Strategies for calibrating models of biology

Mark N. Read [1, 2, *]

Kieran Alden [3]

Jon Timmis [3]

Paul S. Andrews [4]

1. The School of Life and Environmental Sciences, The University of Sydney, Australia.

2. The Charles Perkins Centre, The University of Sydney, Australia.

3. Department of Electronic Engineering, The University of York, UK.

4. SimOmics Ltd, Suite 10 IT Centre, Innovation Way, York, YO10 5NP, UK.

Corresponding Author Contact Details:

* mark.read@sydney.edu.au

+61 416 282 513

Dr. Mark Read,

The Charles Perkins Centre,

John Hopkins Drive, Camperdown NSW 2006,

Australia

Keywords:

Computational Biology

Systems Biology

Model Fitting

Model Selection

Model Calibration

Parameter Tuning

Abstract:

Computational and mathematical modelling has become a valuable tool for investigating biological systems. Modelling enables prediction of how biological components interact to deliver system-level properties, and extrapolation of biological system performance to contexts and experimental conditions where this is unknown. A model's value hinges on knowing that it faithfully represents the biology under the contexts of use, or clearly ascertaining otherwise and thus motivating further model refinement. These qualities are evaluated through *calibration*, typically formulated as identifying model parameter values that align model and biological behaviours as measured through a metric applied to both. Calibration is critical to modelling, but is often under-appreciated. A failure to appropriately calibrate risks unrepresentative models that generate erroneous insights. Here we review a suite of strategies to more rigorously challenge a model's representation of a biological system. All are motivated by features of biological systems, and illustrative examples are drawn from the modelling literature. We examine the calibration of a model against distributions of biological behaviours or outcomes, not only average values. We argue for calibration even where model parameter values are experimentally ascertained. We explore how single metrics can be non-distinguishing for complex systems, with multiple component dynamic and interaction configurations giving rise to the same metric output. Under these conditions, calibration is insufficiently constraining and the model *non-identifiable*: multiple solutions to the calibration problem exist. We draw an analogy to curve fitting and argue that calibrating a biological model against a single experiment or context is akin to curve fitting against a single data point. Though useful for communicating model results, we explore how metrics that quantify heavily emergent properties may not be suitable for use

in calibration. Lastly, we consider the role of sensitivity and uncertainty analysis in calibration and the interpretation of model results. Our goal in this manuscript is to encourage deeper a consideration of calibration, and how to increase its capacity to either deliver faithful models or demonstrate them otherwise.

Main Body

Introduction

Explored in ever greater depth, our appreciation of the sheer inherent complexity of biological systems grows [1]. Important system-level properties, such as disease and health, emerge from local interactions amongst vast numbers of molecular and cellular components [2]. These emergent properties cannot be predicted from examination of system components in isolation [3,4]. As immunologist Irun Cohen reflects: “The more data we have access to, the more confused we have become” [5]. Consequentially, computational and mathematical modelling has arisen as a “constructionist” tool that complements reductionist techniques [6]. Models reveal how local-level component dynamics impact system-level performance, and can predict real system behaviour in a given context. That anything can be manipulated or measured in a model is arguably its greatest upside [7,8]. The greatest downside is the risk that a model is unrepresentative of the biology, and hence misleads rather than informs.

Two aspects of biological modelling can give rise to unrepresentative models, and complicate the interpretation of modelling results. First, models are abstract representations of the systems they seek to capture. Despite advances in multi-scale modelling [9,10], it is currently both technologically and conceptually impossible to fully model a biological system, from the molecule, through the cell and then organism, to the ecosystem. Low-level process dynamics are abstracted and summarily represented, e.g. as

distributions. Many biological components are similarly omitted, amalgamated or abstracted, and together they are represented in a simplified subset of the full spatial environmental. It is rarely known *a priori* what the most appropriate abstractions of a biological system will be [11]. Some mathematical models can capture realistic population sizes of system components. For others, such as agent-based modelling wherein every individual component of the model is explicitly represented [12], this is impractical. Second, the biological system is likely incompletely understood; gaining a greater understanding may in fact motivate the modelling enterprise. How, then, does one construct a representative model of such a system? The implication is that we cannot assume a direct mapping of biological- to model-component(s), neither in terms of component concept, dynamics nor population size. Rather, establishing and testing this mapping is accomplished through model *calibration* [7].

We distinguish between model *mechanics* and *parameters*. A model's components, and the dynamics they are capable of, represent its mechanics. A model's parameters assign rates and probabilities to component dynamics, and values to initial conditions. At the very least, calibration must seek to establish suitable parameter values. Ideally, however, calibration should be capable of detecting inappropriate model mechanics, and thus unrepresentative models.

Calibration approaches typically seek parameter values that align model and real-world behaviours and outcomes under some single context or experimental condition. Approaches range from manual exploration of putative parameter values, drawing heavily on domain expertise [13,14], to automation. Domain experts can help overcome a lack of available

data, or conflicts and inconsistencies therein [15], but can prove inferior to systematic approaches [14,16]. Automated calibration approaches include the use of heuristic search algorithms [17,18], Bayesian methods such as Markov chain Monte Carlo and maximum likelihood [4,19], Kalman filters [20], and standard curve fitting techniques [21,22].

Here we explore strategies aimed at increasing the power of calibration to either deliver representative, accurate models, or demonstrate them otherwise. Our aim is not to review the specific fitting technologies cited above, but to explore the broader contexts through which calibration can be performed.

Diverse sources for model parameter values and mechanics

Modellers can industriously delve through literature seeking experimental sources for model parameter values. For instance, Efroni *et al.* report sourcing parameter values from around 300 papers in their modelling of thymic T cell maturation [23]. We posit that, even where putative values for all model parameters can be extracted from the literature, adjustment and tuning through calibration is still warranted. In reproducing biological system dynamics, components included within a model must compensate for the activities of those that are omitted or abstracted. For instance, as key mediators of the adaptive immune response, T cells feature heavily in immunological models [24,25]. Currently, 29 distinct subsets of T cell are recognised [26], encompassing a functional richness and nuance beyond most models' aspirations. A modelled "T cell" may not correspond exactly with any single subset, but rather an integration of several. Hence, putative model parameters

originating from measurements of a specific subset may still warrant adjustment to account for abstraction. This principle is further compounded if data derives from different experimental systems, be they different organisms or systems of study, e.g. humans, mice or zebra fish, *in vivo*, *in vitro* or other *in silico* models. For instance, *in vivo*, *ex vivo* and *in vitro* studies have found differing effects of stress on NK cell cytotoxicity [27]. Elsewhere, differing transcriptomics results between *in vitro* and *in vivo* toxicology studies pose a challenge for the pharmaceutical industry [28]. Lastly, cell migration patterns can differ substantially between *in vivo* and *in vitro* contexts [29].

Responses to intervention can also differ considerably across organisms, reflecting mechanistic differences in biological components. For instance, in a study of over 100 strains of mouse commonly used in research, the development of insulin resistance, and associated metabolism and physiology, in response to high fat-high sucrose diets was found to vary by over an order of magnitude [30]. Further, the effect of caloric restriction on lifespan, body weight, core body temperature, insulin sensitivity, metabolism, and pathology has been found to differ between mouse strains and sexes [31,32].

When acquired from multiple experimental systems, particularly those other than the specific system being modelled, putative parameter values are better thought of as guidelines than prescriptions. Yet calibration is not always reported in modelling literature, even when estimates for particular parameters vary by several orders of magnitude, e.g. Swerdlin *et al.* cite literature reporting the frequency of B cells specific for a given antigen to lie in the range of 1 in 10k to 1 in 1,000K [33].

An alternative to obtaining parameter values from highly diverse sources is to extensively interrogate the exact experimental system being modelled, and calibrate parameters against these data. For instance, in modelling hematopoietic stem cell expansion dynamics under given cytokine exposure, Gullo *et al*. acquired extensive experimental data of expansion from the exact *in vitro* system they were modelling for use in calibration [34]. To calibrate their model of the immune response in murine tuberculosis, Marino *et al*. quantified four key leukocyte population sizes in the lymph nodes and lungs at 7 time points post-infection obtained from 80 mice [35]. From 5000 samples of model parameter value space, the single set of values most closely (qualitatively) resembling *in vivo* dynamics was selected as the starting point for calibration. 130 parameters, those related to leukocyte dynamics, were calibrated; the remaining 80 retained these starting values. This practice can mitigate the diversity of experimental systems from whence parameter values are drawn as a source of uncertainty or error, though ideally sufficient data is extracted to calibrate *all* model parameters. Putative parameter values sourced from elsewhere can form initial values or boundaries of exploration for calibration.

Calibrating to reproduce distributions of outcomes

Stochasticity is a quality ubiquitous across Biology. Through it, identical interventions can generate different outcomes, even in genetically identical individuals [36]. Recent statistical meta-analyses have highlighted the importance of the *variance* in outcomes, not just their mean average. For instance, a low carbohydrate (LC) dietary weight-loss intervention can outperform caloric restriction (CR) on *average*, but in so doing incurs a larger spread of

outcomes [37]; a greater proportion of individuals *gained* weight on LC diet than CR, despite LC's lower average. Rather than determining LC the superior intervention, we must understand the mechanisms at play, and stratify patients onto interventions optimal for them as individuals. Similar trends exist for the effects of dietary interventions on lifespan, arguably an even more critical outcome [38], and the number of food sources on evolutionary fitness [39]. The spread of data is important, and stochastic models can offer insight into the mechanisms generating it. This requires calibration against the spread, not just the mean. The Kolmogorov-Smirnov (KS) statistic represents a powerful tool for accomplishing this, Figure 1. This non-parametric statistic is sensitive to differences anywhere across two distributions, not only their averages. As illustration, Read *et al.* employed the KS statistic to align *in silico* with *in vivo* distributions of leukocyte motility characteristics [40]. Only by explicitly modelling and tuning cellular heterogeneity could population-level dynamics be reproduced; some cells are inherently faster and more directional than others.

We advocate for effect size measures rather than statistical significance (p values) in stochastic model calibration [41]. Unless the distributions being contrasted arise from the exact same process, statistical significance is always attainable given sufficient replicates, Figure 1C. As abstract representations from which additional replicates are trivially obtained, models will never be statistically indistinguishable from the biology. Rather than aspire to perfect (unattainable) model alignment with biology, we suggest expressing an acceptable tolerance in terms of effect size, which is relatively invariant to increasing replicates, Figure 1C. Example effect size statistics include: the KS statistic, described in the Figure 1 caption; the Vargha-Delaney "A" statistic, which quantifies the probability that a

randomly selected sample from one distribution is larger than a randomly selected sample from the other, and for which guidelines equating A scores to “small”, “medium” or “large” effects exist [42]; and Cohen’s “δ” statistic, the difference between two distributions’ means divided by their pooled standard deviation.

Robust stochastic model calibration requires that any observed disparities can be attributed to inappropriate model parameter values or mechanisms, rather than artefacts arising from an insufficient sample size [15]. However, calibration can entail exploring numerous points in parameter space, and acquiring large sample sizes (model executions) for each point can compound into considerable computational expense, particularly for agent-based models. Modellers may wish to select a statistical precision (sample size) in accordance with the computational capacity available. A technique to quantify how many replicates are needed to reduce the contribution of stochasticity to a desired level was derived in [15], and implemented in [43]. It entailed contrasting *n* (e.g. 20) groups of model execution replicates, all generated under identical parameter values, using the Vargha-Delaney “A” statistic. This procedure was repeated with varying numbers of replicates, after which the maximum “A” statistic score obtained amongst group comparisons was plotted against sample size used. The authors sought the minimum sample size delivering a maximum “A” statistic score indicating a “small” effect [15]. This technique essentially performs “mock” parameter adjustments, therein quantifying the portion of an observed difference in model output (when calibrating proper) attributable to model stochasticity for the sample size used. There is no value in attempting to calibrate beyond this baseline level of alignment: any improvements would likely represent sampling artefacts rather than an improved model.

The “A” statistic was used because it formed the basis of subsequent analyses; any other statistic more relevant to the given calibration effort can be substituted.

Single metrics alone may not fully distinguish complex systems

Most conventional model fitting techniques require a single metric to quantify the difference between modelled and biological dynamics. However, complex biological systems are not necessarily quantifiable through single metrics alone. They constitute numerous types of interacting component with many-to-many mappings between components and functions, and encompass both positive and negative feedback loops [44,45]. A single metric may not be fully distinguishing in such a system. This can render a system, and a putative model thereof, *non-identifiable* [20]: multiple distinct component configurations and dynamics can produce the same value. For instance, motility research has classically employed “mean squared displacement” to quantify how far agents travel on average over time [46]. Typically used to characterise search behaviour on a scale of “localised” to “highly-directional”, this metric is nonetheless non-distinguishing. Slow-moving directional agents can yield the same mean squared displacement as fast-moving non-directional agents, and yet the spatial coverage, and hence interactions, of these agents can differ vastly [47,48].

In lieu of a more distinguishing metric, the only recourse is to instead employ a suite of them [49]. Yet, this is not readily applicable to standard model fitting techniques given their reliance on single metrics. Multi-objective optimisation (MOO) technologies can offer a

solution [50]. The differences between target biological and model dynamics, as measured by each metric, are posed as separate objectives to be minimised through MOO, Figure 2A [14,51]. MOO typically employs a guided search to find solutions, in this case parameter values, that best satisfy the objectives. Given its abstractive nature, no single set of parameter values will likely provide a perfect model alignment across all metrics. Rather, tradeoffs will exist, where given parameter values will provide better alignment on some metrics than others, Figure 2B. MOO provides a *Pareto front* of optimal parameter value sets: those for which performance improvement in any one metric necessitates worsening in another. Knowledge of these tradeoffs can inform modellers' adoption of parameter values from calibration [14,49].

Alternatively, subsequent experiments can be replicated using all Pareto front parameter values, thereby exposing the extent to which results hinge on calibration choices rather than the intervention [51].

MOO-based calibration has found extensive application outside of Biology, in the calibration of hydrological models [52]; water basin properties are not sufficiently constrained by single metrics alone [49]. Elsewhere, Newland *et al.* calibrate land-use models through MOO, employing objectives quantifying capture of absolute location and patterns in land use [14]. In a Biological context, we demonstrated MOO-based calibration for models of murine multiple sclerosis [51] and leukocyte search behaviour [40]. The former employed metrics capturing the proliferation dynamics of several leukocyte populations involved in disease onset and subsequent recovery, recapitulating a prior expert-informed manual calibration effort [24]. The latter overcame the aforementioned mean squared displacement drawbacks by using a complementary suite of motility metrics.

Imperative in model fitting is the detection of *overfitting*, wherein the model captures not only general biological behaviours but also the specific noise nuances of the (*training*) dataset against which it is fitted. The standard approach of detecting overfitting through a separate, independent *validation* dataset not used for training can be ported to the multi-objective context [51]. Separate Pareto fronts are maintained with respect to both datasets, and overfitting is indicated through the proportion of training dataset Pareto front member solutions that *are not* also members of the validation dataset Pareto front, Figure 2C.

MOO-based calibration approach can facilitate model selection by contrasting the Pareto fronts generated under competing putative models [40,51], Figure 2D. The superior model's Pareto front will contain solutions offering better tradeoffs than alternatives. However, there currently exists no formal framework to control for model complexity under this approach, such as in the Akaike information criteria. Hence, the technique may simply select the most complex model. This can be mitigated by formulating the number of model parameters, or some other measure of model complexity, as an objective to be minimised [53]. The tradeoff of better biological capture versus model complexity is thus made explicit, allowing modellers to select accordingly.

The "NSGA-II" algorithm is a very popular MOO implementation [54]. However, the more recent Universal-NSGA-III improves scalability to many objectives, which can otherwise lead to a cumbersomely large number of similar solutions [55], Figure 2E. We provide an open-source implementation of Universal-NSGA-III at *https://github.com/marknormanread/unsga3*. The NSGA algorithms generate putative

solutions through a genetic algorithm, employing natural selection and genetic recombination processes to iteratively improve quality. There exist other solution-generation strategies compatible with the Pareto front concept, as reviewed in [56].

Calibrating against multiple scenarios

In curve fitting and regression, one attempts to fit (well-understood) equations to data, predicting how a dependent variable varies with one or more independent variables/predictors. Curve-fitting canon prescribes two data points to fit a line, three for a curve, and more if the data is noisy. The fitted equation permits extrapolation between and beyond known data points, and reasoning over how predictors drive system response. Herein exist clear parallels to modelling biological systems. With the model we aim to predict the biological system's behaviour (dependent variable) under a given context, and gain mechanistic understanding. "Context" here encompasses the abstract space of possible experiments, conditions or environments to which a biological system can be subjected, which equates to predictors. For simplicity we refer to points in this space as *scenarios*. This analogy highlights the prudence of "fitting" the model against multiple scenarios, Figure 3. Otherwise, how can we be confident that the model will yield accurate predictions when used in predicting biological behaviour beyond what is already known? Any equation, and arguably, incorrect model, can be fitted to a single data point. Incorrect models can be identified as such through the heavy modifications to mechanics and parameter values they require to reproduce behaviour under each scenario. Such modifications compensate for the model's inadequate biological capture.

Yet, rarely does modelling literature report model calibration procedures, let alone against multiple scenarios. Model fitting against single scenarios and demonstrating the ability to reproduce an independent experiment's outcomes, not used in calibration, is sometimes reported [58,59]. This is akin to the training-test/validation splits commonly used in machine learning to detect over-fitting of statistical models, and we consider it good practice. For instance, following calibration, Palumbo *et al*. validated their model of energy homeostasis in relation to physical exercise against six independent real-world datasets, encompassing 69 demographically diverse study participants [60]. One might argue that fewer calibration data points are required in biological modelling, owing to *a-priori* biological knowledge informing model design. We counter this position: one cannot *guarantee* that a model is appropriate, and calibration is a valuable independent test thereof.

Though not reported in the context of model calibration, Bloch and Harel verified that their model of tumour growth matched several known biological cases [61]. These include tumour growth being contingent on angiogenesis, non-cancerous cell viability in the absence of angiogenesis, and tumour angiogenesis being impeded if blood vessels were too distant. Kamal *et al.* calibrated their model of influenza and antiviral recombination therapy in two stages: firstly against placebo patient data, after which additional parameters were estimated by fitting against drug-receiving patient data [62]. Gullo *et al*. also report a multi-stage calibration for their model of *in vitro* hematopoietic stem cell expansion under a given cytokine milieu [34]. Their work is particularly notable for the sheer number of experimental scenarios, representing combinations of cytokine exposures, informing their calibration: 20

combinations, each performed *in vitro*, in triplicate. Successfully calibrating against such a comprehensive exploration of possible cytokine exposures instils considerable confidence that their subsequent *in silico* experiments are trustworthy. In developing a computational model of the murine autoimmune disease Experimental Autoimmune Encephalomyelitis, we simultaneously calibrated against two experimental scenarios [24]: physiological recovery of mice post-induction of disease, and a hindered recovery following targeted ablation of a given T cell population. We evaluated each putative parameter value set against both scenarios simultaneously, seeking values that provided alignment in both experiments. Our calibration effort was manual, guided by a domain expert [15]. In hindsight, we believe a more thorough, and certainly less time-consuming, calibration could have been accomplished through application of MOO, with each metric used in each scenario comprising an objective. Herein, a single set of parameter values that accurately reproduces target behaviours with a given model can be considered a successful calibration. Alternatively, one could use conventional single-metric calibration techniques for each scenario independently, seeking a single set of parameter values providing good alignment across all calibration exercises.

Calibrating with measures of emergent properties

Quantifying a model in the same terms as the real biological system can facilitate adoption and interpretation of modelling results [63]. The development of powerful 3D model visualisation engines attests to this [64]. A model that is “seen” to look like the real system can help biologists appreciate what the model shows, and raise confidence that it is

representative of the biology [2,65]. It has been argued that similar quantifications of the real biology can be applied to models and used in their calibration [63]. Yet the metrics through which biological systems are quantified range from direct assays of individual systems components to observations of whole-system state (emergent properties). For instance, cell counting through flow cytometry is a fairly direct measure of system components. Likewise for histology and imaging, though in all three cases expression measurements can be biased by differing marker affinities for their targets [66]. At the other extreme lies metrics such at those used in disease model quantification. In the murine autoimmune disease Experimental Autoimmune Encephalomyelitis (EAE), disease severity is scored on a scale of 0 to 5, capturing the progression of paralysis from tail through hind-, then fore-, legs, ultimately culminating in death [67]. The K/BxN arthritis mouse model scores the severity in each paw from 0-3 and takes the sum thereof [68].

The degree to which such metrics can enable calibration depends on how directly they relate to a model's components, and here these metrics occupy a spectrum. Exemplifying one extreme, Butler *et al.* developed a tool-chain facilitating flow cytometry, protein expression heatmap and histology analyses for a model of pre-natal lymphoid organ development [63]. These metrics were directly tied to processes explicit in the spatial agent-based model. Further along the spectrum, calibration of Cilfone *et al.* and Warsinske *et al.*s' granuloma models necessitated their development of a transformation from 2D modelled microbial load to 3D real-world equivalents [69,70]. Finally, at the other extreme lies Read *et al.*'s EAE model disease severity scoring metric [24]. EAE paralysis emerges from the effects of molecular-level interference with neurons that spans the entire central nervous system. This level of detail and spatial scale, from the molecular to the whole-organism, far

exceeded the model's scope. Instead, modelled EAE severity scoring was based on the exposure of neurons to pro-inflammatory cytokines, which they believed would serve as a proxy. The result was a complicated fitting exercise requiring data smoothing, Fourier transforms, and, importantly, a calibrated model deemed representative of two real-world experiments where rates and degrees of fluctuations in mouse paralysis were known [71].

All these metrics proved useful for interpreting modelling results. However, the degree to which they can be used in calibration depends on how directly they pertain to directly observable modelled components. Metrics are less readily applicable in model calibration if they *themselves* require calibration, particularly if this necessitates an *a priori* calibrated model.

Sensitivity and uncertainty analyses aid calibration and guide interpretation

Complex models comprising many parameters can prove challenging to calibrate. For instance, if real-world data against which to calibrate is scarce with respect to model complexity, the problem may be non-identifiable: more data is needed to constrain all the model's free variables. *Sensitivity analysis* (SA) can help target the calibration effort at the most influential subset of model parameters [13,16,72]. Less influential parameters are "fixed," assigning values measured in the real system or extracted from the literature where possible.

SA describes a suite of analyses that quantify how variation of a model's output (*response)* relates to variation of its inputs (*samples* of model parameter space) [73,74], Figure 4A. In *global* SA, all (or many) parameters are perturbed simultaneously. Alternative *one at a time* SA approaches vary single parameters whilst holding others to some default value; these are best applied when those default values are well-motivated, often only after calibration has been performed. Global SA can quantify which parameters are most influential on model behaviour, and is sensitive to compound effects where one parameter's influence is dependent on values held by others. SA consists of a strategy to systematically vary parameter values coupled with a way of quantifying these with changes in output. A simple strategy to vary inputs is a *factorial design*, Figure 4B, in which a set of values is selected for individual each parameter, and their cartesian product then forms combinations of parameter values to explore: all possible combinations are utilised. This sampling strategy is robust to sampling artefacts, and its comprehensive exploration of parameter space yields highly representative results. However, it also incurs considerable computational expense when applied to all but the most computationally efficient models. *Latin hypercube design* offers a more efficient, yet still thorough, exploration of parameter space [75], Figure 4C. Representative model behaviour, termed the *response*, is obtained for each sample of parameter space and is then related to each individual parameter, Figure 4D. The *partial rank correlation coefficient* (PRCC) is particularly suitable here, it is sensitive to non-linear trends whilst minimising the effect of confounding co-variates (e.g. correlations between parameters) [76]. It is critical that samples are selected to minimise any correlation between parameters, as these sampling artefacts confound subsequent analysis, Figure 4E. More influential parameters will have larger PRCCs with the response. Many other SA techniques exist, and are extensively reviewed in [74]. The Latin hypercube-PRCC method is

implemented in [43]. Confidence that calibration has delivered an appropriate model is raised if the parameters and associated mechanisms SA highlights as influential are congruent with what is known of the biology.

From a modelling perspective, lack of precise knowledge concerning some aspect of the biology is termed *epistemic uncertainty* [77]. This can concern, for instance, rates, probabilities, population sizes, or any figures pertaining to model parameters. Knowledge, or otherwise, of plausible ranges for parameter values derived from the biology can help contextualise the scientific significance of *in silico* experimentation. This can be done in two ways. First, whilst calibration can assign appropriate parameter values, those values resembling what is well-established biologically will raise trust in the model. This is most relevant for highly influential parameters. Conversely, if the model is highly sensitive to parameters about which very little is known biologically, caution when interpreting results is advisable. Second, one might moderate the interpreted scientific significance of an intervention if the resultant changes in model behaviour lie well within the range of behaviours possible under epistemic uncertainty. This can be gauged through an *uncertainty analysis* [74,78], which quantifies the diversity of model behaviours possible within the ranges of parameter values that current biological knowledge supports. Such an analysis can be performed in a *one at a time* fashion. Alternatively, building a factorial design or Latin hypercube around biologically-supported ranges of parameter values equates to a *global* UA, Figure 5A & B. An *in silico* intervention can be considered scientifically significant if it yields behavioural changes far exceeding that explainable under biological uncertainty, Figure 5C. For ease of illustration this discussion assumes discrete boundaries of "biologically plausible values." However, in many cases biologically uncertainty may be

better expressed in terms of probability distributions, and probabilistic uncertainty analyses do exist [79].

Conclusion

As the old adage goes, “all models are wrong but some are useful” [80]. No model will ever perfectly reflect a complex biological system under all circumstances. Given sufficient effort, calibration will always reveal inconsistencies between model and reality. The importance of a model accurately reflecting a biological system under given scenarios is problem-specific. We would demand greater fidelity of a model directing clinical practice than one employed in notional, preliminary exploration [81]. The cost of capturing finer biological nuance is model complexity, which poses conceptual and calibration challenges; consider another adage, “everything should be made as simple as possible, but not simpler.” Accordingly, modellers should document their calibration activities and the justifications therefor. This can take the form of evidencing in MIRIAM [82], or comprising part of a structured argument that the model is fit for its designated purpose [83].

Rigorous calibration can instil confidence that a model appropriately captures a biological system, or, if unsuccessful, motivate further model development. The strategies we have explored here aim to increase the rigour of calibration. In curve fitting, the further away one extrapolates from known data points, the less accurate predictions will be. The same principle applies to biological models, and the scenarios their calibration covered (Figure 3). Being conceptual, this space of scenarios has no objective notion of distance. However,

selecting calibration scenarios that exercise the full dynamic range of model components and their possible interactions should maximise the range of scenarios under which the model proves accurate.

The ultimate validation for a model is arguably that an *in silico* prediction be experimentally verified. For instance, Pappalardo *et al.*'s prediction that beta sitosterol, a citrus-derived compound, would serve as a powerful adjuvant for influenza A virosome vaccindation a was subsequently verified experimentally in mice [84]. A goal of calibration is to supply evidence and confidence of a model's quality, which can help make the case for investment in such end-stage experimental validations.

Supporting information

S1 Fig Methods.

Methodological details for Figure 1.

Acknowledgements

MNR's Fellowship is supported by the David and Judith Coffey LifeLab: a philanthropic donation made to the University of Sydney. The funders had no part in directing any of the present work, or the preparation of this manuscript.

References

1. Ross J, Arkin AP. Complex Systems: From chemistry to systems biology. *Proc Natl Acad Sci* 2009;**106**:6433–4.

2. Cohen IR, Harel D. Explaining a complex living system: dynamics, multi-scaling and emergence. *J R Soc Interface* 2007;**4**:175–82.

3. Germain RN, Meier-Schellersheim M, Nita-Lazar A *et al.* Systems biology in immunology: a computational modeling perspective. *Annu Rev Immunol* 2011;**29**:527–85.

4. Chen Y, Lawless C, Gillespie CS *et al.* Cali bayes and BASIS: integrated tools for the calibration, simulation and storage of biological simulation models. *Brief Bioinform* 2010;**11**:278–89.

5. Cohen IR. Modeling immune behavior for experimentalists. *Immunol Rev* 2007;**216**:232–6.

6. Ahn AC, Tewari M, Poon CS *et al.* The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Med* 2006;**3**:0709–13.

7. Andrews PS, Stepney S, Timmis J. Simulation as a scientific instrument. *Proc 2012 Work Complex Syst Model Simul* 2012:1–10.

8. Williams RA, Greaves R, Read M *et al.* In silico investigation into dendritic cell regulation of CD8Treg mediated killing of Th1 cells in murine experimental autoimmune encephalomyelitis. *BMC Bioinformatics* 2013;**14 Suppl 6**:S9.

9. Hoekstra A, Chopard B, Coveney P. Multiscale modelling and simulation: a position paper. *Philos Trans R Soc A Math Phys Eng Sci* 2014;**372**:20130377–20130377.

10. Karabasov S, Nerukh D, Hoekstra A *et al.* Multiscale modelling : approaches and

challenges. *Philos Trans R Soc A* 2014;**372**:20130390.

11. Forrest S, Beauchemin C. Computer immunology. *Immunol Rev* 2007;**216**:176–97.

12. Cosgrove J, Butler J, Alden K *et al.* Agent-based modeling in systems pharmacology. *CPT Pharmacometrics Syst Pharmacol* 2015;**4**:615–29.

13. Kim K, Gaudet S, Kim KA *et al.* Systematic calibration of a cell signaling network model Systematic calibration of a cell signaling network model. *BMC Bioinformatics* 2010;**11**:202.

14. Newland CP, Maier HR, Zecchin AC *et al.* Multi-objective optimisation framework for calibration of Cellular Automata land-use models. *Environ Model Softw* 2018;**100**:175–200.

15. Read M, Andrews PS, Timmis J *et al.* Techniques for Grounding Agent-Based Simulations in the Real Domain: a case study in Experimental Autoimmune Encephalomyelitis. *Math Comput Model Dyn Syst* 2012;**18**:67–86.

16. Ruano MV, Ribes J, De Pauw DJW *et al.* Parameter subset selection for the dynamic calibration of activated sludge models (ASMs): Experience versus systems analysis. *Water Sci Technol* 2007;**56**:107–15.

17. Calvez B, Hutzler G. Automatic Tuning of Agent-Based Models Using Genetic Algorithms. In: Sichman JS, Antunes L (eds.). *Multi-Agent-Based Simulation VI*. Springer Berlin Heidelberg, 2006, 41–57.

18. Fabretti A. On the problem of calibrating an agent based model for financial markets. *J Econ Interact Coord* 2012;**8**:277–93.

19. Stockdale JE, Kypraios T, O'Neill PD. Modelling and Bayesian analysis of the Abakaliki smallpox data. *Epidemics* 2017;**19**:13–23.

20. Lillacci G, Khammash M. Parameter estimation and model selection in computational biology. *PLoS Comput Biol* 2010;**6**, DOI: 10.1371/journal.pcbi.1000696.

21. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical*

Information-Theoretic Approach (2nd Ed). Springer-Verlag New York, 2002.

22. Motulsky HCA. Fitting Models to Biological Data using Linear and Nonlinear Regression. *Prism Man* 2003:1–351.

23. Efroni S, Harel D, Cohen IR. Toward rigorous comprehension of biological complexity: modeling, execution, and visualization of thymic T-cell maturation. *Genome Res* 2003;**13**:2485–97.

24. Read M, Andrews PS, Timmis J *et al.* Determining disease intervention strategies using spatially resolved simulations. *PLoS One* 2013;**8**:e80506.

25. Rapin N, Lund O, Castiglione F. Immune system simulation online. *Bioinformatics* 2011;**27**:2013–4.

26. Golubovskaya V, Wu L. Different subsets of T cells, memory, effector functions, and CAR-T immunotherapy. *Cancers (Basel)* 2016;**8**, DOI: 10.3390/cancers8030036.

27. Kaphingst KA, Persky S, Lachance C. The misleading nature of in vitro and ex-vivo findings in studying the impact of stress hormones on NK cell cytotoxicity. *Brain, Behav Immun* 2015;**45**:277–86.

28. Otava M, Shkedy Z, Talloen W *et al.* Identification of in vitro and in vivo disconnects using transcriptomic data. *BMC Genomics* 2015;**16**:1–10.

29. Wu P-H, Giri A, Sun SX *et al.* Three-dimensional cell migration does not follow a random walk. *Proc Natl Acad Sci* 2014;**111**:3949–54.

30. Parks BW, Sallam T, Mehrabian M *et al.* Genetic architecture of insulin resistance in the mouse. *Cell Metab* 2015;**21**:334–46.

31. Mitchell SJ, Madrigal-Matute J, Scheibye-Knudsen M *et al.* Effects of Sex, Strain, and Energy Intake on Hallmarks of Aging in Mice. *Cell Metab* 2016;**23**:1093–112.

32. Liao CY, Rikke BA, Johnson TE *et al.* Genetic variation in the murine lifespan response to

dietary restriction: From life extension to life shortening. *Aging Cell* 2010;**9**:92–5.

33. Swerdlin N, Cohen IR, Harel D. The lymph node B cell immune response: Dynamic analysis in-silico. *Proc Ieee* 2008;**96**:1421–43.

34. Gullo F, Van Der Garde M, Russo G *et al.* Computational modeling of the expansion of human cord blood CD133+ hematopoietic stem/progenitor cells with different cytokine combinations. *Bioinformatics* 2015;**31**:2514–22.

35. Marino S, Myers A, Flynn JL *et al.* TNF and IL-10 are major factors in modulation of the phagocytic cell environment in lung and lymph node in tuberculosis: a next-generation two-compartmental model. *J Theor Biol* 2010;**265**:586–98.

36. King NJC, Getts DR, Getts MT *et al.* Immunopathology of flavivirus infections. *Immunol Cell Biol* 2007;**85**:33–42.

37. Senior AM, Gosby AK, Lu J *et al.* Meta-analysis of variance: An illustration comparing the effects of two dietary interventions on variability in weight. *Evol Med Public Heal* 2016;**2016**:244–55.

38. Senior AM, Nakagawa S, Raubenheimer D *et al.* Dietary restriction increases variability in longevity. *Biol Lett* 2017;**13**:20170057.

39. Senior AM, Nakagawa S, Lihoreau M *et al.* An Overlooked Consequence of Dietary Mixing: A Varied Diet Reduces Interindividual Variance in Fitness. *Am Nat* 2015;**186**:649–59.

40. Read MN, Bailey J, Timmis J *et al.* Leukocyte Motility Models Assessed through Simulation and Multi-objective Optimization-Based Model Selection. *PLOS Comput Biol* 2016;**12**:e1005082.

41. Kirk RE. Effect magnitude: A different focus. *J Stat Plan Inference* 2007;**137**:1634–46.

42. Vargha A, Delaney HD. A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. *J Educ Behav Stat* 2000;**25**:101–32.

43. Alden K, Read M, Timmis J *et al.* Spartan: A Comprehensive Tool for Understanding Uncertainty in Simulations of Biological Systems. *PLoS Comput Biol* 2013;**9**:e1002916.

44. Kitano H. Biological robustness. *Nat Rev Genet* 2004;**5**:826–37.

45. Cohen IR. *Tending Adam's Garden: Evolving the Cognitive Immune Self*. Elsevier Academic Press, 2004.

46. Beltman JB, Maree AF, de Boer RJ. Analysing immune cell migration. *Nat Rev Immunol* 2009;**9**:789–98.

47. Beltman JB, Maree AFM, Lynch JN *et al.* Lymph node topology dictates T cell migration behavior. *J Exp Med* 2007;**204**:771–80.

48. Rubinstein M, Colby R. *Polymer Physics*. Oxford University Press, New York, 2003.

49. Mostafaie A, Forootan E, Safari A *et al.* Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data. *Comput Geosci* 2018:1–26.

50. Chiandussi G, Codegone M, Ferrero S *et al.* Comparison of multi-objective optimization methodologies for engineering applications. *Comput Math with Appl* 2012;**63**:912–42.

51. Read MN, Alden K, Rose LM *et al.* Automated multi-objective calibration of biological agent-based simulations. *J R Soc Interface* 2016;**13**:20160543.

52. Efstratiadis A, Koutsoyiannis D. One decade of multi-objective calibration approaches in hydrological modelling: a review Andreas. *Hydrol Sci J* 2010;**55**:58–78.

53. Handl J, Kell DB, Knowles J. Multiobjective optimization in bioinformatics and computational biology. *IEEE/ACM Trans Comput Biol Bioinform* 2007;**4**:279–92.

54. Deb K, Pratap A, Agarwal S *et al.* A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans Evol Comput* 2002;**6**:182–97.

55. Seada H, Deb K. A Unified Evolutionary Optimization Procedure for Single, Multiple, and

Many Objectives. *IEEE Trans Evol Comput* 2016;**20**:358–69.

56. Cui Y, Geng Z, Zhu Q *et al.* Review: Multi-objective optimization methods and application in energy saving. *Energy* 2017;**125**:681–704.

57. Deb K, Jain H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach , Part I : Solving Problems With Box Constraints. *IEEE Trans Evol Comput* 2014;**18**:577–601.

58. Beauchemin C, Samuel J, Tuszynski J. A simple cellular automaton model for influenza A viral infections. *J Theor Biol* 2005;**232**:223–34.

59. Warrender C, Forrest S, Koster F. Modeling intercellular interactions in early Mycobacterium infection. *Bull Math Biol* 2006;**68**:2233–61.

60. Palumbo MC, Morettini M, Tieri P *et al.* Personalizing physical exercise in a computational model of fuel homeostasis. *PLoS Comput Biol* 2018;**14**:1–23.

61. Bloch N, Harel D. The tumor as an organ: Comprehensive spatial and temporal modeling of the tumor and its microenvironment. *BMC Bioinformatics* 2016;**17**:1–15.

62. Kamal MA, Gieschke R, Lemenuel-Diot A *et al.* A Drug-disease model describing the effect of oseltamivir neuraminidase inhibition on influenza virus progression. *Antimicrob Agents Chemother* 2015;**59**:5388–95.

63. Butler JA, Alden K, Veiga-Fernandes H *et al.* Novel Approaches to the Visualization and Quantification of Biological Simulations by Emulating Experimental Techniques. *ALIFE 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*. New York: MIT Press, 2014, 614–21.

64. Bloch N, Weiss G, Szekely S *et al.* An interactive tool for animating biology, and its use in spatial and temporal modeling of a cancerous tumor and its microenvironment. *PLoS One* 2015;**10**:1–11.

65. Efroni S, Harel D, Cohen IR. Emergent dynamics of thymocyte development and lineage determination. *PLoS Comput Biol* 2007;**3**:127–36.

66. De Vita M, Catzola V, Buzzonetti A *et al.* Unexpected interference in cell surface staining by monoclonal antibodies to unrelated antigens. *Cytom Part B - Clin Cytom* 2015;**88**:352–4.

67. Kumar V, Stellrecht K, Sercarz E. Inactivation of T cell receptor peptide-specific CD4 regulatory T cells induces chronic experimental autoimmune encephalomyelitis (EAE). *J Exp Med* 1996;**184**:1609–17.

68. Monach PA, Mathis D, Benoist C. The K/BxN arthritis model. *Curr Protoc Immunol* 2008;**81**:15.22.1-15.22.12.

69. Cilfone NA, Ford CB, Marino S *et al.* Computational modeling predicts IL-10 control of lesion sterilization by balancing early host immunity-mediated antimicrobial responses with caseation during mycobacterium tuberculosis infection. *J Immunol* 2015;**194**:664–77.

70. Warsinske HC, Pienaar E, Linderman JJ *et al.* Deletion of TGF-β1 increases bacterial clearance by cytotoxic t cells in a tuberculosis granuloma model. *Front Immunol* 2017;**8**, DOI: 10.3389/fimmu.2017.01843.

71. Read MN. Statistical and Modelling Techniques to Build Confidence in the Investigation of Immunology through Agent-Based Simulation. PhD Thesis, The University of York. 2012.

72. Zhu A, Guo J, Ni B-J *et al.* A Novel Protocol for Model Calibration in Biological Wastewater Treatment. *Sci Rep* 2015;**5**:8493.

73. Zi Z. Sensitivity analysis approaches applied to systems biology models. *IET Syst Biol* 2011;**5**:336–46.

74. Saltelli, A., Ratto, M., Andres, T., Campolongo, F. C, J., Gatelli, D., Saisana, M., Tarantola S. *Global Sensitivity Analysis. The Primer*. Wiley, 2008.

75. McKay MD, Beckman RJ, Canover WJ. A Comparison of Three Methods for Selecting

Values of Input Variables in the Analysis of Output From a A Computer Code. *Technometrics* 1979;**21**:239–45.

76. Marino S, Hogue IB, Ray CJ *et al.* A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J Theor Biol* 2008;**254**:178–96.

77. Helton JC. Uncertainty and sensitivity analysis for models of complex systems. In: Barth TJ, Griebel M, Keyes DE, et al. (eds.). *Computational Methods in Transport: Verification and Validation*. Springer Berlin Heidelberg, 2008, 207–-228.

78. Helton JC, Johnson JD, Oberkampf WL *et al.* Representation of analysis results involving aleatory and epistemic uncertainty. *Int J Gen Syst* 2010;**39**:605–46.

79. Coleman HW, Steele WG. *Experimentation, Validation and Uncertainty Analysis for Engineers*. Third. John Wiley & Sons, Inc., 2009.

80. Box G. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN (eds.). *Robustness in Statistics*. Academic Press, 1979, 201--236.

81. Mangion K, Gao H, Husmeier D *et al.* Advances in computational modelling for personalised medicine after myocardial infarction. *Heart* 2018;**104**:550–7.

82. Juty N, Le Novere N, Laibe C. Identifiers.org and MIRIAM Registry: Community resources to provide persistent identification. *Nucleic Acids Res* 2012;**40**:580–6.

83. Alden K, Andrews PS, Polack FAC *et al.* Using argument notation to engineer biological simulations with increased confidence. *J R Soc Interface* 2015;**12**:20141059.

84. Pappalardo F, Fichera E, Paparone N *et al.* A computational model to predict the immune system activation by citrus-derived vaccine adjuvants. *Bioinformatics* 2016;**32**:2672–80.

Key Points:

- The value of biological models hinges on either knowing they are accurate representations of the biology, or establishing otherwise; this is ascertained through *calibration.*
- Stochastic models calibrated against the *distribution* of biological outcomes can help ascertain how these diverging outcomes arise from system components and they dynamics.
- Even where all model parameter values are experimentally ascertained, calibration is still warranted to account for the simplifications and abstractions that models make.
- Single metrics alone may be non-distinguishing for complex systems, resulting in multiple solutions to the calibration problem. Multi-metric calibration can overcome this and make explicit trade-offs in accurate biological capture across metrics.
- Model calibration is ideally performed against multiple experimental conditions or contexts. To do otherwise is akin to curve fitting against a single data point.

Biographical Notes:

Mark Read is a Research Fellow at the University of Sydney. He studies biological systems through computational, mathematical and statistical methods. He focuses primarily on the immune system and gut microbiome.

Kieran Alden is a Research Fellow in Intelligent and Adaptive Systems and the University of York, specialising in the engineering processes through which models help us understand biological systems.

Jon Timmis is Professor of Intelligent and Adaptive Systems at the University of York, with a wide interest in simulation of complex systems. His work spans immunology and swarm robotics.

Paul S. Andrews is a computer scientist and systems engineer, with experience in working with interdisciplinary teams to model and simulate immune system function.

Figures and Captions:

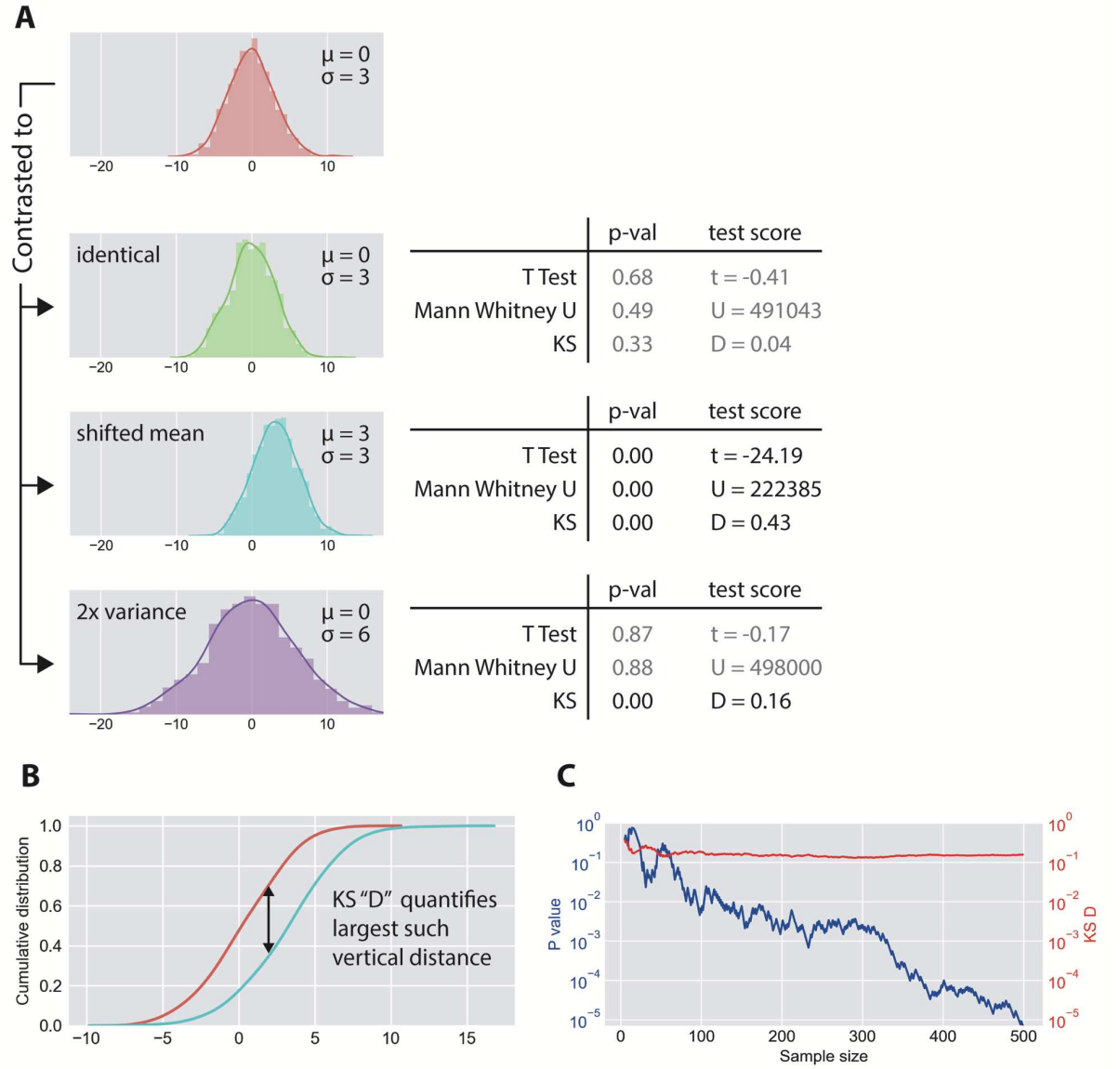

Figure 1: The Kolgmogorov-Smirnov (KS) "D" statistic permits model calibration against distributions of biological outcomes. Such calibration requires sensitive detection of discrepancies between two distributions: that of the target biological behaviour and that of the model. (**A**) An evaluation of the sensitivity of two commonly used statistics, the Mann Whitney U (also called the Wilcoxon rank-sum test) and the T test, and the KS statistic. All three fail to determine statistically significant differences (p-values) between statistically identical Gaussian distributions, as would be expected, and all three are sensitive to changes in distribution mean values. However, only the KS test is sensitive to changes in variance in distributions with identical mean values. We advocate for calibrating

stochastic models with the KS “D” value, which quantifies the largest difference between the proportions of two distributions occurring at or below a given value. This is intuitively visualised (**B**) as the vertical arrow on a cumulative distribution plot. (**C**) The KS “D” is an effect size measure, and is thus relatively insensitive to the number of samples in the distributions being contrasted. Conversely, statistical significance (p values) quantify the probability that the given difference between two distributions could occur through random chance. Given two non-identical distributions, statistical significance is always obtainable with a sufficient sample size. Non-significant p-values are shown in light grey. We provide methodological details for this figure in the supplementary materials.

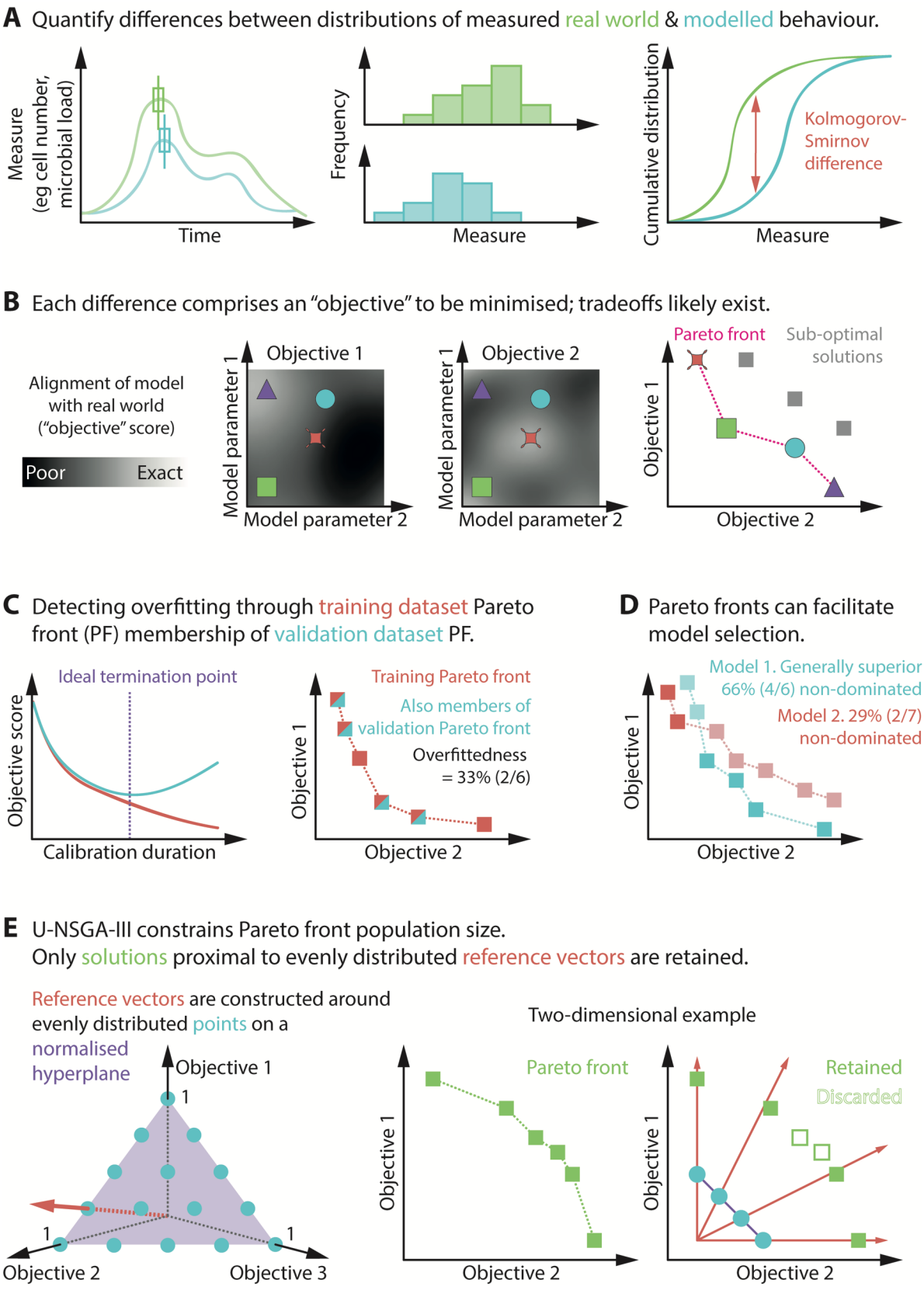

Figure 2: Model calibration and selection using multi-objective optimisation (MOO). (**A**) The divergence between modelled and real biological system behaviours must be quantified. MOO-

calibration can accommodate multiple such quantifications using different measures of system behaviour (only one shown). The example shows how distributions of modelled and real-world behaviour can be aligned using the Kolmogorov-Smirnov (KS) statistic. Cumulative distribution plots are related to histograms, they show the proportion of data in a distribution less than or equal to a given value. Other forms of quantification, such as mean squared error or alternative statistics, may be substituted if more appropriate. (**B**) Each quantification of the difference between modelled and real-world behaviour comprises an *objective* for MOO to minimise by exploring model parameter values. It is highly unlikely that that any single set of parameter values perfectly satisfies all objectives; rather, tradeoffs in performance are typical. A *Pareto front* comprises the subset of solutions for which an improvement in any one objective necessitates a worsening in another; the remaining solutions are sub-optimal. For clarity, only two objectives and model parameters are depicted; many more of both are possible. (**C**) Imperative in model fitting is the detection of overfitting, wherein models capture not only general trends but the specific noise nuances of the training dataset, therein reducing the generality of the model. Overfitting is typically detected through an independent validation dataset not used in model training. Model capture of general biological properties will lead to closer alignment with both training and validation datasets. Conversely, when overfitting occurs model alignment with the training dataset will continue to improve, but alignment with the validation dataset will worsen. This overfitting detection strategy can be ported to the multi-objective context by maintaining Pareto fronts with respect to both training and validation datasets. Training dataset Pareto front member solutions that *are not* also members of the validation dataset Pareto front are indicative of overfitting, and calibration should be terminated when a predetermined portion of the training dataset Pareto front satisfies this criteria. (**D**) The Pareto front concept can facilitate model selection. The superior model will better align with real world behaviours, manifesting as smaller tradeoffs in objective scores. This model's Pareto front will dominate those of inferior alternatives. If modellers have reason to preference particular metrics/objectives, they can focus on sub-regions of the models' Pareto fronts. (**E**) The

number of solutions comprising the Pareto front typically grows exponentially with the number of objectives used; “Universal Non-dominated Sorting Genetic Algorithm” (U-NSGA-III) constrains the Pareto front whilst encouraging a diversity of solutions. This is accomplished by preferentially retaining Pareto front solutions closest to an evenly distributed set of *reference vectors* in objective space. These vectors intersect each of the evenly distributed points on a unit hyperplane and it’s origin. The hyperplane is normalised using the maximum and minimum objective scores across the Pareto front. For more information on MOO and its use in calibration, see references [40,51,55,57].

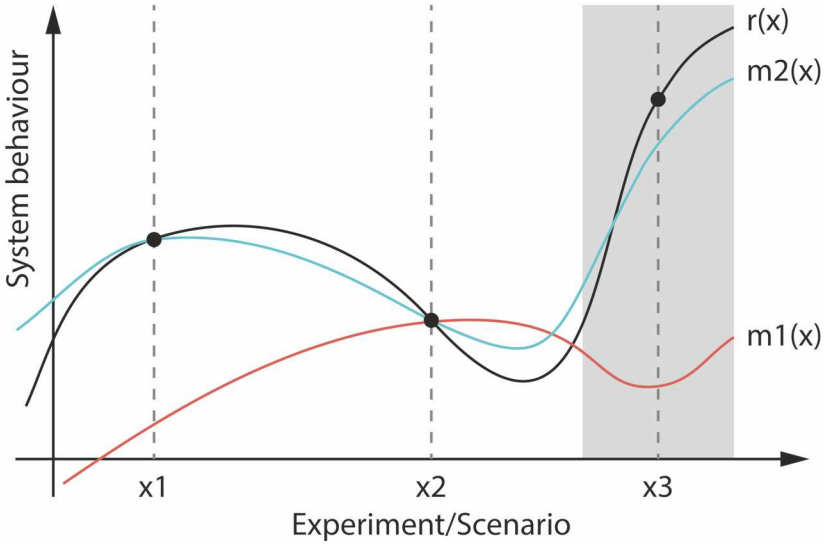

Figure 3: Fitting models against biological behaviour under multiple experiments, conditions or environments (e.g. x1, x2) affords greater confidence that the model will accurately predict biological system performance (r(x)) in scenarios for which this is not known (grey box, x3). Almost any model can be fitted against a single data point. However, unrepresentative models (m1(x)) will fail to reproduce system behaviour across multiple scenarios without heavy alterations to parameter values or mechanics to account for their inadequacy. Employing multiple experiments or scenarios in calibration can highlight inadequate models.

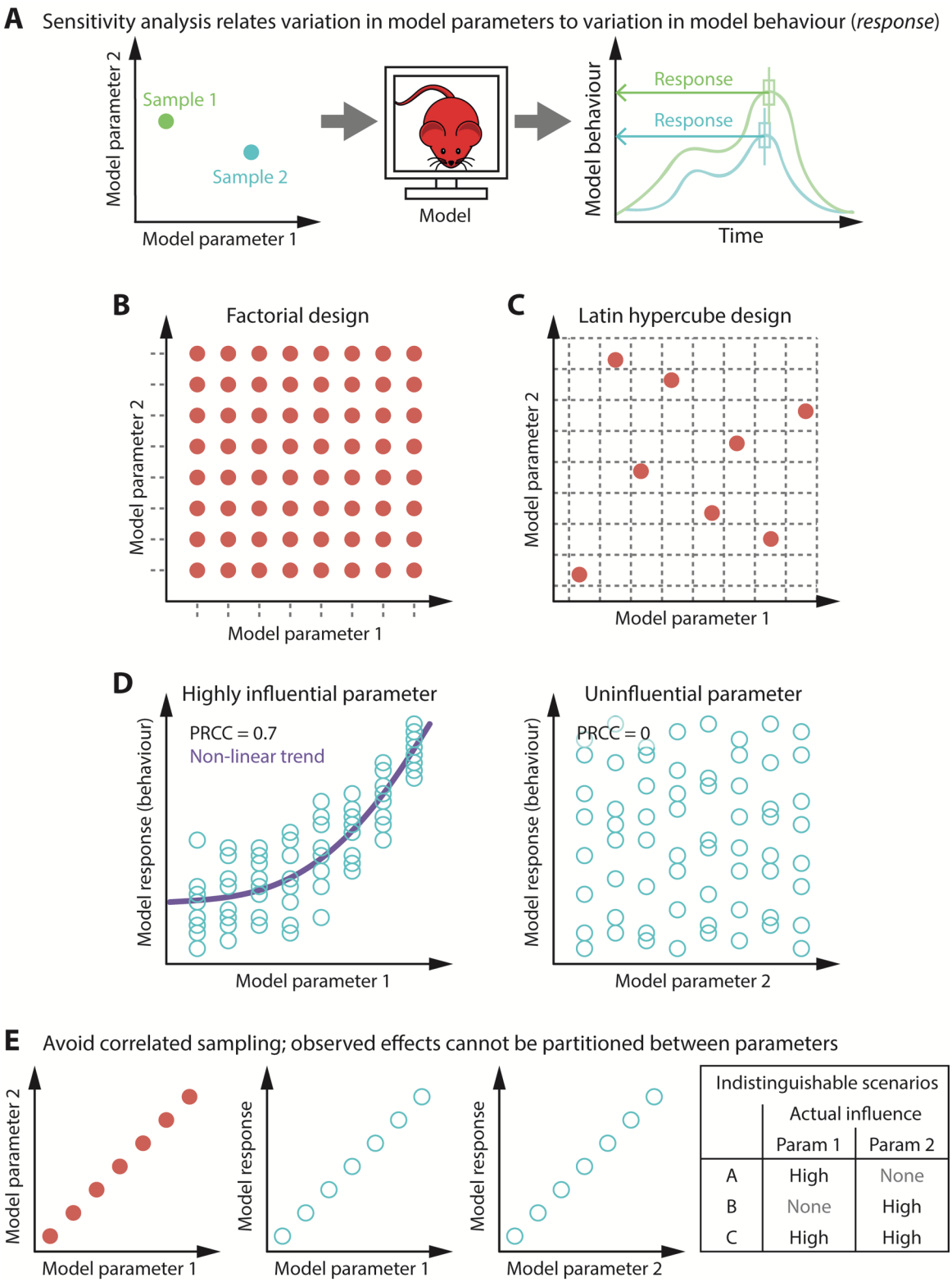

Figure 4: (**A**) Sensitivity analysis quantifies how variation of a model's output relates to variation of its inputs (parameters). It commences with a systematic sampling of model parameter space. (**B**) Given a set of values to explore for each parameter, *factorial design* samples every possible combination thereof. Obtaining representative model behaviours for each sample, particularly for

stochastic models where repeat executions are required, can incur considerable computation expense under factorial design. (**C**) Latin hypercube design represents a more efficient sampling of space, whilst still encouraging an extensive coverage. The range of values of interest for each parameter are segregated into intervals, and each is sampled once. Note that intervals need not be evenly spaced, and can instead concentrate samples around areas of particular interest. (**D**) Parameter influences on the response can be quantified through the partial rank correlation coefficient, which is sensitive to non-linear effects. (**E**) It is essential to minimise correlations between parameters when sampling (e.g. Latin hypercube design), these artefacts confound the partitioning of observed effects between parameters that are correlated. In this example, three distinct possibilities are non-distinguishable because parameters 1 and 2 are sampled in a correlated manner: both parameters are equally influential on model output, parameter 1 is influential and parameter 2 has no effect, or vice versa. This figure depicts only two model parameters for clarity.

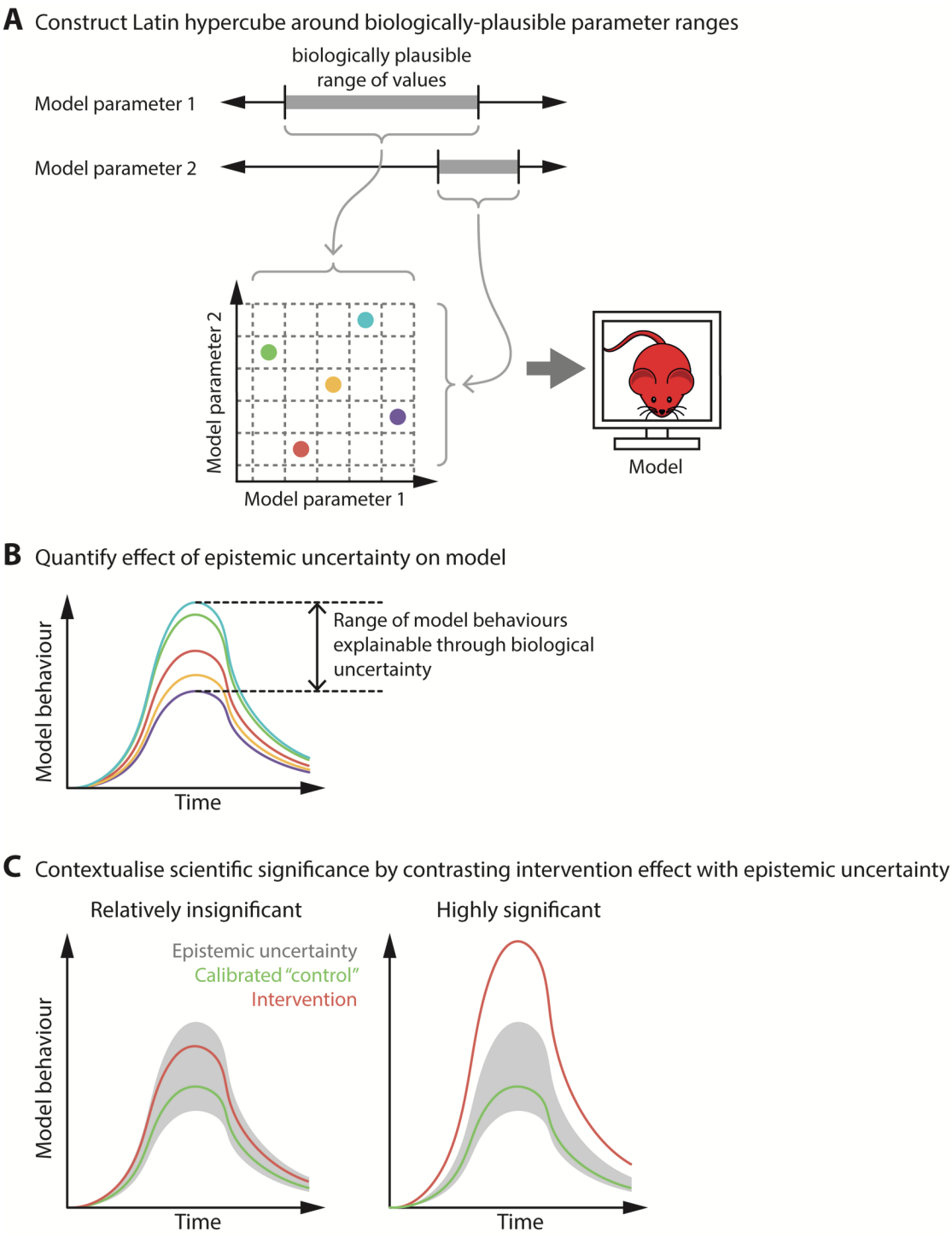

Figure 5: Contextualising *in silico* experimental scientific significance by contrasting effect size against range of model behaviours possible given biological uncertainty (*epistemic uncertainty*). (**A**) Sample model parameter space within region of biologically plausible parameter values. Shown is a Latin hypercube, but other sampling schemes can be substituted. Gather representative model behaviours within this region, and (**B**) quantify range of model behaviours explainable given current

epistemic uncertainty. (**C**) Interventions resulting in model behaviour changes lying within the range explained by epistemic uncertainty have relatively little scientific significance. Conversely, interventions delivering effects far exceeding this range may be highly significant.