to being selected and tested on the same outputs. On real data their test χ^2/n did not differ from that of the main subsets (median ratio 1.00, Wilcoxon $P = 0.17$). On synthetic data they reached $e_{\text{pred}} \leq \tau_e$ in 51%, 48% and 55% of the cases for the three families of truths, not significantly different from Brun's (55%, 50% and 52%) and significantly more often than profile likelihood and standard pivoted QR in every family (Supplementary Table S47).

(ii) *Null controls.* Re-estimating the calibration parameters alone was admissible in 9 of 28 problems at both perturbation sizes, and random k-subsets in a median of 16 of 28 problems over the 20 draws (14–19 at ±5%, 14–18 at ±20%). In 16 problems at least half of the random draws are admissible, so in these models many subsets suffice, whereas in others, such as Lucarelli (0 of 20 draws admissible), the choice of parameters is decisive (Supplementary Table S46). The main proposed subsets (28 of 28), Brun's (26) and the training-output subsets (24 of 27) are well above both controls; Morris (23), profile likelihood (18 of 27) and standard pivoted QR (17) are close to the random baseline.

(iii) *Standard pivoted QR.* The Golub–Klema–Stewart selection chose a different subset from the earlier variant in 23 of 29 problems. It was admissible in 17 of 28 problems at ±5% (earlier variant: 14) and

16 at ±20% (15), reached the synthetic threshold in 30% of the uniform cases (34%), and its real-data test error did not differ significantly from the proposed one (median ratio 1.00, $P = 0.16$). The earlier variant differs in one respect: on real data its test error was slightly lower than the proposed one (median ratio 0.99, $p = 0.026$, unadjusted; Supplementary Table S26). The main text reports the standard variant.

(iv) *Other families of truths.* Supplementary Table S47 gives, for the three families, the success rate of every criterion and the difference from the proposed method with 95% confidence intervals from resampling problems. The ordering is the same in all three: the proposed method and Brun's are not significantly different, and both are above profile likelihood and standard pivoted QR; the difference from Morris is significant for uniform and sparse truths but not for log-uniform ones.

Supplementary Table S46. Admissibility at ±5% per problem with the fixed scenario set: median e_{fit} of the subset of each criterion (bold, admissible; for $k = 0$ without the rule $|S| \geq 2$) and, for random subsets, the number of the 20 draws that are admissible. –, not evaluated.

Problem	Proposed	Proposed (train)	Brun	Morris	Profile	QR (GKS)	QR (earlier)	$k = 0$	Random
Alkan	**0.12**	**0.13**	**0.13**	**0.12**	**0.13**	0.89	0.95	0.95	1/20
Armistead	**0.33**	**0.33**	**0.33**	**0.33**	0.57	0.60	0.74	1.00	0/20
Bachmann	**0.34**	**0.31**	**0.31**	**0.34**	**0.29**	**0.35**	0.50	0.51	8/20
Bertozzi	**0.33**	–	0.94	0.94	0.94	0.94	0.94	1.00	12/20
Boehm	**0.09**	**0.09**	**0.09**	**0.09**	0.57	0.46	0.67	1.00	4/20
Borghans	**0.04**	**0.17**	**0.17**	**0.36**	**0.10**	**0.04**	0.53	0.55	4/20
Brannmark	**0.08**	**0.08**	**0.10**	**0.11**	**0.09**	**0.10**	**0.12**	**0.16**	20/20
Bruno	**0.38**	0.71	0.71	0.86	0.71	0.89	0.89	1.00	1/20
Elowitz	**0.36**	**0.36**	**0.36**	0.65	0.65	0.69	0.72	0.77	8/20
Fiedler	**0.15**	**0.13**	**0.13**	**0.13**	**0.35**	**0.34**	**0.13**	**0.42**	20/20
Fujita	**0.29**	**0.29**	**0.29**	**0.27**	**0.22**	**0.18**	**0.18**	**0.30**	20/20
Giordano	**0.06**	**0.05**	**0.05**	**0.06**	0.82	0.78	0.91	1.00	10/20
Isensee	**0.08**	**0.08**	**0.10**	**0.10**	**0.10**	**0.08**	**0.10**	**0.10**	20/20
Laske	**0.10**	**0.15**	**0.15**	**0.15**	**0.15**	**0.15**	**0.12**	**0.22**	20/20
Liu	**0.26**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**	1.00	2/20
Lucarelli	**0.44**	0.49	**0.41**	0.55	0.74	0.77	0.91	1.00	0/20
Okuonghae	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.16**	**0.11**	1.00	20/20
Oliveira	**0.13**	**0.02**	**0.08**	**0.08**	**0.05**	**0.04**	**0.04**	1.00	16/20
Perelson	**0.00**	–	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	1.00	20/20
Rahman	**0.04**	**0.04**	**0.04**	**0.08**	**0.04**	**0.13**	**0.11**	1.00	11/20
Raia	**0.03**	**0.06**	**0.03**	**0.03**	**0.05**	**0.06**	**0.04**	**0.08**	20/20
Raimundez	**0.19**	**0.19**	**0.19**	**0.28**	–	**0.25**	**0.28**	**0.28**	20/20
Salazar-Cavazos	**0.19**	**0.19**	**0.19**	**0.19**	**0.40**	0.52	0.52	1.00	20/20
Schwen	**0.17**	**0.17**	**0.19**	**0.17**	**0.17**	**0.15**	**0.19**	**0.27**	20/20
Smith	**0.08**	**0.08**	**0.08**	**0.08**	**0.06**	**0.11**	**0.09**	**0.14**	20/20
Sneyd	**0.38**	**0.38**	**0.38**	0.50	0.68	0.90	0.74	1.00	1/20
Weber	**0.04**	**0.04**	**0.04**	**0.04**	**0.07**	**0.41**	0.48	0.48	19/20
Zhao	**0.30**	0.61	0.44	0.61	0.97	0.44	0.97	1.00	1/20
Zheng	**0.30**	**0.29**	**0.29**	**0.29**	0.67	0.93	0.70	1.00	0/20

Supplementary Note: Identifiability over the admissible set

The hypothesis of `practically_identifiable_grad` on the nonlinear remainder must hold for every parameter vector of the admissible set (the box of the free parameters within the bounds of the study), not only at the computed profile points. This note checks it over that set, tests whether the classification of the profiles depends on their resolution, and profiles the subsets selected on the training outputs.

Design. (a) The profiles of the subsets selected on the training outputs were computed with the same procedure, for the problems in which they differ from the main ones. (b) The profiles of the main subsets were recomputed with 20 displacements per side (from 0.02 to 4.6 in $\log|\theta_j|$, geometric spacing) instead of 10, from the same optimum. (c) Two variants of the theorem were proved in Lean (`practically_identifiable_grad_tail` and `practically_identifiable_rel_tail`): the remainder bound is needed only on the tail of the profile beyond the guaranteed half-width, either in absolute form ($\leq \varepsilon$) or relative to the linear change ($\leq \eta\|A\delta\|$, $\eta < 1$). For every selected parameter and

side, the tail bounds were estimated over the region of the free parameters within a factor r of the optimum (and within the bounds of the study), for $r = 10$, 5, 3, 2 and 1.5, from 512 Sobol points, the profile points of the region and an adversarial search (L-BFGS-B maximisation of the remainder or of the ratio on the tail, from the worst sampled points, repeated until the half-width found was stable; `analysis/identifiability_box.py`). A parameter satisfies the hypotheses in the region when one of the two conditions holds on both sides.

Results. (a) For the subsets selected on the training outputs, 35 of the 43 selected parameters were identifiable. (b) With twice as many profile points, the same 65 of 74 parameters were identifiable; the only change of category was `reaction_6_k1` of Fujita, from flat to non-identifiable on one side. (c) Over the region, the hypotheses held for 9 of the 74 parameters with $r = 1.5$, 7 with $r = 2$, 3 with $r = 3$ and none with $r = 5$; with $r = 10$ (the whole range of the study) the sampling found one parameter (`a_basal` of Blasi), but the region of $r = 5$ is contained in that of $r = 10$, so the counterexample found with $r = 5$ also lies in the larger region and the defensible count over the whole range is none (Supplementary Table S48). The Lean condition therefore supports practical identifiability only in a neighbourhood of the optimum, because the models are far from linear over a factor of ten in every free parameter; identifiability over the range of the study rests on the computed profiles. The sampled bounds are estimates of suprema, not proofs.

Supplementary Table S48. Identifiability of the selected parameters: profile classification with 10 and 20 points per side, and whether the hypotheses of the Lean theorems with the remainder bound on the tail (`practically_identifiable_grad_tail` or `_rel_tail`) hold over the whole region within a factor r of the optimum, by sampling (512 points) and adversarial search; a numerical check of the hypotheses, not a certificate.

Problem	Parameter	Profile (10)	Profile (20)	$r = 10$	5	3	2	1.5
Alkan	k_rep_nhej	identifiable	identifiable	–	–	–	–	–
Alkan	kt	identifiable	identifiable	–	–	–	–	–
Armistead	k00	identifiable	identifiable	–	–	–	–	yes
Armistead	k_d	identifiable	identifiable	–	–	–	yes	yes
Bachmann	CISRNADelay	identifiable	identifiable	–	–	–	–	–
Bachmann	JAK2ActEpo	identifiable	identifiable	–	–	–	–	–
Bertozzi	I0_NY	identifiable	identifiable	–	–	–	–	–
Bertozzi	R0_NY	identifiable	identifiable	–	–	–	–	–
Blasi	a_basal	identifiable	identifiable	yes	–	yes	yes	yes
Blasi	a_k16_k12k16	identifiable	identifiable	–	–	–	yes	yes
Boehm	k_imp_hetero	identifiable	identifiable	–	–	–	yes	yes
Boehm	k_phos	identifiable	identifiable	–	–	–	–	–
Borghans	K_par	identifiable	identifiable	–	–	–	–	–
Borghans	Kz	identifiable	identifiable	–	–	–	–	–
Brannmark	k1f	identifiable	identifiable	–	–	–	–	–
Brannmark	km2	identifiable	identifiable	–	–	–	–	–
Bruno	init_bcar1	identifiable	identifiable	–	–	–	–	–
Bruno	init_bcar2	identifiable	identifiable	–	–	–	–	–
Bruno	init_bcry_1	identifiable	identifiable	–	–	–	–	–
Bruno	init_zea_1	identifiable	identifiable	–	–	–	–	–
Bruno	kb1	identifiable	identifiable	–	–	–	–	–
Chen	k106	identifiable	identifiable	–	–	–	–	–
Chen	k44	identifiable	identifiable	–	–	–	–	–
Elowitz	n_Hill	identifiable	identifiable	–	–	–	–	–
Elowitz	tau_mRNA	identifiable	identifiable	–	–	–	–	–
Fiedler	k6	identifiable	identifiable	–	–	–	–	–
Fiedler	tau2	identifiable	identifiable	–	–	–	–	–
Fujita	init_AKT	flat	flat	–	–	–	–	–
Fujita	reaction_6_k1	flat	one side	–	–	–	–	–
Giordano	alpha_28	flat	flat	–	–	–	–	–
Giordano	alpha_4	one side	one side	–	–	–	–	–
Isensee	kf_RIIp_C_2__RII_C_2	identifiable	identifiable	–	–	–	–	–
Isensee	xi_kf_RII_C_2__RII_2	identifiable	identifiable	–	–	–	–	–
Laske	k_imp	identifiable	identifiable	–	–	–	–	–
Laske	k_syn_R_M	identifiable	identifiable	–	–	–	–	–
Liu	kappa	identifiable	identifiable	–	–	–	–	–
Liu	phi	identifiable	identifiable	–	–	–	–	–
Lucarelli	S2tot	identifiable	identifiable	–	–	–	–	–
Lucarelli	S3tot	identifiable	identifiable	–	–	–	–	–
Lucarelli	S4tot	identifiable	identifiable	–	–	–	–	–

Problem	Parameter	Profile (10)	Profile (20)	$r = 10$	5	3	2	1.5
Lucarelli	S_dephosphos	identifiable	identifiable	–	–	–	–	–
Lucarelli	S_phos	identifiable	identifiable	–	–	–	–	–
Lucarelli	geneF_turn	identifiable	identifiable	–	–	–	–	–
Lucarelli	geneH_turn	identifiable	identifiable	–	–	–	–	–
Lucarelli	geneK_turn	identifiable	identifiable	–	–	–	–	–
Lucarelli	init_Rec	identifiable	identifiable	–	–	–	–	–
Lucarelli	k_234	identifiable	identifiable	–	–	–	–	–
Okuonghae	gamma_i	identifiable	identifiable	–	–	–	–	–
Okuonghae	transmission_rate_effective	flat	flat	–	–	–	–	–
Oliveira	beta_0	flat	flat	–	–	–	–	–
Oliveira	h_hosp_rate	identifiable	identifiable	–	–	–	–	–
Perelson	c	flat	flat	–	–	–	–	–
Perelson	delta	flat	flat	–	–	–	–	–
Rahman	infected_moderate_transmission_rate	identifiable	identifiable	–	–	–	yes	yes
Rahman	infected_normal_worsen_rate	identifiable	identifiable	–	–	–	–	yes
Raia	SOCS3mRNA_production	identifiable	identifiable	–	–	–	–	–
Raia	init_Rec_i	identifiable	identifiable	–	–	–	–	–
Raimundez	ka_MAPK__MKN1	identifiable	identifiable	–	–	–	–	–
Raimundez	kdeg_membran__MKN1	identifiable	identifiable	–	–	–	–	–
Salazar-Cavazos	ratio_kpkd_Y1068__FREE	identifiable	identifiable	–	–	–	–	–
Salazar-Cavazos	ratio_kpkd_YN__FREE	identifiable	identifiable	–	–	–	–	–
Schwen	ini_R1	identifiable	identifiable	–	–	–	–	–
Schwen	ini_R2fold	identifiable	identifiable	–	–	–	–	–
Smith	k8	one side	one side	–	–	–	–	–
Smith	kminus7b	identifiable	identifiable	–	–	–	–	–
Sneyd	l2	identifiable	identifiable	–	–	–	–	–
Sneyd	l4	identifiable	identifiable	–	–	–	–	–
Weber	a21	identifiable	identifiable	–	–	–	–	–
Weber	s12	identifiable	identifiable	–	–	–	–	–
Zhao	R_Stage_II_Wuhan	identifiable	identifiable	–	–	–	–	–
Zhao	R_Stage_I_China	identifiable	identifiable	–	–	–	–	–
Zhao	gamma_1_Stage_I_Hubei	identifiable	identifiable	–	–	–	–	–
Zheng	k01_02	identifiable	identifiable	–	–	yes	yes	yes
Zheng	k13_12	identifiable	identifiable	–	–	yes	yes	yes

Supplementary Note: Identifiability of the subsets of other criteria

Question. Brun's collinearity index, profile likelihood and the conditioning tests of the proposed method all aim, in different ways, at subsets of identifiable parameters. Brun's index and the conditioning tests of the proposed method judge identifiability from the sensitivity matrix at the starting point, that is, locally and to first order; profile likelihood judges it from the profiles of the full model; Morris screening and pivoted QR do not include an explicit redundancy test against compensation. The main text shows that the subsets of these criteria differ in whether they reproduce the full model. This note asks whether they also differ in the practical identifiability of the parameters they select, measured in the same way for every criterion, and whether any criterion other than the proposed one obtains subsets that are at the same time identifiable and sufficient.

Design. For every problem and criterion, the subset of the criterion (selected on the training outputs, with the same number $k = |S|$ of biological parameters as the proposed method; Methods, *Comparison with other selection criteria*) was profiled with exactly the procedure used for the proposed subsets (Supplementary Note *Identifiability of the selected parameters*): the reduced problem has the selected and the calibration parameters free and all other parameters at θ_0; it is fitted to all measurements from θ_0; for every selected parameter, $\log|\theta_j|$ is fixed at ten displacements per side, from 0.05 to 4.6, and the other free parameters are re-estimated; points outside the bounds $[|\theta_0|/10, 10|\theta_0|]$ are not used; if a point improves the optimum the problem is re-fitted from it and the profile repeated (at most three rounds); and a parameter is identifiable if $\Delta\chi^2$ exceeds 3.84 on both sides. Where a criterion chose exactly the proposed subset, its profiles were reused (10 problems for Brun, 6 for Morris, 2 for profile likelihood and 1 for pivoted QR). The profiles were computed on a 128-core server (`scripts/run_identifiability_other_criteria.sh`, which calls `analysis/prepare_criteria_subsets.py` and `analysis/identifiability_profiles.py`); the tables are produced by `figures_tables/identif_criteria_tables.py`.

The comparison is made on the 27 problems in which every criterion has a subset: Blasi has no training/test split on which the other criteria could select, Chen was not included in the selections of the other criteria (Methods), Raimundez has no profile-likelihood subset, and Bertozzi is excluded as in the comparison of admissibility. For each criterion, the counts are (i) the selected parameters that are identifiable, (ii) the problems in which every selected parameter is identifiable, and (iii) the problems in which every selected parameter is identifiable *and* the subset is admissible at ±5% (Table 2 of the main text), and (ii) and (iii) are compared with the proposed subsets by the exact McNemar test over problems, with Holm adjustment over the four existing criteria. The proposed subsets selected on the training outputs are reported as well, on the 26 of these problems in which that selection gave an admissible subset.

Results. Supplementary Table S49 gives the summary and Supplementary Table S50 the result for every problem and criterion.

- *Identifiability alone separates only the criteria without a redundancy test.* 86% of the parameters selected by the proposed method were identifiable (57 of 66), 85% of Brun's (56), 86% of those of profile likelihood (57), 80% of Morris's (53) and 76% of those of pivoted QR (50). The number of problems in which every selected parameter is identifiable was 21 for the proposed subsets, 20 for Brun, 22 for profile likelihood, 18 for Morris and 14 for pivoted QR; none of these differences from the proposed method is significant after adjustment (Holm-adjusted $P = 1$, 1, 0.75 and 0.16).
- *The same identifiability is reached with different parameters.* Brun's subsets share two thirds of the parameters selected by the proposed method (46 of 70 over its 29 problems) and those of profile likelihood about two fifths (26 of 68), yet both are as identifiable as the proposed subsets: a model usually contains several groups of parameters that the data determine, and the criteria find different ones.
- *Requiring identifiability and sufficiency together separates the criteria.* 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 Morris's, 14 with those of profile likelihood and 7 with those of pivoted QR (Holm-adjusted McNemar $P = 0.38$, 0.047, 0.078 and 0.002; problems in which only the proposed subsets have both properties against problems in which only the other criterion has them: 4 against 1, 7 against 0, 8 against 1 and 15 against 1). Selected on the training outputs, the proposed subsets had both properties in 16 of 26 problems.

Interpretation. Profile likelihood chooses, by construction, the parameters with the narrowest confidence intervals, and its subsets are therefore as identifiable as the proposed subsets; but in a third of the problems they do not reproduce the full model when all parameters change (18 of 27 admissible). Brun's collinearity index, which judges identifiability only locally and to first order, gives subsets that are in practice as identifiable as the proposed subsets and almost as often sufficient, consistent with the close similarity between its ranking and the proposed one (Discussion). Morris screening and pivoted QR give less identifiable subsets that are also less often sufficient. The property that distinguishes the subsets of the proposed method is thus not identifiability by itself, which several criteria achieve, but the combination of identifiability with the verified ability to reproduce the full model, which is what the method tests by re-estimation.

Computing time. The profiles of the subsets of the four criteria took 41 h of processor time (Brun 4.7 h, Morris 13.1 h, profile likelihood 8.6 h, pivoted QR 14.4 h), most of it for Isensee, whose profile points take up to an hour each; on the 128-core server the whole analysis took about 2 h of elapsed time.

Limitations. The comparison uses the profiles of the reduced problem with the non-selected parameters at θ_0, as for the proposed subsets; it does not assess identifiability in the full model. The subsets of the other criteria were selected on the training outputs, whereas the profiles use all measurements, as for the proposed subsets. Pivoted QR, Morris screening and profile likelihood were applied as they are usually implemented (Supplementary Note *Brun and Morris selections*); other implementations could select different parameters.

Supplementary Table S49. Practical identifiability of the parameters selected by every criterion, with the profile-likelihood procedure of the proposed subsets (reduced problem, all data, $\Delta\chi^2 = 3.84$ on both sides), over the 27 problems in which every criterion has a subset. Identifiable: selected parameters whose 95% profile-likelihood interval is bounded on both sides within the bounds of the study. All identifiable: problems in which every selected parameter is identifiable. Admissible: the subset reproduces the full model at ±5% (Table 2 of the main text). Both: all parameters identifiable *and* the subset admissible. P: exact McNemar test against the proposed subsets (all outputs) over problems, for "all identifiable" and for "both", Holm-adjusted over the four existing criteria (for the proposed subsets selected on the training outputs, unadjusted). Bertozzi is excluded, as in the comparison of admissibility. Same as proposed: problems in which the criterion chose exactly the proposed subset, whose profiles are then those of the proposed subsets. The proposed subsets selected on the training outputs are shown only for the problems in which that selection gave an admissible subset.

Criterion	Problems	Identifiable	All identifiable	Admissible	Both	P (all id.; both)	Same as proposed
Proposed (all outputs)	27	57/66 (86%)	21	27	21	–	–
Proposed (training)	26	50/61 (82%)	19	23	16	0.5; 0.062	9
Brun	27	56/66 (85%)	20	25	18	1; 0.38	9
Morris	27	53/66 (80%)	18	22	14	0.75; 0.047	6
Profile likelihood	27	57/66 (86%)	22	18	14	1; 0.078	2
Pivoted QR	27	50/66 (76%)	14	16	7	0.16; 0.0021	1

Supplementary Table S50. Identifiable parameters / selected parameters for every problem and criterion (profile likelihood of the reduced problem). =, same subset as the proposed method (profiles shared); *, the subset is admissible at ±5%; –, no subset. Problems below the line are not common to all criteria and are not included in the summary of Supplementary Table S49.

Problem	Proposed (all outputs)	Proposed (training)	Brun	Morris	Profile likelihood	Pivoted QR
Alkan	2/2*	2/2*	2/2*	2/2*	2/2*	2/2
Armistead	2/2*	= 2/2*	= 2/2*	= 2/2*	2/2	2/2
Bachmann	2/2*	2/2*	2/2*	= 2/2*	2/2*	1/2*
Boehm	2/2*	= 2/2*	= 2/2*	= 2/2*	2/2	1/2
Borghans	2/2*	2/2*	2/2*	2/2*	2/2*	2/2*
Brannmark	2/2*	2/2*	2/2*	0/2*	2/2*	1/2*
Bruno	5/5*	5/5	5/5	5/5	5/5	5/5
Elowitz	2/2*	= 2/2*	= 2/2*	2/2	2/2	2/2
Fiedler	2/2*	1/2*	1/2*	1/2*	2/2*	1/2*
Fujita	0/2*	= 0/2*	= 0/2*	1/2*	2/2*	2/2*
Giordano	0/2*	0/2*	0/2*	0/2*	0/2	0/2
Isensee	2/2*	= 2/2*	1/2*	2/2*	2/2*	2/2*
Laske	2/2*	2/2*	2/2*	2/2*	2/2*	2/2*
Liu	2/2*	2/2*	2/2*	2/2*	2/2*	2/2*
Lucarelli	10/10*	8/8	10/10*	10/10	10/10	10/10
Okuonghae	1/2*	0/2*	1/2*	0/2*	0/2*	1/2*
Oliveira	1/2*	0/2*	1/2*	1/2*	1/2*	0/2*
Perelson	0/2*	–	= 0/2*	= 0/2*	= 0/2*	= 0/2*
Rahman	2/2*	2/2*	2/2*	2/2*	2/2*	1/2*
Raia	2/2*	1/2*	= 2/2*	2/2*	2/2*	2/2*
Salazar-Cavazos	2/2*	= 2/2*	= 2/2*	= 2/2*	2/2*	2/2
Schwen	2/2*	= 2/2*	2/2*	= 2/2*	= 2/2*	2/2*
Smith	1/2*	= 1/2*	2/2*	1/2*	0/2*	1/2*
Sneyd	2/2*	= 2/2*	= 2/2*	2/2	2/2	1/2
Weber	2/2*	2/2*	= 2/2*	2/2*	2/2*	1/2*
Zhao	3/3*	2/2	3/3	2/3	3/3	3/3
Zheng	2/2*	2/2*	2/2*	2/2*	2/2	1/2
Bertozzi	2/2	–	2/2	1/2	2/2	2/2

Problem	Proposed (all outputs)	Proposed (training)	Brun	Morris	Profile likelihood	Pivoted QR
Blasi	2/2	= 2/2	–	–	–	–
Chen	2/2	= 2/2	–	–	–	–
Raimundez	2/2*	= 2/2*	= 2/2*	2/2*	–	2/2*

Supplementary Note: Admissibility with the calibration parameters fixed

Question

In the admissibility test every parameter is perturbed, the calibration parameters Θ_C (scalings and offsets of the observables) included, and Θ_C is always re-estimated together with the subset. Part of the output change that the reduced model has to reproduce is therefore produced by Θ_C itself and can be undone by re-estimating it, which could make the test easier than it should be. The null baselines point in this direction: re-estimating Θ_C alone was admissible in 9 of the 14 comparable problems that have calibration parameters, and in these problems all random pairs were admissible too (main text, Table 2). This note measures how much of the output change is due to Θ_C and repeats the test without it.

Design

For the 14 problems with calibration parameters among the 28 of the comparison (Chen, which also has them, is excluded from the comparison because of its cost), the admissibility test at ±5% was repeated with the calibration parameters kept at θ_0: the 15 scenarios are those of the standard test with the components of Θ_C reset to their nominal values, so the perturbations of the biological parameters are exactly the same and the comparison is paired. The subset and Θ_C are still re-estimated, with the same settings and the same rule as in the comparison of criteria (fixed scenario set; nonlinear re-estimation, first order in Alkan, Bachmann, Isensee and Raimundez). The test was applied to the proposed subsets, to the proposed subsets selected on the training outputs, to the calibration parameters alone ($k = 0$, without the rule $|S| \geq 2$), to the 20 random k-subsets and to the subsets of Brun, Morris, profile likelihood and pivoted QR (27 subsets per problem, 378 single-core tasks; `analysis/criteria_tests.py` with `CT_CAL_FIXED=1`, summary by `analysis/calibration_fixed_summary.py`). For the proposed subsets, the fraction of $\|\Delta y_{\text{full}}\|^2$ of each standard scenario that the calibration parameters alone produce (all other parameters at θ_0) was also computed. In the other 14 problems of the comparison there are no calibration parameters and the test is unchanged.

Results

The calibration parameters produce a median of 30% of the output change of the standard scenarios (range 0–93%; Supplementary Table S51); their share is largest in Laske (93%), Isensee (84%), Smith (75%) and Schwen (71%). Removing their contribution raises the error of every subset, that of the proposed subsets by a median of 4.4 percentage points (from 7.6% to 23.0% in Isensee and from 10.1% to 36.6% in Laske).

The proposed subsets remained admissible in 13 of the 14 problems (Supplementary Table S52). The exception is Bachmann, whose median error, evaluated to first order, rose from 33.7% to 44.1%, just above $\tau_e = 43.6\%$. Calibration-only re-estimation, in contrast, became inadmissible in three of the nine problems in which it passed (Fiedler, Laske and Schwen): there it had reproduced mainly its own perturbation. Brun's subsets and the proposed training-output subsets lost one problem each, as the proposed subsets did (14 to 13), whereas Morris lost two (13 to 11), profile likelihood one (12 to 11 of 13) and pivoted QR three (12 to 9); and the admissible random draws fell from 220 to 192 of 280.

The clearest case is Laske: with the standard scenarios all 20 random pairs and the calibration parameters alone were admissible, so the test did not separate subsets there; with the calibration parameters fixed only 2 of the 20 random pairs and not the calibration parameters alone are admissible, whereas the proposed pair is (36.6%). Random draws also became rarer in Bachmann (8 to 1 of 20) and Elowitz (8 to 4). With the calibration parameters fixed, the test therefore separates subsets in five of the 14 problems (Alkan, Bachmann, Borghans, Elowitz, Laske; four with the standard scenarios), and the proposed subset is admissible in four of them.

In six problems (Brannmark, Fujita, Isensee, Raia, Raimundez and Smith) the calibration parameters alone remain admissible although they are no longer perturbed, and every random pair passes. In these models re-estimating the scalings of the observables compensates for most of the change produced by the biological parameters, so the measured outputs change mainly in scale; the test does not distinguish between subsets there, and the admissibility of the proposed subset in these problems says little about the choice of its parameters.

Conclusion

The perturbation of the calibration parameters makes the test easier, by a few percentage points of e_{fit} in most problems and substantially where they produce most of the output change. Without this help the proposed subsets remain admissible in 13 of 14 problems, and where the test separates subsets the selected pair is among the admissible ones. The counts of the null baselines in the main text should be read with this in mind: part of the admissibility of calibration-only re-estimation reflects its own perturbation, and in six problems almost any subset is admissible.

Programs and reproduction

The whole test is run, from the folder `reproducibility/` of the package, with `NPROC=90 bash scripts/run_test_calibration_fixed.sh`. The script launches `analysis/criteria_tests.py` with `CT_CAL_FIXED=1`, `CT_PRUEBAS=adm` and the 27 subsets of every problem, distributing the 378 tasks over `NPROC` single-core processes with the slowest problems first; it is resumable, because finished tasks are skipped. Each task writes one file `results/criteria_tests_calfixed/<problem>__<subset>__adm.json`, whose fields are those of the standard test plus `calibracion_fija` and, for the proposed subsets, `frac_cambio_calibracion` (`DATA_DICTIONARY.md`). When all tasks have finished, `analysis/calibration_fixed_summary.py` compares them with the standard results of `results/criteria_tests_fixedscen/` and writes `results/calibration_fixed_summary.json` and `results/logs_calibration_fixed/summary.txt`. Supplementary Tables S51 and S52 are produced from these files by `figures_tables/calibration_fixed_tables.py` (part of `build_all.py`), and the computing time by `figures_tables/make_timing_notes.py` from the field `segundos` of each file.

Computing time

The 378 tasks took 16.7 h of processor time and 34 min of elapsed time on 90 cores of a 192-core server (`scripts/run_test_calibration_fixed.sh`, `NPROC=90`); Smith, the slowest problem (34 min per task in the standard test), set the elapsed time. Per problem, the processor time ranged from about 2 min (Elowitz) to 10.6 h (Smith; median 8.7 min; Supplementary Table S41, column *Calibration fixed*); this test adds 16.7 h, 2.4% of the processor time of all the analyses of the paper (697 h; Supplementary Table S37).

Supplementary Table S51. Admissibility at ±5% with the standard scenarios (Std.: all parameters perturbed) and with the calibration parameters kept at θ_0 (Fixed: only the biological parameters perturbed, with the same biological perturbations), for the 14 problems with calibration parameters. $|\Theta_C|$: calibration parameters; Share: median over the standard scenarios of the fraction of $\|\Delta y_{\text{full}}\|^2$ produced by the calibration parameters alone (%); the proposed subset and calibration only ($k = 0$, without the rule $|S| \geq 2$): median e_{fit} over the 15 scenarios (%; bold, admissible, $\leq \tau_e = 43.6\%$; first order in Alkan, Bachmann, Isensee and Raimundez, as for every criterion); random: admissible draws of the 20 random k-subsets.

			The proposed subset		Calibration only		Random	
Problem	$\|\Theta_C\|$	Share	Std.	Fixed	Std.	Fixed	Std.	Fixed
Alkan	10	7	**12.1**	**12.5**	94.8	98.6	1/20	1/20
Bachmann	74	30	**33.7**	44.1	50.5	68.7	8/20	1/20
Borghans	2	0	**4.3**	**4.5**	54.8	54.0	4/20	7/20
Brannmark	3	22	**7.8**	**12.1**	**15.9**	**21.9**	20/20	20/20
Elowitz	2	11	**35.8**	**40.5**	76.9	89.1	8/20	4/20
Fiedler	8	33	**14.7**	**14.7**	**42.4**	50.3	20/20	19/20
Fujita	3	22	**29.4**	**35.6**	**29.6**	**38.0**	20/20	20/20
Isensee	11	84	**7.6**	**23.0**	**10.3**	**26.7**	20/20	20/20
Laske	2	93	**10.1**	**36.6**	**21.7**	78.1	20/20	2/20
Raia	5	50	**3.3**	**3.4**	**7.6**	**8.3**	20/20	20/20
Raimundez	79	30	**18.7**	**21.5**	**28.3**	**32.2**	20/20	20/20
Schwen	15	71	**17.1**	**29.2**	**27.3**	57.1	20/20	20/20
Smith	9	75	**7.8**	**17.4**	**13.7**	**36.5**	20/20	20/20
Weber	5	4	**4.1**	**3.9**	48.3	54.7	19/20	18/20

Supplementary Table S52. Admissible subsets at ±5% in the 14 problems with calibration parameters, with the standard scenarios and with the calibration parameters kept at θ_0 (profile likelihood: 13 problems; no subset for Raimundez).

Subsets	Standard	Calibration fixed
The proposed method (all outputs)	14/14	13/14
The proposed method (training outputs)	14/14	13/14
Brun	14/14	13/14
Morris	13/14	11/14
Profile likelihood	12/13	11/13
Pivoted QR	12/14	9/14
Calibration only ($k = 0$)	9/14	6/14
Random k-subsets (all draws)	220/280	192/280

Supplementary Table S47. Synthetic prediction for three families of truths. Fraction of the cases with signal (common to all criteria of each block) in which $e_{\text{pred}} \leq \tau_e$; difference between the reference and each criterion in percentage points, with a 95% confidence interval from resampling whole problems (10,000 replicates) and the Holm-adjusted two-sided P over the four existing criteria (the earlier QR variant and the comparison between the two proposed selections are unadjusted). Uniform: $\theta^* = \theta_0(1 + 0.2u)$, 20 seeds; log-uniform: $\theta_j^* = \theta_{0,j}10^{v_j}$ with v_j uniform in $[-0.3, 0.3]$, ten seeds; sparse: 15% of the estimable parameters changed by ±50%, ten seeds.

Family	Criterion	Rate	Difference (95% CI)	Holm P
Uniform ±20%, reference: the proposed method (all outputs); 281 cases, 24 problems				
	The proposed method (all outputs)	56%	–	–
	Brun	55%	+1 (−1 to +3)	0.433
	Morris	48%	+8 (+2 to +16)	0.010
	Profile likelihood	39%	+17 (+7 to +28)	0.002
	Pivoted QR (GKS)	30%	+26 (+12 to +40)	< 0.001
	Pivoted QR (earlier)	34%	+22 (+10 to +34)	< 0.001 (unadj.)
Uniform ±20%, reference: the proposed method (training outputs); 280 cases, 23 problems				
	The proposed method (training outputs)	51%	–	–
	Brun	55%	−4 (−12 to +1)	0.412
	Morris	48%	+3 (−7 to +13)	0.603
	Profile likelihood	39%	+12 (+3 to +21)	0.022
	Pivoted QR (GKS)	30%	+21 (+6 to +35)	0.022
	Pivoted QR (earlier)	34%	+17 (+2 to +30)	0.023 (unadj.)
	The proposed method (all outputs)	56%	−5 (−12 to 0)	0.038 (unadj.)
Log-uniform, reference: the proposed method (all outputs); 217 cases, 25 problems				
	The proposed method (all outputs)	50%	–	–
	Brun	50%	0 (−3 to +2)	1.000
	Morris	47%	+2 (−3 to +9)	1.000
	Profile likelihood	37%	+13 (+4 to +23)	0.018
	Pivoted QR (GKS)	24%	+25 (+14 to +37)	< 0.001
	Pivoted QR (earlier)	29%	+21 (+9 to +34)	< 0.001 (unadj.)
Log-uniform, reference: the proposed method (training outputs); 217 cases, 25 problems				
	The proposed method (training outputs)	48%	–	–
	Brun	50%	−2 (−8 to +1)	0.958
	Morris	47%	0 (−7 to +8)	0.958
	Profile likelihood	37%	+11 (+2 to +20)	0.046
	Pivoted QR (GKS)	24%	+24 (+12 to +36)	< 0.001
	Pivoted QR (earlier)	29%	+19 (+5 to +34)	0.007 (unadj.)
	The proposed method (all outputs)	50%	−2 (−8 to +3)	0.614 (unadj.)
Sparse, reference: the proposed method (all outputs); 147 cases, 24 problems				
	The proposed method (all outputs)	54%	–	–
	Brun	52%	+2 (−1 to +6)	0.312
	Morris	46%	+8 (+3 to +14)	0.003
	Profile likelihood	40%	+14 (+4 to +24)	0.008
	Pivoted QR (GKS)	39%	+15 (+4 to +27)	0.019
	Pivoted QR (earlier)	44%	+10 (−1 to +21)	0.101 (unadj.)
Sparse, reference: the proposed method (training outputs); 147 cases, 24 problems				
	The proposed method (training outputs)	55%	–	–
	Brun	52%	+3 (−3 to +11)	0.420
	Morris	46%	+10 (+2 to +18)	0.028
	Profile likelihood	40%	+15 (+5 to +26)	0.009
	Pivoted QR (GKS)	39%	+16 (+5 to +28)	0.023
	Pivoted QR (earlier)	44%	+11 (−3 to +24)	0.140 (unadj.)
	The proposed method (all outputs)	54%	+1 (−6 to +10)	0.862 (unadj.)

Supplementary Note: Concluding remarks

This Supplementary Information was written so that every claim of the main text can be traced to its source and checked independently. Supplementary Tables S1 and S2 give the subsets selected for each benchmark problem and the notation of the method, and Supplementary Note *Verification of the simulator* compares the simulation of the measured outputs with the reference simulations of the benchmark. Supplementary Note *Prediction of held-out experiments* and Supplementary Fig. S1 show the prediction of held-out experiments on an absolute scale. These materials connect the results reported in the paper with the data and code from which they were obtained.

The remaining material documents the formal verification and, just as importantly, its scope. Supplementary Fig. S3 and Supplementary Table S5 separate three levels of evidence: general mathematical guarantees proved in Lean 4 (G1–G4 and G6), hypotheses proved for specific models (G5), and numerical conditions computed in Python, on which the application of the guarantees to a particular result depends. Supplementary Table S4 and Supplementary Note *Logical sequence of the formal proofs* follow the 32 modules of the formalisation in the order in which they depend on each other, from the energy bound on the error of the reduced model and the relation between $\cos\Delta$ and the best rescaled fit, through the differentiability of ODE solutions with respect to the parameters, to certificates for ODE models and the hypotheses of the benchmark models; they also state what the proofs do not establish, such as identifiability of the full nonlinear models or the convergence of the numerical optimisation; for the selected parameters of the reduced problem, step 32 gives a sufficient condition for practical identifiability whose hypotheses are checked numerically at the computed profile points (Supplementary Note *Identifiability of the selected parameters*). Supplementary Note *Statements of the formal theorems* and Supplementary Tables S6 and S7 give the statement and meaning of each of the 309 theorems and lemmas, so that every formal statement can be compared with the mathematics it encodes. Supplementary Note *Verified benchmark models* and Supplementary Tables S8 and S9 show, model by model, the data on which the Lean kernel executed its checks, a complete worked example, an independent check that the translated equations agree with the SBML models, and the reasons why seven benchmark problems were not verified. Supplementary Note *Reproducibility package* explains how to reproduce all of this. The numerical analyses cited in the main text are documented in full: Supplementary Note *Verification of the simulator* compares the simulator of this work with the reference simulations of the benchmark, Supplementary Note *Robustness to the size of parameter changes* gives the distribution of the fitting error over the paired random scenarios at every perturbation size, Supplementary Note *Synthetic prediction experiment* reports the prediction of held-out data when the truth is known, Supplementary Note *Single-parameter subsets* examines the minimum subset size of two parameters, a design choice (a single parameter is already admissible in 15 of 27 problems), Supplementary Note *Sensitivity to the method settings* gives the effect of every setting of the method on every problem, and Supplementary Note *Computational cost* the computational cost. Supplementary Note *System-by-system summary* gives a summary of each of the 34 benchmark problems: what the model describes, what the selected parameters mean in its reactions, how well the reduced model reproduces the full one and how it predicted held-out data.

Taken together, the formalisation establishes that the steps of the method are mathematically sound under explicit hypotheses and that 28 benchmark models satisfy the hypotheses on the dynamics; the numerical quantities reported in the paper are computed in floating point by the accompanying Python code, and the formal results state precisely which conclusions follow from them. Everything described here can be reproduced from the repository: the proofs compile with a single command, the supplementary tables and notes are generated from the Lean sources and the result files, and the analyses can be rerun from the published benchmark. This combination of a public benchmark, open code and machine-checked mathematics, together with a clear statement of what each part guarantees, makes the reduction method easy to verify, to reuse and to extend to other dynamical models.