Multi-Model and Network Inference Based on Ensemble Estimates: Avoiding the Madness of Crowds

Michael P.H. Stumpf[1,2]

1 School of BioSciences and School of Mathematics and Statistics, University of Melbourne, Parkville VIC 3010, Australia
2 Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK
mstumpf@unimelb.edu.au

Abstract

The development of mathematical models of biological systems has largely relied on a mix of biological intuition, mathematical expediency, and comparisons with data. Models are hard to develop and hard to validate. Recent progress in theoretical systems biology, applied mathematics and computational statistics has, however, led to a proliferation of models and methods that allow us to compare quantitatively the performance of different candidate models at describing a particular biological system or process. Model selection has been applied with great success to problems where a small number — typically less than 10 — of models are compared, but recently studies have started to consider thousands and even millions of candidate models. As formal computational and automated systems increasingly aim to replace or augment biological intuition this trend is likely to continue. Often no single model is chosen as the best, but a large number of models is found to be compatible with data, and these are then often studied as an ensemble. In other situations several prediction algorithms are used to make ensemble predictions, for example, about the presence or absence of molecular interactions. Using a set of test-cases I outline when such ensemble estimators can be relied upon, and when they are likely to fail. If model or predictor ensembles are not carefully constructed and curated, predictions and estimates based on ensembles can perform worse than good individual predictors.

1 Introduction

Legend has it that Leonardo de Pisa — otherwise known by the name of *Fibonacci* — laid the foundations of mathematical biology (Schroeder, 2008). This would mean a 200 year head start compared to theoretical physics, the beginnings of which can be loosely traced to his compatriot, Gallileo Gallilei (Segrè, 1984). While such a statement would be an arguably incorrect historical account of the roots of theoretical biology and theoretical physics, there is widespread agreement that, while physics has become far more mathematical than biology, quantitative and mathematical approaches are becoming ever more popular in the life- and biomedical sciences(May, 2004). Following pioneering work at the start of the 20th Century(Michaelis & Menten, 1913), the second half in particular saw tremendous success in e.g. mathematical developments — often intimately tied to statistical analyses — in the context of evolutionary and population biology, which have then branched out. Bioinformatics, and then systems biology have become important tools supporting research across the life sciences in the 21st century, and increasingly *in silico* biology complements the traditional *in vivo* and *in vitro* lines of investigation.

In physics simple and elegant symmetry relationships have often led the way to theoretical models(Thorne & Blandford, 2017). Most importantly Emmy Noether's theorem has been pivotal in establishing the correspondence between continuous symmetries and conservation lows in physics(Neuenschwander, 2017); it together with suitable generalisations to quantum and quantum field theory has been pivotal in the derivation of physical *laws of nature*. Biology has not been able to draw on such fundamental principles, to a large

degree because most processes are intrinsically dissipative (meaning energy is produced and consumed) and hence the conditions where Noether's theorem holds simply do not apply. Instead biological models have often had a heuristic element, or described the rates of change in (often macroscopic) observables (e.g. species of plans or animals, or molecular species).

Mathematical models are more precise than verbal models — which have permeated biological literature for a long time — because they require us to state our assumptions explicitly and at the outset. Thus a statement of the type "*X increases over time*" could correspond to mathematical models of the type

$$
\begin{aligned}
\frac{dx}{dt} &= \alpha \\
\frac{dx}{dt} &= \alpha x \\
\frac{dx}{dt} &= \alpha x^2 + \beta x \\
\frac{dx}{dt} &= \frac{\alpha x}{K + x} \\
\vdots &= \vdots
\end{aligned}
$$

and many others. Writing down the set of equations, such as here in the form of differential equations, is an important starting point in modelling as it forces us to express our assumptions in a precise form. Which form to choose is, however, not unambiguously obvious. Instead we often rely on data to decide between the different options. *Statistical model selection* (Burnham & Anderson, 2013) provides the tools to make such decisions, balancing the ability of a model to fit the data with the model's complexity. As larger and larger models, even models for whole cells(Karr *et al.*, 2012, 2015; Lang & Stelling, 2016; Babtie & Stumpf, 2017), are being considered model selection problems will presumably become the norm, especially when models are constructed exhaustively or automatically (Ma *et al.*, 2009; Barnes *et al.*, 2011; Szederkényi *et al.*, 2011; Sunnåker *et al.*, 2014; Babtie *et al.*, 2014; Leon *et al.*, 2016; Gerardin *et al.*, 2019), see Fig. 1.

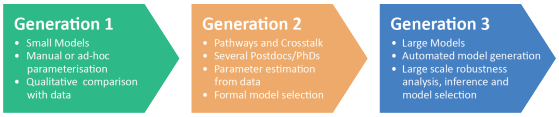

Figure 1: *It can be helpful to divide modelling in biology into three different categories: 1st generation modelling was carried out by individual researchers and considered small systems with parameters determined either from the literature or in an* ad hoc *manner. In the 2nd generation, modelling relied on comparisons with data and involved teams of researchers working on larger systems and networks; we are now entering a realm where 3rd generation models are introduced that may involve automated model generation and sophisticated model selection or robustness analysis.*

Whereas in the past model selection has dealt with a relatively small number of alternative models, biological modelling increasingly considers many models as representations for biological systems(Babtie *et al.*, 2014). In some, probably many, situations model selection will not be able to decide on a single best model. Instead many models may have equally strong support. In such a situation we may then base analysis or predictions on the ensemble of models that are supported by the data. Each model's contribution to the prediction etc. is weighted by the relative support it has received. Such estimates or predictions based on ensembles have been referred to as exploiting the "*wisdom of crowds*" (Marbach *et al.*, 2012). This refers to the notion that groups of individuals/models are more likely to the be better than those based on a single individual/model. This concept, however, also references a much earlier work, Charles Mackay's "*Extraordinary Popular Delusions and the Madness of Crowds*"(Mackay, 1841), a 19th century account of how popular opinion can support quite extraordinary and plainly wrong opinions and concepts.

Below we provide a self-contained overview over model selection and model averaging. We then study in turn: model selection and its susceptibility to poorly defined sets of candidate models; and the behaviour of ensemble approaches to network reconstruction. Network inference has long and rightfully been regarded as a statistically hard problem, and ensemble estimators have become popular to overcome the limitations of individual methods and generate better, more reliable networks from data. Both problems share in common a dependence on the quality of the models that are being used in the ensemble averaging and we map out and quantify these influences; we also provide self-contained `Julia` code for further *in silico* experimentation and analysis by

the reader.

2 Model Selection and Multi-Model Inference

Increasingly we have many competing models to explain any given biological system. Model selection (Burnham & Anderson, 2013; Kirk *et al.*, 2013) provides the framework in which we can compare and contrast the performance of different models in light of available data, balancing, as needed, the model's complexity with its ability to capture the data, $\mathcal{D}$. We assume that we have a *universe* of models, $\mathcal{M}$,

$$\mathcal{M} = \{M_1, M_2, \ldots, M_N\}.$$

For simplicity we assume a finite number of models, N. Furthermore, for each model, M_i, we assume that we have a data generating function, $f_i(\theta_i)$, parameterised by a parameter vector θ_i which is chose from some suitable continuous parameter space,

$$\theta_i \in \Omega_i \subseteq \mathbb{R}^n$$

Coping with the size of the parameter space (Cox, 2006) is one of the essential challenges of model selection (and parameter inference, of course, as well), and a range of methods have been developed to tackle or sidestep this problem.

Our discussion aims for flexibility, but in practice we are opting to stay in a Bayesian framework(Robert, 2007). In the context of parameter estimation we seek to determine the *posterior distribution* over parameters,

$$\Pr(\theta_i|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\theta_i)\pi_i(\theta_i)}{\Pr_i(\mathcal{D})}; \tag{1}$$

here $\Pr(\mathcal{D}|\theta_i)$, is the *likelihood*, defined as the probability of observing the data for a given choice of θ; $\pi_i(\theta_i)$ is *prior* over the parameters for model M_i, and $\Pr_i(\mathcal{D} = \int \Pr(\mathcal{D}|\theta)\pi_i(\theta)d\theta$ (here we make the dependence on the choice of model explicit through an index) is a normalisation constant often referred to as the *evidence*.

The parameter posterior is a probability distribution over the parameter space(Robert, 2007), which takes into account the data, $\mathcal{D}$, and any knowledge from previous experiments or background information via the prior, $\pi_i(\theta_i)$. If we are interested in a *credible interval* (Robert, 2007) for an individual parameter, $\theta^{(k)}$ we obtain this by integrating over the other parameter to obtain a *marginal distribution*

$$\Pr(\theta_i^{(k)}|\mathcal{D}) = \int_{\theta_i^{(1)}} \ldots \int_{\theta_{(k-1)}} \int_{\theta_i^{(k+1)}} \ldots \int_{\theta_i^{(n)}} \Pr(\theta_i|\mathcal{D}) d\theta_i^{(1)} \ldots d\theta_i^{(k-1)} \theta_i^{(k+1)} \ldots \theta_i^{(n)} \tag{2}$$

Typically we seek to determine an interval that contains the parameter (conditional on model, data, and prior knowledge) with greater than probability q, where conventionally we may, for example, choose $q = 0.95$. However, going from global to marginal distributions may distort the true posterior function, as has been well established for example for so-called *sloppy systems*(Gutenkunst *et al.*, 2007), but this has been discussed elsewhere(Secrier *et al.*, 2009; Apgar *et al.*, 2010; Erguler & Stumpf, 2011; Tönsing *et al.*, 2014; Kirk *et al.*, 2015; Dufresne *et al.*, 2018).

In the Bayesian framework model selection is a (relatively) straightforward extension, and the *model posterior* is given by

$$\begin{aligned}\Pr(M_i|\mathcal{D}) &= \frac{\Pr(\mathcal{D}|M_i)\pi(M_i)}{\Pr(\mathcal{D})} \\ &= \frac{\int_{\Omega_i} \Pr(\mathcal{D}|\theta)\pi_i(\theta)d\theta\pi(M_i)}{\Pr(\mathcal{D})},\end{aligned} \tag{3}$$

where analogously to Eqn. (1), we have the *model posterior* and $\Pr(M_i, \mathcal{D})$, *model prior* $\pi(M_i)$. The denominator terms in Eqns. (1) and (3) are notoriously hard to evaluate for all but the simplest cases, and a large amount of ingenuity and work has been invested into computational schemes(Green *et al.*, 2015), including *Markov Chain Monte Carlo*, *Sequential Monte Carlo*, and related approaches. Often even the likelihood is prohibitively expensive to evaluate and so-called *approximate Bayesian computation* (ABC) schemes have been devised to make Bayesian statistical inference possible (Toni & Stumpf, 2010).

There are several alternatives to the Bayesian framework, and likelihood-based inference and optimisation of cost-functions are very popular in the literature (Raue *et al.*, 2014). These methods result in *point estimates* for the parameters, e.g. the value of θ' that maximises the probability of observing the data, i.e. the *maximum likelihood estimate* (Cox, 2006),

$$\hat{\theta}'_L = \text{argmax}_{\theta'} \Pr(\mathcal{D}|\theta') \tag{4}$$

Similarly, we can determine the *maximum aposteriori estimate* by finding the mode of the posterior distribution (Robert, 2007; MacKay, 2003),

$$\hat{\theta}'_B = \text{argmax}_{\theta'} \Pr(\theta'|\mathcal{D}). \tag{5}$$

Clearly, compared to analysis of the posterior distribution, such local estimates loose a lot of relevant information, but some characteristics can still

be recovered by considering the local curvature of the likelihood, i.e. the *Fisher Information*, or cost-function surface around the (local) extremum identified in this manner(Gutenkunst *et al.*, 2007; Erguler & Stumpf, 2011).

As far as model selection is concerned, seminal work by Akaike (Akaike, 1974) and others has led to model selection frameworks that are based on likelihood inference and provide criteria by which the *explanatory power* of a candidate model can be balanced against its *complexity* — principally the number of model parameters — to find an *Ockam's Razor* (MacKay, 2003)t ype most parsimonious model among a set of candidate models. The *Akaike Information criterion* (AIC) (Burnham & Anderson, 2013; Kirk *et al.*, 2013) for model M_i is given by

$$\mathrm{AIC}_i = -2\log(\Pr(\mathcal{D}|\hat{\theta}, M_i)) + 2n_i \tag{6}$$

with $\hat{\theta}$ given by Eqn. (4), and n_i the number of parameters of model M_i. The AIC is probably the most widely used model selection criterion, despite the fact that it is biased in favour of overly complicated models as the amount of available data increases. The *Bayesian information criterion* does not suffer in the same way; it is defined as

$$\mathrm{BIC}_i = -2\log(\Pr(\mathcal{D}|\hat{\theta}, M_i)) + n_i\log(\nu) \tag{7}$$

with ν representing the size of the data or number of samples. Several other information criteria exist (discussed e.g. in (Burnham & Anderson, 2013; Kirk *et al.*, 2013)), but they all share in common the purpose of balancing model complexity with model fit. The BIC can be understood as and derived from an approximation to the full Bayesian model posterior probability, which achieves this balance implicitly.

Model selection, by whichever mechanisms, works if one model is substantially better than its alternatives(Burnham & Anderson, 2013; Kirk *et al.*, 2013). If several models appear to have similar explanatory power, and model selection cannot pick out a clear winner, then either (i) further analysis should be used to design better, more informative experiments that can discriminate among these models(Liepe *et al.*, 2013; Busetto *et al.*, 2013; Silk *et al.*, 2014); or (ii) these models should be considered as an ensemble (Burnham & Anderson, 2013; Thorne & Stumpf, 2012). The former has definite attractions as it will lead to an increase in our understanding if we can discard some of the models/mechanistic hypotheses.

The latter approach, basing analysis and especially predictions on an ensemble of models has become a popular concept in machine learning. Most notably, in the context of biological network inference the "*Wisdom of Crowds*" concept (Marbach *et al.*, 2012) has been important in popularising inference based on several models. Here we are considering model averaging where contributions from different models may be weighted by their relative explanatory performance. In the Bayesian framework(Hoeting *et al.*, 1999; Thorne & Stumpf, 2012), we can use the posterior probability directly. In the context of an information criterion I_i for model i we define (Burnham & Anderson, 2013)

$$\Delta_i = I_i - \mathrm{argmin}_i I_i \tag{8}$$

and then determine the model weight as

$$w_i = \frac{\exp(-\Delta_i)}{\sum_{i=1}^{N}\exp(-\Delta_i)}. \tag{9}$$

The model weights (e.g. the Akaike weight if I_i is the AIC) provide the relative probability of model M_i to be the correct model conditional on the data $\mathcal{D}$ and the set of models, $\mathcal{M}$, considered.

Model averaging in this framework can serve as a *variance reduction technique*(Burnham & Anderson, 2013; Murphy, 2012). In extreme cases where we have a set of models (or predictors) but no recourse to weights based on e.g. information criteria, we may instead choose to use a voting scheme among the set of models. That is, if the majority of models supports the existence of a given feature, e.g. an edge in a network, then the majority vote will decide as to whether such a feature should be represented. We return to such simple schemes in the penultimate section.

3 Statistical Physics of Model Selection and Ensemble Estimation

Below we are interested in investigating cases where the size of the model universe, $\mathcal{M}$, can be very large (Ma *et al.*, 2009; Babtie *et al.*, 2014; Zheng *et al.*, 2016; Scholes *et al.*, 2019; Qiao *et al.*, 2019). We assume that we either have posterior model probabilities, Eqn. (3), or relative model probabilities, Eqn. 9. In order to simplify and unify the discussion we define now a relationship between the model probability (always implicitly understood to be either a posterior or relative model probability), p_i, and a *cost* or *energy*, ϵ_i, as

$$p_i = \frac{\exp(-\beta\epsilon_i)}{Z} \tag{10}$$

with the normalisation constant Z (the *partition function*) given by

$$Z = \sum_{i=1}^{N}\exp(-\beta\epsilon_i)$$

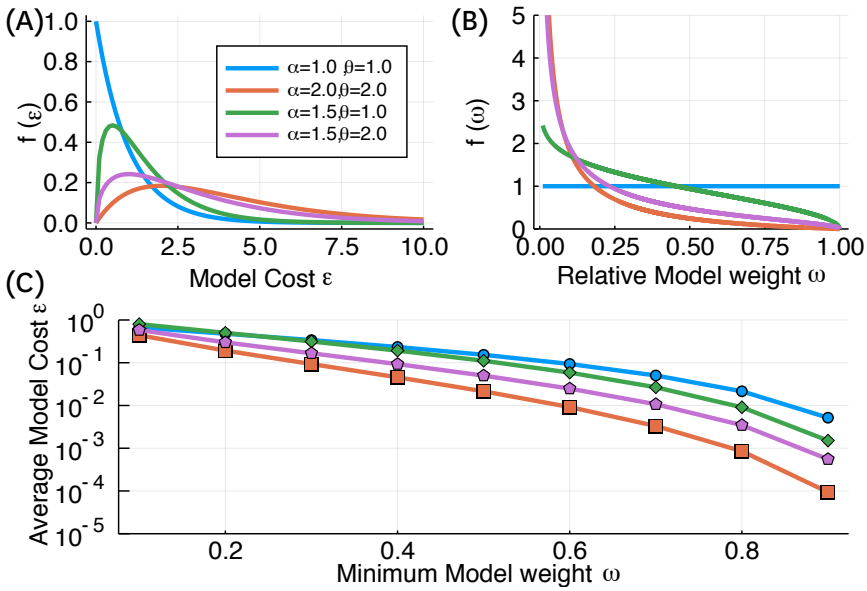

Figure 2: *Different distributions over the model costs, ϵ_i (top left) and the resulting distributions over the relative model weights (top right). Models seen as good (low cost) correspond to those with higher weights, ω. Note that for $\alpha = \theta = 1$ the Gamma distribution reduces to an Exponential distribution, and therefore the distribution over model weights becomes uniform. The bottom panel shows the average cost, ϵ, if models above a given minimum relative weight are included.*

With this in hand we can straightforwardly consider different model selection/averaging frameworks in a similar manner.

In general the true data-generating model (a natural system) will not be represented perfectly by any of the models in $\mathcal{M}$; we denote this true model by $\aleph \notin \mathcal{M}$.

But if we are interested in finding out whether $\aleph$ has a certain characteristic ϕ we would obtain this from the appropriate ensemble average

$$\Pr(\phi \in \aleph) = \sum_{i=1}^{N} p_i \mathbb{1}(\phi \text{ is part of } M_i). \tag{11}$$

where $\mathbb{1}()$ is the conventional indicator function, i.e.

$$\mathbb{1}(x) = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Eqn. (11) is based on three, potentially strong assumption.

1. The model universe, $\mathcal{M}$, is "complete" in the sense that we expect $\mathcal{M}$ to contain any model, M_k, which we might expect to be a reasonable description of $\aleph$ (always remembering that $\aleph \notin \mathcal{M}$.
2. It is decidable if ϕ is part of M_i, $\forall M_i \in \mathcal{M}$.
3. ϕ, or every ϕ of interest has plaid no part, or only a negligible part in the construction of the model probabilities, p_i, Eqn. (10).

The first assumptions is arguably the strongest assumption. In principle we can update $\mathcal{M} \longrightarrow \mathcal{M}'$ by adding additional models in light of new data; specifying new priors, π' for the models $M_i' \in \mathcal{M}'$, will require some care, of course. One important condition that $\mathcal{M}$ (and $\mathcal{M}'$) must fulfil is that models must not appear more than once in the set. This is important to keep in mind as we are increasingly relying on automated generation of models.

With fixed $\mathcal{D}$ and $\mathcal{M}$ Eqn. (11) is, however, a good approach for ensemble-based prediction and estimation. It also encompasses Bayesian model averaging(Hoeting *et al.*, 1999) and multi-model inference based on information criteria (Burnham & Anderson, 2013).

Multi-model inference and model averaging have received considerable positive attention in the systems biology and medicine literature(see e.g. (Choobdar *et al.*, 2019; Kueffner *et al.*, 2019; Camacho *et al.*, 2018; Castro *et al.*, 2019; van der Wijst *et al.*, 2018)), although problems in its applications are frequently overlooked (Cade, 2015). In the context of biological network inference, for example, the DREAM consortium has shown that inference based on community consensus tends to outperform individual algo-

rithms for network inference (Marbach *et al.*, 2012). We next consider two illustrative cases of model averaging (or *ensemble averaging*) in the rest of this section, before turning to network inference in the following section.

The effective model space

Here we first turn our attention to a simple case where we assume that all models in our universe have associated energies, or costs, drawn from a probability distribution, $q(\eta)$ (Babtie & Stumpf, 2017). Furthermore we assume that this is a sufficiently well behaved probability distribution, with positive support and associated density, $f_q(\epsilon)$, over the model energies, ϵ_i, such that

$$\epsilon_i \sim q(\eta). \tag{13}$$

Because ϵ_i is a random variable, the relative weight, $\omega_i = \exp(-\epsilon_i)$, will also be a random variable, and we can obtain the probability density function, $f(\omega)$ via straightforward change of variables as

$$f(\omega) = \frac{q(-\log(\omega))}{|\omega|}. \tag{14}$$

With this in hand we can, for different choices of $q(\epsilon)$, investigate the distribution over model weights. For example if $\epsilon_i \sim \text{Gamma}(\alpha, \theta)$ (where α and θ denote the shape and scale parameters, respectively) then

$$f(\omega) = \frac{-\log(\omega)^{\alpha-1}\omega^{1/\theta}}{\omega\Gamma(\alpha)\theta^{\alpha}}, \tag{15}$$

with $\Gamma(.)$ the Gamma function. The Gamma distribution is a very flexible but generic distribution and was chosen for its generality rather than any particular property and our discussion here does not depend on its specifics. Some representative distributions over ϵ and the corresponding distributions over ω are shown in Fig. 2, panels (A) and (B).

We see that the distribution over model costs, ϵ, affects the distributions over model weights, ω. This is important to realise when deciding on how to triage model sets for further analysis and prediction (Cade, 2015): generally, inference based on all models weighted by ω_i is neither practical nor desirable, as many models with low weight will mask the information available in the good models. If, for example, we only include models with $\omega \in [0.9, 1.0]$ then the average model cost

$$\bar{\epsilon}_{\omega>\omega'} = \int_{\omega'}^{1} -log(\omega)f(\omega)d\omega$$

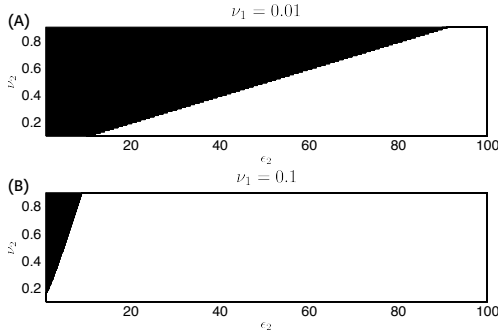

Figure 3: *For $\epsilon_1 = 0.1$ we indicate the areas in the ϵ_2, ν_2 space, where the contribution to ensemble estimators coming from the nuisance models exceeds that coming from the useful models (black). On the top the frequency of useful models is $\nu_1 = 0.01$; in the bottom panel we set $\nu_1 = 0.1$.*

(because $\epsilon = -\log(\omega)$) for these sets will be 0.005 (blue), 0.9×10^{-4} (orange), 0.0015 (green), and 0.0006 (purple).

These results, and more generally those contained in Fig. 2(C), indicate that the — of course unknown — distribution over costs can affect multi-model inference quite profoundly. For model universes that are enriched for good models (i.e. many models M_i with low values of ϵ_i) selecting a subset of models based on even a fairly conservative threshold ω' on the model weights ω_i, can result in large model sets for further prediction.

A simple test case for multi-model inference

Here we study a very simplistic scenario in which we have three types of models, borrowing and adapting Box's (Box, 1976) terminology:

Useful Models which capture important aspects of $\aleph$ and which have an associated cost ϵ_1.

Useless Models which are qualitatively and quantitative poor descriptions of reality and have an associated cost $\epsilon_2 \gg \epsilon_1$.

Nuisance Models which are qualitatively different from reality, but which can quantitatively capture aspects of $\aleph$ by chance; their costs are random variables, $\eta \sim [0, \tau]$.

Nuisance models are here assumed to be models where the quantitative agreement with data is unrelated to the data generating mechanism $\aleph$. Purely machine-learning based models are one way in which we can realise such nuisance models (Camacho *et al.*,

2018); for small datasets $\mathcal{D}$, poorly designed experiments, or simply lack of prior knowledge, there are many ways in which model fit may only be a poor reflection of reality, and this can also give rise to what amounts to nuisance models (Baker *et al.*, 2018).

For concreteness let these three different model classes have sizes ν_1, ν_2 and $\nu_3 = N - \nu_1 - \nu_2$; again in reality we will not these numbers, nor the class membership of any given model *a priori*. For simplicity we consider the case $\eta \sim \mathcal{U}_{[0,\epsilon_2]}$ i.e. nuisance models are at worst as bad as useless models. Then the fraction of nuisance models that have lower associated costs than the *useful models* is given by

$$\frac{\epsilon_1}{\epsilon_2}\nu_3$$

The relative influence of nuisance models can be studied by contrasting three features, ϕ_1, ϕ_2, and ϕ_3, with the following properties

ϕ_1 : equally represented with frequency ξ among models of all classes.

ϕ_2 : only represented among useful models.

ϕ_3 : only represented among "nuisance" models.

With Eqn. (11) we can obtain $\Pr(\phi_i \in \aleph)$ for any property with frequencies ξ_i in the classes $i = 1, 2, 3$,

$$\begin{aligned}\Pr(\phi_1 \in \aleph) &= \frac{\nu_1\xi_1 e^{-\epsilon_1} + \nu_2\xi_2 e^{-\epsilon_2} + \frac{\nu_3\xi_3}{\epsilon_2}\int_0^{\epsilon_2} e^{-\eta} d\eta}{\nu_1 e^{-\epsilon_1} + \nu_2 e^{-\epsilon_2} + \frac{\nu_3}{\epsilon_2}\int_0^{\epsilon_2} e^{-\eta} d\eta} \\ &= \frac{\nu_1\xi_1 e^{-\epsilon_1} + \nu_2\xi_2 e^{-\epsilon_2} + \frac{\nu_3\xi_3}{\epsilon_2}(1 - e^{-\epsilon_2})}{\nu_1 e^{-\epsilon_1} + \nu_2 e^{-\epsilon_2} + \frac{\nu_3}{\epsilon_2}(1 - e^{-\epsilon_2})}\end{aligned} \tag{16}$$

First, for ϕ_1 we trivially obtain

$$\Pr(\phi_i \in \aleph) = \xi. \tag{17}$$

For the more interesting probability for ϕ_2 under the model averaging framework we obtain,

$$\Pr(\phi_2 \in \aleph) = \frac{\nu_1\xi_1 e^{-\epsilon_1}}{\nu_1 e^{-\epsilon_1} + e^{-\epsilon_2}(\nu_2 + \frac{\nu_3}{\epsilon_2}(e^{\epsilon_2} - 1))}, \tag{18}$$

and finally, for a characteristic shared by and confined to the set of nuisance models, we obtain,

$$\Pr(\phi_3 \in \aleph) = \frac{\frac{\nu_3}{\epsilon_2}(1 - e^{-\epsilon_2})}{\nu_1 e^{-\epsilon_1} + e^{-\epsilon_2}(\nu_2 + \frac{\nu_3}{\epsilon_2}(e^{\epsilon_2} - 1))}. \tag{19}$$

From Eqns. (18) and (19) we already see that multimodal averaging is prone to propagate wrong results if nuisance models are sufficiently frequent and some receive good quantitative support (i.e. low model costs, ϵ). Equally, worrying, the same scenario can make it hard for true aspects of $\aleph$ to receive sufficient support through Eqn. (18) if there are many nuisance and useless models included in the data.

To illustrate this further we can consider the case where $\xi_2 = 1$ and $\phi_3 = 1$ (meaning every useful model exhibits characteristic ϕ_2, and every nuisance model characteristic ϕ_3) and ask when is $\Pr(\phi_3 \in \aleph) > \Pr(\phi_2 \in \aleph)$? We obtain

$$\nu_3 > \epsilon_2\nu_1\frac{e^{-\epsilon_1}}{1 - e^{-\epsilon_2}} > \epsilon_2\nu_1 e^{-\epsilon_1}. \tag{20}$$

Thus if useful models are sufficiently rare in the model set (say $\nu_1 < 0.1$) the nuisance models' characteristics will have high weight in the ensemble average, see also Figure 3. Neither of the parameters, $\nu_1, \nu_2, \nu_3, \epsilon_1, \epsilon_2$ are, of course, known, and we cannot know which class a model belongs to *a priori*. Thus model averaging is not a panacea and requires careful consideration of which models are included in the set.

4 Ensemble Estimation and Network Inference

In the preceding section we have looked at generic problems arising in the context of multi-model/ensemble inference in the case of potentially large model spaces. Here we address a related problem: inference of networks using a pool of different network inference methods. In the context of the DREAM initiative (Marbach *et al.*, 2012) basing inference on a set of models was found to be advantageous compared to individual methods; other studies have confirmed this in the setting of protein interaction networks (Thorne & Stumpf, 2012).

Network inference based on ensemble methods is prone to similar pitfalls as discussed above and we illustrate these here. We are looking at networks with V vertices or nodes, and L links or edges. We are considering directed networks (Kolaczyk, 2009; Koller & Friedman, 2009), i.e. every link has a start and an end indicated by an arrow; the *adjacency matrix*, A, is a convenient way to specify such networks, if we indicate the presences and absences of edges, by one and zero, respectively, in its entries,

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

we do not comment on the relative merits or lack thereof of static network descriptions for biological

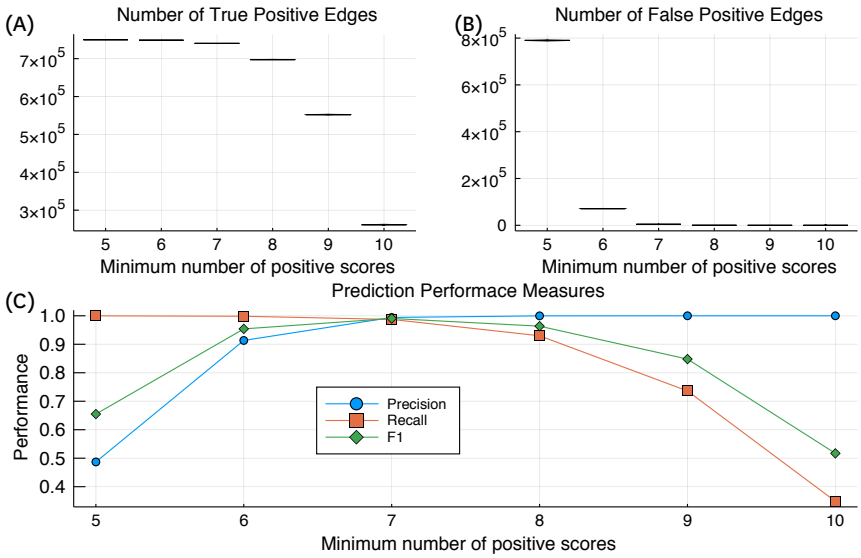

Figure 4: *Illustration of the performance of ensembles of network inference methods with with false positive and false negative probabilities $s = t = 0.1$. Real biological networks are sparse unless false-positives are controlled in the ensemble(Stumpf & Wiuf, 2010; Penfold & Wild, 2011); once false-positives are over-controlled (by demanding a larger number of methods to score an edge), the recall deteriorates.This is reflected in the top panels (the number of true and false positives (we generated 1000 random inferred networks but because of the large network size and the binomial nature of the process the distributions around the mean are tight. The lower panel shows the precision, recall and F1 statistics (Murphy, 2012; Mc Mahon* et al.*, 2014) (see also Appendix A) as a function of the minimum number of methods that need to positively score an edge (again confidence intervals are very tight to be effectively hidden by the symbols.*

systems which need to be considered carefully and separately for each problem being considered (Ingram *et al.*, 2006; Jaeger & Monk, 2015). Instead we use these static networks as a vehicle to study when ensemble predictions will improve on individual predictions and when not. The lessons learned here will translate to more realistic network descriptions as we will argue.

Again we start from some data, $\mathcal{D}$, but now we seek to determine whether the structure of $\mathcal{D}$ suggests functional relationships among pairs of nodes v_i, v_j that may be explained by an edge from i to j, $_{ij}$, or from j to i, $l_j i$. We have access to k different prediction algorithms for the existence of such edges, $O_\kappa \kappa = 1, \ldots, k$, which predict the presence, $A_{ij}^{(\kappa)} = 1$ or absence $A_{ij}^{(\kappa)} = 0$, of a link from i to j etc.. If we have for the false positive and false-negative probabilities for method κ,

$$s_\kappa = \Pr(1|\text{link is not part of the true network})) \tag{21}$$

and

$$t_\kappa = \Pr(0|\text{link is part of the true network}) \tag{22}$$

we can assess how beneficial the ensemble estimators are for the quality of the inferred networks.

Ensembles of Identical Estimators

The simplest case, which is already instructive as a baseline, is where all methods have identical performance, $s_\kappa = s$ and $t_\kappa = t$, $\forall \kappa = 1, \ldots, k$. If the performance of the inference methods is statistically independent, then the number of agreeing inference methods is a binomial random variable. If we set a threshold on the minimum number of concordances among inference methods, k_0, we then have for the overall probability of scoring a true edge from the ensemble,

$$\tilde{\Pr}(1|\text{edge is in } \aleph) = \sum_{\kappa=k_0}^{k} \binom{k}{\kappa} (1-t)^\kappa t^{k-\kappa}, \tag{23}$$

while the probability of a false negative is

$$\tilde{\Pr}(1|\text{edge is not in } \aleph) = \sum_{\kappa=k_0}^{k} \binom{k}{\kappa} s^\kappa (1-s)^{k-\kappa}, \tag{24}$$

To illustrate the outcome of such a simple estimation procedure we assume a network loosely based on expected *Homo sapiens* network sizes(Stumpf *et al.*, 2008) ($22,000$ nodes and $750,000$ interactions). In figure 4 we consider 10 ensembles and base the ensemble estimator on the minimum number, k_0, of methods that have to predict an edge. If

k_0 is too small then there will be too many positives as is the case here; a *majority-vote* vote, i.e. here $k_0 = 6$, would do an acceptable job in terms of precision, recall and F1 statistic (Murphy, 2012), but this does depend on the false positive ratio in particular (as biological networks are sparse), as well as the size of the ensemble of inference algorithms, $\mathcal{O} = \{O_1, O_2, \ldots, O_k\}$.

Ensemble Estimators Are not Always Better than Individual Estimators

In many practical settings we may want to use ensemble estimators because intuitively we recognise that our individual estimators are far from perfect. We illustrate the level to which this holds true by comparing an ensemble of equally well performing estimators (like in the example above) with a single superior estimator, characterised by false positive and false negative probabilities s^* and t^*, subject to $s^* < s$ and $t^* < t$.

We start by investigating the ensemble estimator false negative probability

$$T = \sum_{\kappa=k_0}^{k} \binom{k}{\kappa} t^{\kappa}(1-t)^{k-\kappa} = \binom{k}{k_0} t^{k_0}(1-t)^{k-k_0}\, {}_2F_1(1, k_0-k; k_0+1; \frac{t}{t-1}), \tag{25}$$

where ${}_2F_1$ is the Hypergeometric function (Arfken *et al.*, 2013) (which can be conveniently approximated for small arguments, see Appendix B). From this we can determine when the ensemble false negative rate, T, will be greater than t for the individual estimators, by tabulating the values of T and identifying when $T > t$, or

$$t < \binom{k}{k_0} t^{k_0}(1-t)^{k-k_0}\, {}_2F_1(1, k_0-k; k_0+1; \frac{t}{t-1}).$$

Equally for the false positive rate, of the ensemble estimate we obtain

$$S = \binom{k}{k_0} t^{k_0}(1-s)^{k-k_0}\, {}_2F_1(1, k_0-k; k_0+1; \frac{s}{s-1}), \tag{26}$$

We see that for low thresholds, k_0, the ensemble error rate can be greater than that of the individual estimator; this is because the stringency of the ensemble prediction is then reduced, because the cumulative probability of a sufficiently small number of estimators to "accidentally" agree is greater than the error rate of the individual estimator. We show this for two false negative rates, $t = 0.1$, and $t = 0.05$, in figure 5 (A) and (B).

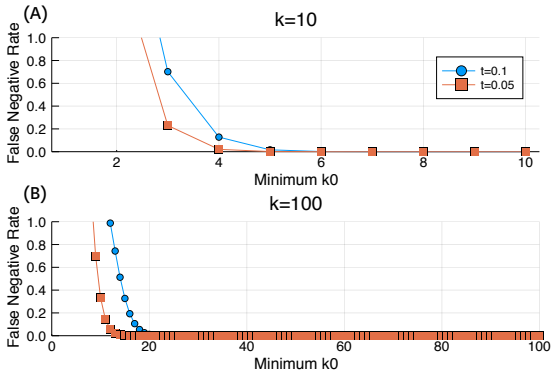

Figure 5: *Illustrative cases where the error rate of an ensemble estimator exceeds the error rate of an individual estimator for $k = 10$ and $k = 100$. Reassuringly, this only happens for small values of the threshold k_0.*

4.1 Heterogeneous Ensembles of Network Inference Methods

Above we have considered model averaging of heterogeneous sets of models, and later network inference of ensembles of statistically identical methods. To conclude we now turn to what happens when our ensemble of network predictors contains useful and less useful or useless predictors: under what circumstances can we trust an ensemble that contains good as well as bad estimators? We should issue a note of caution, that most network inference algorithms have severe shortcomings. Nevertheless some are worse than others and the original paper by Marbach et al only payed little attention to this.

We focus on the case of a small set of predictors, $k = 10$, and two classes of methods: a set of *useful* methods with error rates t_1 and s_1; and a set of *useless* methods (to stick to the terminology used above) with error rates $t_2 > t_1$ and $s_2 > s_1$. In figure 6 we show, again for a case modelled on a likely human gene regulation network(Stumpf *et al.*, 2008), the likely true and false positive predictions arising from ensembles with different numbers of good vs bad link prediction methods. The basic lesson is that the good methods have to outnumber the bad methods, otherwise especially the precision will suffer. Here we have chosen a simple majority-vote criterion, and to bring precision and F1 statistic to a satisfying level (say in excess of 0.7) requires essentially purging the ensemble of the weakly performing methods (i.e. $k_1 \gtrsim 8$). This only points to the extent to which poor methods can compromise the performance of ensemble estimators (and the `Jupyter` notebook can be used to explore this more widely).

The usual caveats referred to above — primarily the

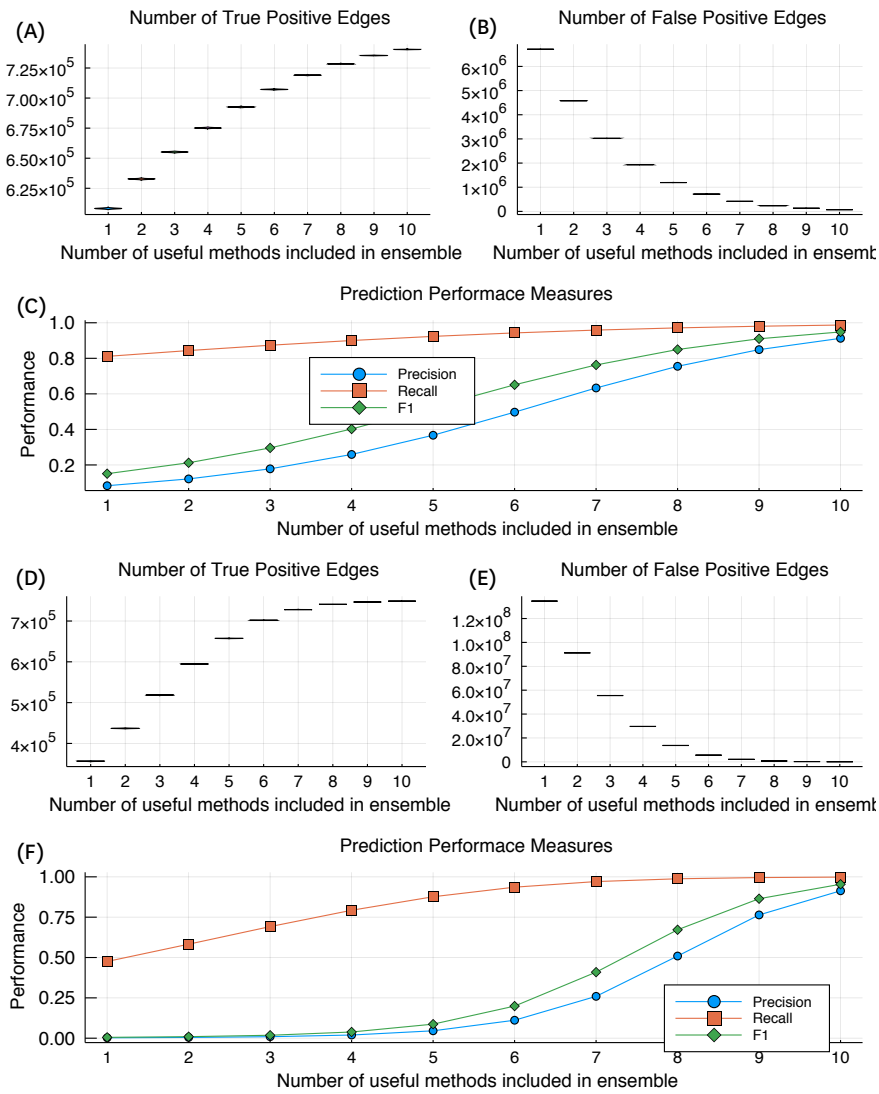

Figure 6: *Illustration of the performance of ensembles of network inference methods with with false positive and false negative probabilities* $s = t = 0.1$. *Real biological networks are sparse and unless false-positives are controlled in the ensemble; once false-positives are over-controlled (by demanding a larger number of methods to score an edge), the recall deteriorates.This is reflected in the top panels (the number of true and false positives (we generated 1,000 random inferred networks but because of the large network size and the binomial nature of the process the distributions around the mean are tight. The lower panel shows the precision, recall and F1 statistics (Murphy, 2012) (see also Appendix A) as a function of the minimum number of methods that need to positively score an edge (again confidence intervals are very tight to be effectively hidden by the symbols.*

fact that *a priori* we do not know the statistical performance of each method — apply, and will likely exacerbate any performance decrease. In sparse networks especially, poor estimators will result in inflation of false positive results, in particular when $k_1 < k_0$the Especially for the false-positive rate in very sparse networks poor network inference methods can lead to overall poor performance: in a directed network there are $N \times (N-1)$ interactions possible; correlation and other methods based on undirected graph structures would only have half that number. But this is still vastly greater than the number of existing edges; e.g. in the case of the human network we would have on the order of 4.8×10^8 potential interactions of which only some 750,000 are expected to exist(Stumpf *et al.*, 2008). So even for $s_2 = 0.5$ there would be about 470,000 cases where ten such methods would agree and score an edge (and 1.8×10^8 if a majority vote rule were applied).

5 Conclusion

In the preceding sections we have shown that ensemble estimators are not as robust as has sometimes been claimed, or incorrectly surmised from the success of community average results in the DREAM competition (Marbach *et al.*, 2012); there, of course, it had already been shown that certain subsets of estimators can give better results than others. The relative balance between the *Madness of Crowds* and the *Wisdom of Crowds* has to our knowledge not been discussed in these concrete terms; but here, too, we have only scratched the surface.

Returning first to the results on multi-model inference from mechanistic models with information criteria or formal model selection, we can distill two points: (i) ensembles of mechanistic models that are reasonably defined (Aijo & Bonneau, 2016) (i.e. their construction incorporates any available mechanistic insights; duplicate models are avoided; the model is predictive and can be used to generate data that can be compared with real observations/data) can be combined with the aid of model selection criteria or Bayesian posterior model probabilities with relative ease and safety; (ii) the inclusion of what we referred to as nuisance models can hamper model averaging and ensemble approaches if they come to predominate the model universe $\mathcal{M}$. Such situations could become more likely as model spaces are explored exhaustively(Scholes *et al.*, 2019) or automatically (Sunnåker *et al.*, 2014). Because of the formalism connecting different model selection criteria, Eqn. (10), these are general results. We note that we have deliberately chosen an abstract representation of the model averaging procedure. On top of the problems addressed here there are a host of technical and computational problems linked to statistical inference and model-fitting procedures (see e.g. Cade (2015)). These will add further to the problems discussed here, which really represents a best-case scenario.

So in essence the construction of the models in $\mathcal{M}$ is of the essence (May, 2004; Kirk *et al.*, 2015). As far as the robustness of model averaging or ensemble approaches to prediction and analysis are concerned nothing is to be gained by increasing the size of $\mathcal{M}$ beyond the (already quite large) set of models which may represent mechanisms at work in reality.

In the context of network inference the situation is similar; here we consider the inference of pairwise interactions among nodes, again under idealised conditions. But rather than using weights based on suitable statistical criteria, which are typically absent in the context of network reconstruction (with some exceptions), we base the ensemble predictor on some voting criteria applied to the set of predictions from the different methods. Majority vote will typically be a sound criterion for a set of reasonable estimators, though not necessarily optimal from the perspective of precision and F1 statistics. This is because biological networks are sparse and false-positives will predominate the inferred networks unless they are carefully controlled. For a set of statistically similar powerful inference methods, a conservative criterion for scoring edges will improve on the overall performance of the individual estimators, however.

But again we find that the poor performance of some methods can drag down the performance of an ensemble estimator or network predictor. So like in the construction of the model universe *ethicalM* before, the make-up of the ensemble of network inference methods, $O_\kappa \kappa = 1, \ldots, k$, does matter considerably (as was also found in the empirical study of Marbach *et al.* (2012)).

The problem of network inference has long been known to be challenging. One reason for this (in addition to the large-scale testing problem) is that we do not have a fair way to score and compare the performance of different network inference approaches. The most promising existing approaches are typically computationally expensive and rigorous *in silicon* assessment of performance as well as the factors influencing performance is often seen as computationally prohibitively expensive. There is also a danger of biasing simulated data in favour of a given method and the DREAM competition has aimed — with some success we feel — to avoid this, and other approaches have followed suite. Clearly

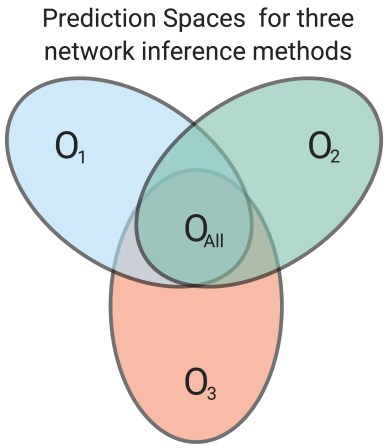

Figure 7: *Illustration of different predictors which capture different aspects of the data. If predictors, O_1, O_2 and O_3 have particular performance preferences for certain types of interactions then combining them may improve the ensemble estimation. But this depends crucially on the false-positive rate.*

more effort in this domain is needed and computational cost should not preclude rigorous assessment; this situation is mirrored in other areas of genomics and systems biology, e.g. pseudo-temporal ordering approaches have until recently (Saelens *et al.*, 2019) rarely been rigorously tested. But what is also needed are approaches which allow us to assign confidence to inferred networks, or, more specifically, predicted interactions without recourse to a gold-standard (Chan *et al.*, 2017). This would most likely have to be indirect and based on biological expectations/knowledge related to the scenario under investigation; or it could be directly, if other data are collected simultaneously alongside mRNA quantifications.

One of the potential initial attractions of using a panel of network inference algorithms is that different methods may capture different aspects of the data and in concert may provide a more complete representation of a "true" network of functional relationships among the genes in an organism under a given scenario. While appealing this notion needs to be viewed with caution. First, we would have to be able to tell by some *a priori* way for which type of candidate interactions, say predictor O_1, is superior to the other predictors and we should accordingly weigh it higher. Second, and by the same token, we would no longer be able to use some form of majority voting to filter out the abundant false-positives. Third, information in data is finite and while different methods can perhaps tap into slightly different aspects of variability (e.g. mutual information vs regression) they all have their own limitations, which are still only poorly understood for many methods. So in essence looking at different methods is most powerful, when we can base the ensemble prediction on a consensus or majority vote.

In summary, unless we know the constituency of the model universe, $\mathcal{M}$, or the ensemble of predictors, O_κ, we have little way of telling whether we are dealing with a *Madness of crowds* or a *Wisdom of crowds* scenario.

Software

A Jupyter Notebook containing the Julia code to reproduce all the computational results here, and to explore the effects of the *madness of crowds* in network inference and model averaging is available at `https://github.com/MichaelPHStumpf/Madness-Of-Crowds`

A Assessing the performance of network inference methods

Real biological networks are sparse (Lèbre *et al.*, 2010); this means that a predictor which scores each candidate edge to be absent can have high performance if the number of false-positives is heavily influencing how we quantity performance of a network inference methods. We therefore focus here on *precision* and *recall* and a derived statistic, the $F1$ statistic (Lèbre *et al.*, 2010; Murphy, 2012). We denote the numbers of true and false positive inferences (i.e. scored edges in the context of network inference) by TP and FP, and the true and false negatives by TN and FN. Then the *precision*, P, is given by

$$P = \frac{TP}{TP + FP},$$

and the *recall* by

$$R = \frac{TP}{TP + FN}.$$

The $F1$ statistic is given by

$$F1 = 2\frac{P \times R}{P + R}.$$

These statistics are confined to the $[0, 1]$ range, with larger values indicating better performance, and the $F1$ statistic in particular becomes maximal at the point where the curves for P and R intersect.

B Approximating the Hypergeometric function

The Hypergeometric function (Arfken *et al.*, 2013), ${}_2F_1$, appearing in Eqn. (25) can be unwieldy to evaluate for large k and k_0 and for sufficiently small of t we can Taylor-expand it around $t = 0$ and then obtain (if we restrict the expansion of the Hypergeometric function to third order)

$$T = \binom{k}{k_0} t^{k_0}(1-t)^{k-k_0}\left(1 - t\frac{k-k_0}{k_0+1} \times \left(1 + \frac{t(2k_0-k-3)}{(k_0+2)} - \frac{t^2\left(k^2 - k(4k_0+9) + 4k_0(k_0+4) + 14\right)}{(k_0+2)(k_0+3)}\right) + O\left(t^4\right)\right) \tag{27}$$

From this we can determine when T will be greater than t by solving

$$t > \binom{k}{k_0} t^{k_0}(1-t)^{k-k_0},$$

because the term in the brackets in Eqn. (27) is ≤ 1.

References

Aijo T & Bonneau R (2016). *Biophysically Motivated Regulatory Network Inference: Progress and Prospects. Human heredity*, **81**:62–77.

Akaike H (1974). *A new look at the statistical model identification.* In *Selected Papers of Hirotugu Akaike*, pp. 215–222. Springer.

Apgar J F, Apgar J F, Witmer D K, Witmer D K, *et al.* (2010). *Sloppy models, parameter uncertainty, and the role of experimental design. Molecular Biosystems*, **6**:1890–1900.

Arfken G, HJ W, & Harris F (2013). *Mathematical Methods for Physicists.* Academic Press.

Babtie A C & Stumpf M P H (2017). *How to deal with parameters for whole-cell modelling. Journal of the Royal Society, Interface / the Royal Society*, **14**:20170,237.

Babtie A C, Kirk P, & Stumpf M P H (2014). *Topological sensitivity analysis for systems biology. Proceedings of the National Academy of Sciences of the United States of America*, **111**:18,507–18,512.

Baker R E, Peña J M, Jayamohan J, & Jérusalem A (2018). *Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biology Letters*, **14**:20170,660.

Barnes C P, Silk D, Sheng X, Stumpf M P H, *et al.* (2011). *Bayesian design of synthetic biological systems. Proceedings of the National Academy of Sciences of the United States of America*, **108**:15,190–15,195.

Box G E P (1976). *Science and Statistics. Journal of the American Statistical Association*, **71**:791–799.

Burnham K P & Anderson D R (2013). *Model Selection and Inference.* A Practical Information-Theoretic Approach. Springer Science & Business Media.

Busetto A G, Hauser A, Krummenacher G, Sunnåker M, *et al.* (2013). *Near-optimal experimental design for model selection in systems biology. Bioinformatics (Oxford, England)*, **29**:2625–2632.

Cade B S (2015). *Model averaging and muddled multimodel inferences. Ecology*, **96**:2370–2382.

Camacho D M, Collins K M, Powers R K, Costello J C, *et al.* (2018). *Next-Generation Machine Learning for Biological Networks. Cell*, **173**:1581–1592.

Castro D M, de Veaux N R, Miraldi E R, & Bonneau R (2019). *Multi-study inference of regulatory networks for more accurate models of gene regulation. PLoS computational biology*, **15**:e1006,591.

Chan T E, Stumpf M P H, & Babtie A C (2017). *Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. Cell Systems*, **5**:251–267.e3.

Choobdar S, Ahsen M E, Crawford J, Tomasoni M, *et al.* (2019). *Assessment of network module identification across complex diseases. Nature methods*, **16**:843–852.

Cox D (2006). *Principles Of Statistical Inference.* Cambridge University Press.

Dufresne E, Harrington H A, & Raman D V (2018). *The geometry of Sloppiness. Journal of Algebraic Statistics*, **9**:30–68.

Erguler K & Stumpf M P H (2011). *Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. Molecular Biosystems*, **7**:1593–1602.

Gerardin J, Reddy N R, & Lim W A (2019). *The Design Principles of Biochemical Timers: Circuits That Discriminate Between Transient and Sustained Stimulation. Cell Systems*, **102**:100,651.

Green P J, Latuszynski K, Pereyra M, & Robert C P (2015). *Bayesian computation: a summary of the current state, and samples backwards and forwards. Statistics and Computing*, **25**:835–862.

Gutenkunst R N, Gutenkunst R N, Waterfall J J, Waterfall J J, *et al.* (2007). *Universally sloppy parameter sensitivities in systems biology models. PLoS computational biology*, **3**:1871–1878.

Hoeting J, Hoeting J A, Madigan D, Madigan D, *et al.* (1999). *Bayesian model averaging: a tutorial. Statistical science.*

Ingram P J, Stumpf M P H, Stumpf M P H, & Stark J (2006). *Network motifs: structure does not determine function. BMC Genomics*, **7**:108.

Jaeger J & Monk N (2015). *Everything flows: A process perspective on life. EMBO reports*, **16**:1064–1067.

Karr J R, Sanghvi J C, Macklin D N, Gutschow M V, *et al.* (2012). *A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell*, **150**:389–401.

Karr J R, Williams A H, Zucker J D, Raue A, *et al.* (2015). *Summary of the DREAM8 Parameter Estimation Challenge: Toward Parameter Identification for Whole-Cell Models. PLoS computational biology*, **11**:e1004,096.

Kirk P D W, Kirk P, Thorne T W, & Stumpf M P H (2013). *Model selection in systems and synthetic biology. Current opinion in biotechnology*, **24**:767–774.

Kirk P D W, Babtie A C, & Stumpf M P H (2015). *Systems biology (un)certainties. Science*, **350**:386–388.

Kolaczyk E D (2009). *Statistical Analysis Of Network Data: Methods And Models*. Springer.

Koller D & Friedman N (2009). *Probabilistic Graphical Models: Principles And Techniques*. MIT Press.

Kueffner R, Zach N, Bronfeld M, Norel R, *et al.* (2019). *Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. Scientific reports*, **9**:690–14.

Lang M & Stelling J o r (2016). *Modular Parameter Identification of Biomolecular Networks. SIAM Journal on Scientific Computing*, **38**:B988–B1008.

Lèbre S, Becq J, Devaux F, Stumpf M P H, *et al.* (2010). *Statistical inference of the time-varying structure of gene-regulation networks. Bmc Systems Biology*, **4**:130.

Leon M, Woods M L, Fedorec A J H, & Barnes C P (2016). *A computational method for the investigation of multistable systems and its application to genetic switches. Bmc Systems Biology*, **10**:130–12.

Liepe J, Filippi S, Komorowski M, Stumpf M P H, *et al.* (2013). *Maximizing the information content of experiments in systems biology. PLoS computational biology*, **9**:e1002,888.

Ma W, Trusina A, El-Samad H, Lim W A, *et al.* (2009). *Defining network topologies that can achieve biochemical adaptation. Cell*, **138**:760–773.

Mackay C (1841). *Extraordinary Popular Delusions and the Madness of Crowds*. Harriman House Limited.

MacKay D J C (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Marbach D, Costello J C, Küffner R, Vega N M, *et al.* (2012). *Wisdom of crowds for robust gene network inference. Nature Biotech*, **9**:796–804.

May R M (2004). *Uses and abuses of mathematics in biology. Science*, **303**:790–793.

Mc Mahon S S, Sim A, Filippi S, Johnson R, *et al.* (2014). *Information theory and signal transduction systems: from molecular information processing to network inference. Seminars in Cell & Developmental Biology*, **35**:98–108.

Michaelis L & Menten M L (1913). *The kenetics of the inversion effect. Biochemische Zeitschrift*, **49**:333–369.

Murphy K P (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Neuenschwander D E (2017). *Emmy Noether's Wonderful Theorem*. JHU Press.

Penfold C A & Wild D L (2011). *How to infer gene networks from expression profiles, revisited. Interface Focus*, **1**:857–870.

Qiao L, Zhao W, Tang C, Nie Q, *et al.* (2019). *Network Topologies That Can Achieve Dual Function of Adaptation and Noise Attenuation. Cell Systems*, **9**:271–+.

Raue A, Karlsson J, Saccomani M P, Jirstrand M, *et al.* (2014). *Comparison of approaches for parameter identifiability analysis of biological systems. Bioinformatics (Oxford, England)*, **30**:1440–1448.

Robert C P (2007). *The Bayesian Choice*. From Decision-Theoretic Foundations to Computational Implementation. Springer Verlag.

Saelens W, Cannoodt R, Todorov H, & Saeys Y (2019). *A comparison of single-cell trajectory inference methods. Nature biotechnology*, **541**:1.

Scholes N S, Schnoerr D, Isalan M, & Stumpf M P H (2019). *A Comprehensive Network Atlas Reveals That Turing Patterns Are Common but Not Robust. Cell Systems*, **9**:243–257.e4.

Schroeder M (2008). *Number Theory in Science and Communication*. With Applications in Cryptography, Physics, Digital Information, Computing, and Self-Similarity. Springer Science & Business Media.

Secrier M, Toni T, Stumpf M P H, & Stumpf M P H (2009). *The ABC of reverse engineering biological signalling systems. Molecular Biosystems*, **5**:1925–1935.

Segrè E (1984). *From Falling Bodies to Radio Waves*. Classical Physicists and Their Discoveries.

Silk D, Kirk P, Kirk P D W, Barnes C P, *et al.* (2014). *Model Selection in Systems Biology Depends on Experimental Design. PLoS computational biology*, **10**:e1003,650.

Stumpf M P H & Wiuf C (2010). *Incomplete and noisy network data as a percolation process. Journal of the Royal Society, Interface / the Royal Society*, **7**:1411–1419.

Stumpf M P H, Thorne T W, de Silva E, Stewart R, *et al.* (2008). *Estimating the size of the human interactome. Proceedings of the National Academy of Sciences of the United States of America*, **105**:6959–6964.

Sunnåker M, Zamora-Sillero E, López García de Lomana A, Rudroff F, *et al.* (2014). *Topological augmentation to infer hidden processes in biological systems. Bioinformatics (Oxford, England)*, **30**:221–227.

Szederkényi G, Banga J R, & Alonso A A (2011). *Inference of complex biological networks: distinguishability issues and optimization-based solutions. Bmc Systems Biology*, **5**:177.

Thorne K S & Blandford R D (2017). *Modern Classical Physics*. Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics. Princeton University Press.

Thorne T W & Stumpf M P H (2012). *Graph spectral analysis of protein interaction network evolution. Journal of the Royal Society, Interface / the Royal Society*, **9**:2653–2666.

Toni T & Stumpf M P H (2010). *Simulation-based model selection for dynamical systems in systems and population biology. Bioinformatics (Oxford, England)*, **26**:104–110.

Tönsing C, Timmer J, & Kreutz C (2014). *Cause and cure of sloppiness in ordinary differential equation models. Physical review E, Statistical, nonlinear, and soft matter physics*, **90**:023,303.

van der Wijst M G P, de Vries D H, Brugge H, Westra H J, *et al.* (2018). *An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Medicine*, **10**:96–15.

Zheng M M, Shao B, & Ouyang Q (2016). *Identifying network topologies that can generate turing pattern. Journal of theoretical biology*, **408**:88–96.