Model criticism based on likelihood-free inference, with an application to protein network evolution

Oliver Ratmann1, Christophe Andrieu2, Carsten Wiuf3, and Sylvia Richardson4

1Department of Public Health and Epidemiology, Imperial College London, London W2 1PG, United Kingdom; 2Department of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom; 3Bioinformatics Research Center, University of Aarhus, 8000 Aarhus C, Denmark; and 4Centre for Biostatistics, Imperial College London, London W1 1PG, United Kingdom

Mathematical models are an important tool to explain and comprehend complex phenomena, and unparalleled computational advances enable us to easily explore them without any or little understanding of their global properties. In fact, the likelihood of the data under complex stochastic models is often analytically or numerically intractable in many areas of sciences. This makes it more important to simultaneously investigate the adequacy of these models—in absolute terms, against the data, rather than relative to the performance of other models—but no such procedure has been formally discussed when the likelihood is intractable. We provide a statistical interpretation to current developments in likelihood-free Bayesian inference that explicitly accounts for discrepancies between the model and the data, termed Approximate Bayesian Computation under model uncertainty (ABCµ). We augment the likelihood of the data with unknown error terms that correspond to freely chosen checking functions, and provide Monte Carlo strategies for sampling from the associated joint posterior distribution without the need of evaluating the likelihood. We discuss the benefit of incorporating model diagnostics within an ABC framework, and demonstrate how this method diagnoses model mismatch and guides model refinement by contrasting three qualitative models of protein network evolution to the protein interaction datasets of Helicobacter pylori and Treponema pallidum. Our results make a number of model deficiencies explicit, and suggest that the T. pallidum network topology is inconsistent with evolution dominated by link turnover or lateral gene transfer alone.

Bayesian inference | intractable likelihoods | Markov chain Monte Carlo | Approximate Bayesian Computation | model uncertainty

In the quest to comprehend complex observations, hypotheses about underlying mechanisms are formalized in terms of precise mathematical models (1). Much of statistical reasoning then proceeds in an iterative process between data acquisition, data analysis, and model development (2). At the i-th iteration, the interpretation of observed data x0 in terms of some target parameters θ conditional on a tentative, probabilistic model M_i has a long tradition in Bayesian inference (3). The focus is typically on the posterior density f(θ | x0, M_i), which is related to the likelihood f(x0 | θ, M_i) and the prior π_0(θ | M_i) via Bayes’ Theorem:

f(θ | x0, M_i) = f(x0 | θ, M_i)π_0(θ | M_i) / f(x0).

To explore whether the current model M_i is consonant with x_0, and to guide further model development, Bayesian predictive diagnostics (4,5) ask whether x_0 can be viewed as a random observation from a predictive distribution m(x | M_i) in terms of a chosen discrepancy function ρ(x, x_0), x ~ m; interpretation of such diagnostics has been the subject of lively debate (6, 7). Application of this machinery for complex models is provided by the workhorses of Bayesian inference (8), such as Markov Chain Monte Carlo (MCMC), as long as the likelihood is readily evaluable up to a normalizing constant.

In many areas of science, such as econometrics (9), molecular genetics (10), epidemiology (11), and evolutionary systems biology (12), the likelihood is sometimes intractable. Nevertheless, given a value of θ, it is typically easy to simulate data from f(θ, M_i). Approximate Bayesian Computation (ABC), reviewed in ref. 10, proposes to infer θ by comparing simulated data x to the observed data x_0, in terms of a (real-valued) univariate discrepancy ρ that combines a set of (computationally tractable) summaries S = (S_1,...,S_n). In its simplest form, values of θ for which the discrepancies are within τ ≥ 0 are retained to define the “approximate likelihood”

\[ t_τ(θ) = \frac{1}{τ} \int 1(|ρ(S(x), S(x_0))| ≤ τ/2) f(x | θ, M_i) dx \]  \[ (2) \]

in the sense that as τ → 0, \( t_τ(θ) \) should approach the likelihood of the summaries, \( f(S(x_0) | θ, M_i) \); see Fig. 1A and the supporting information (SI). Appendix, subsection S1.2. ABC may be embedded into Bayesian methods to formally select one model from a specified collection of models, or to average them (13–15); see Fig. 1B. However, relative comparisons between models do not convey whether models correspond adequately to the observed data and, without exploring the adequacy of models to explain the data, the meaning of reporting θ from Eq. 2 remains unclear, see Fig. 2.

We continue in believing that “all models are wrong but some are useful” (2), which prompts us to interpret several \( ρ(S_i(x), S_i(x_0)) \) as realizations of real-valued error terms, denoted by ε = (ε_1, ..., ε_n) (16). Error terms are not observed, and must be estimated from the data; we develop a theoretical framework and provide an algorithm, ABCµ, for this purpose when the likelihood is intractable. We intentionally focus on the posterior distributions of components of ε to make probabilistic statements of mismatch between the model and the data (17) and hence to facilitate model criticism, as summarized in Fig. 1C.

Postgenome data such as protein interaction networks (PINs) are now available for a growing number of organisms, (e.g., refs. 18 and 19). They offer a new perspective on the function of all organisms, and are, in addition to individual gene or genomic approaches, increasingly useful to elucidate the evolution of living systems, (e.g., refs. 12, 20, and 21), despite being noisy, incomplete, and static descriptions of the real, transient protein network (22). To elucidate the network evolution of prokaryotes, we here analyze the compatibility of the Treponema pallidum and Helicobacter pylori PIN datasets with a set of competing models inspired by fundamentally different modes of network evolution.

Author contributions: O.R. and S.R. designed research; O.R. performed research; O.R., C.A., C.W., and S.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

1To whom correspondence should be addressed. E-mail: oliver.ratmann@imperial.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0807882106/DCSupplemental.
under one model relative to that number under all other models (is intractable, define the unknown error
quantity as an augmented joint posterior density of model parameters and summary errors.

ABC under Model Uncertainty

Joint Posterior Density of Model Parameters and Summary Errors. For the purpose of model criticism in situations where the likelihood is intractable, define the unknown error \( \varepsilon \) as the random variable with conditional probability distribution

\[
P_{\theta, x_0}(\varepsilon \leq \varepsilon) = \int_{\mathcal{X}} 1(p(S(x), S(x_0)) \leq ef(x|\theta, M_0)) dx. \tag{3}
\]

Next, we assume that \( P_{\theta, x_0}(\varepsilon \leq \varepsilon) \) has a density \( \xi_{\varepsilon, x_0} \) with respect to a proper measure of \( \varepsilon \). It is natural to suggest using this quantity as an augmented likelihood for \( x_0 \) under \( \theta \) while adhering to the current model,

\[
\theta, \varepsilon \rightarrow f_{\theta}(x_0|\theta, M_0) = \xi_{\varepsilon, x_0}(\varepsilon). \tag{4}
\]

We thus capture the direct information brought by the discrepancies \( \rho \) on \( \theta \) and/or model \( M_0 \) in a scalar value. For a given prior \( \pi_{\theta, x}(\theta|\varepsilon, M_0) \), we embrace two aspects of statistical reasoning,

\[
\text{parameter inference and model criticism, simultaneously by the joint posterior density}
\]

\[
f_{\theta}(\theta|\varepsilon, x_0, M_0) = \xi_{\theta, x_0}(\varepsilon)\pi_{\theta, x}(\theta|M_0)f_{\theta}(x_0| M_0), \tag{5}
\]

using the data once.\(^{4}\) In practice, we take \( \pi_{\theta, x} = \pi_{\varepsilon, x} \), reflecting our inability to quantify a priori model adequacy for a value of \( \theta \). The posterior relationship Eq. 5 exploits the dependence between model error and model parameterization. ABC only infers model parameterization from realized model errors after simulation and does not question the adequacy of the likelihood model.

The simplest algorithm to sample from Eq. 5 is:

**Std-ABCp1** Sample \( \theta \sim \pi_{\theta}(M_0) \), simulate \( x \sim f(\cdot|\theta, M_0) \) and compute \( \varepsilon = \rho(S(x), S(x_0)) \).

**Std-ABCp2** Accept \((\theta, \varepsilon)\) with probability proportional to \( \pi_{\theta}(\varepsilon|M_0) \), and go to Std-ABCp1.

**Interpretation of the Marginals: Parameter Inference and Model Criticism.** The thrust of this article is to recognize the utility of the unknown error \( \varepsilon \) for model criticism. By design, nonzero values of \( \rho \) indicate discrepancies between the model and the data, so that intuitively, only if the model matches the data, we expect the mode of \( \xi_{\varepsilon, x_0}(\varepsilon) \) to be on average zero for some value of \( \theta \) if the summaries behave sufficiently well. Parameter inference based on the marginal posterior distribution \( f_{\theta}(\theta|x_0, M_0) \) is justified in an “approximate likelihood” sense, because, under regularity assumptions (SI Appendix S1.1) on \( \xi_{\varepsilon, x_0}(\varepsilon) \) which we assume throughout this section,

\[
f_{\theta}(\theta|x_0, M_0) \propto \int \pi_{\theta}(p(S(x), S(x_0))|M_0)f(x|\theta, M_0)\pi_{\varepsilon}(\varepsilon|M_0)dx. \tag{6}
\]

Setting \( \pi_{\varepsilon}(\varepsilon|M_0) = 1|\varepsilon| \leq \tau/2)/\tau \), we recover the “standard” ABC approximation Eq. 2; please see SI Appendix S1.2 for more details and examples. We can interpret the variety of ABC kernels as exerting a particular prior belief on the adequacy of the current model (23). In agreement with methods of ABC (SI Appendix S1.2), we always choose a prior \( \pi_{\varepsilon}(\varepsilon|M_0) \) with mode at zero to accommodate a prior belief that the model is plausible. For the purpose of parameter inference, it is sufficient to “plug-in” realized errors in Eq. 6, but here we also focus on the marginal posterior error \( f_{\theta}(x_0|M_0) \). For the prior predictive error density \( L_{\theta}(\varepsilon) = \int \delta(\rho(S(x), S(x_0)) = \varepsilon)\pi_{\varepsilon}(\varepsilon|M_0)dx \) we have that

\[
f_{\theta}(\theta|x_0, M_0) = L_{\rho}(\varepsilon)\pi_{\theta}(\varepsilon|M_0)f_{\theta}(x_0|M_0) \tag{7}
\]

(see SI Appendix S1.1). Hence, \( f_{\theta}(x_0|M_0) \) can be understood as an error density under the prior predictive distribution that is weighted according to error magnitude. Small error boosts the prior belief for a particular value of \( \theta \), see Eq. 6. We thus prefer model criticism based on Eq. 7 rather than \( L_{\rho}(\varepsilon) \) as it focuses on those \( \theta \) actually inferred from the perspective of Eq. 5, and attenuates the dependence of \( L_{\rho}(\varepsilon) \) on \( \pi_{\varepsilon}(\varepsilon|M_0) \). This dependence is undesirable in that a faultless model could appear questionable under unfortunate prior choice (SI Appendix S1.3). For the practitioner, we provide a computationally feasible method for model criticism within the prior predictive setting as an alternative to using data-splitting techniques that are here difficult or too expensive to construct (5, 7, 17).

**Multidimensional Error Terms \( \varepsilon \).** The complexity of the settings to which ABC is typically applied makes it difficult to think of a universal discrepancy function \( \rho \). The joint posterior distribution of multiple errors \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_K) \), corresponding to \( K \) discrepancies

\[\varepsilon_1 \leq \varepsilon_2 \leq \ldots \leq \varepsilon_K \]

\[\text{for clarity, we subscript } n_{\theta} \text{ and } x_{\varepsilon} \text{ to denote the priors in } \theta \text{ and } \varepsilon, \text{ respectively. From now on, we drop the conditioning of } x_{\varepsilon} \text{ on } \varepsilon. \text{ Finally, we denote with } m(x|M_0) \text{ the prior predictive density } f(x|\theta, M_0)\pi_{\theta}(\theta|M_0)\text{.} \]
that better controls the volatility of the simulated datasets, we might capture aspects of model discrepancy. To control severe, and instead a linear combination of summaries is typically addressed the codependencies of ABC, that the likelihood model. For reference, we applied standard smoothing approximation

\[ f_\theta(\theta, \mathbf{x}|\mathbf{x}_0, M_0) \propto \pi_0(\theta|M_0) \pi_0(\mathbf{x}|M_0) \min_k \hat{\epsilon}_k(\mathbf{x}_0, \mathbf{x} )/f(\mathbf{x}_k, M_0). \]

on the auxiliary space (\( \theta, \epsilon, x \)). Various Monte Carlo strategies (8, 24) may be devised to sample from Eq. 9 (SI Appendix S1.8); our MCMC implementation (SI Appendix S1.9), particularly addresses the co-dependencies of \( \rho_k \) with a careful choice of \( q(\epsilon \to \epsilon') \). Suppose an initial sample \( (\theta, \epsilon) \) and prior specifications (SI Appendix S1.10);

\[ \rho_k(S_k(x_k), S_k(x_{0})) \propto \pi_0(\theta|M_k) \pi_0(\mathbf{x}|M_k) \min_k \hat{\epsilon}_k(S_k(x_k))/f(\mathbf{x}_k, M_k) \]

is in the multivariate case exacerbated by the fact that the unknown error terms are correlated by design, and easily outnumber the \( \theta \)’s. To obtain a smoothed, stabilized approximation to \( \hat{\epsilon}_k(\mathbf{x}_0; \mathbf{x} ) \) that better controls the volatility of the simulated datasets, we employ kernel density estimates \( \hat{\epsilon}_k(\mathbf{x}_0; \mathbf{x} ) : = 1/(Bh_k) \sum_{b=1}^{B} K(||x_k - \mathbf{x}||/h_k) \) in line with ABC (SI Appendix S1.15). In theory, this corresponds to replacing \( \hat{\epsilon}_k(\mathbf{x}_0; \mathbf{x} ) \) in Eq. 8 with min \( \int h_k^{-1} K(|x_k - x|)/h_k \hat{\epsilon}_k|_{x, x}(v)dv \). In practice (SI Appendix S1.7), we set \( B = 50 \) and attain under technical modifications (see Material and Methods, section 3) a smoothing approximation

\[ \hat{\epsilon}_k(\theta, \mathbf{x}_0; \mathbf{x} )|_{\mathbf{x}_0, \theta} \]

that models mismatch with posterior densities of multiple error terms when the likelihood is intractable. Suppose we have obtained a dataset \( x_0 \) of 200 independent samples. We believe each sample of \( x_0 \) to be generated from \( f(\theta, M_0) = N(\theta, 1) \) with \( \theta \) unknown, whereas in reality \( x_0 \) is exponential with rate \( \theta = 0.2 \) (denoted with \( M_0 \)). By construction, the sample mean (MEAN) is a sufficient statistic to estimate both \( \theta \) and \( \theta_0 \). To illustrate one iteration of ABC-MCMC, suppose we sample \( \theta = 3, \pi(\theta) = 0.94 \) from the priors (respectively, uniform and exponential). We generate 50 errors \( x - x_0 \) (A, red points for \( x_0 \) and \( f(\theta, M_0) \), estimate the associated error density \( \hat{\epsilon}(\mathbf{x}) \) (A, black line) with a biweight kernel, and compute \( \hat{\epsilon}_{\text{MEAN}}(\mathbf{x}) \) (A, blue point). Using algorithm ABC, we estimated posterior quantities of \( \theta \) and \( \epsilon \) under various summary statistics. Summarizing \( \mathbf{x} \) only with MEAN, the method indicates no model mismatch (B, row 1, column 2; 95% high posterior density interval (HPD)) and \( \hat{\epsilon} \) is estimated as if \( x_0 \) were indeed \( N(\theta, 1) \) (B, row 1, column 1; posterior mean). Reminding ourselves that likelihood-free inference is honest in that inference is here based on \( x_0 \), we see that the algorithm samples correctly from Eq. 9. Observing that under \( M_1 \), the standard deviation (SD) is 1 independently of \( \theta \) and that SD is 1/\( \theta \) under \( M_2 \), we recognize that progress is possible when a comprehensive set of summaries is employed. Repeating our method based on MEAN, SD, and the 0.25 quantile (QU0.25), targeting Eq. 9, we find that all error terms indicate model mismatch (B, row 2, column 2-4; and C, posterior error densities for 4 runs of ABC, starting from oversampled initial values (colored solid lines) versus the (dashed) prior \( \pi(\epsilon|\theta(M_0)) \).

If posterior quantities of comprehensive error terms correctly diagnose model mismatch as in this example, we recommend questioning the interpretability of \( \theta(x_0) \) in terms of the likelihood model. For reference, we applied standard (Rejection) ABC to this example; conditioning on the summaries MEAN, SD, and QU0.25, we find that the numerical estimates of \( \theta(x_0) \) agree between ABC and ABC-MCMC (with \( \epsilon \) fixed as in B, third row).

**Table 1. Acceptance rate and average mixing quality in \( \theta \)**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>B</th>
<th>acc.prob</th>
<th>Burn-in</th>
<th>10^3 ( \text{n_eff}/n )</th>
<th>( \text{n_eff}/n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC-MCMC (26)</td>
<td>1</td>
<td>0.002</td>
<td>10^5</td>
<td>0.7</td>
<td>13.4</td>
</tr>
<tr>
<td>Zoom-ABC-MCMC (12)</td>
<td>50/1</td>
<td>0.002</td>
<td>800</td>
<td>0.7</td>
<td>23.8</td>
</tr>
<tr>
<td>AUX-ABC (27)</td>
<td>1</td>
<td>0.01</td>
<td>10^5</td>
<td>0.6</td>
<td>4.5</td>
</tr>
<tr>
<td>ABC-MCMC with asymmetric walk in ( \epsilon ), Eq. S12</td>
<td>50</td>
<td>0.36</td>
<td>77.1</td>
<td>25.6</td>
<td>24.6</td>
</tr>
</tbody>
</table>

Performance results are obtained from inference from the \( H. pylori \) PIN dataset; tuning parameters have been optimized for each algorithm separately (SI Appendix S1.15). The effective sample size \( \text{n_eff} \) is taken as an indicator of mixing quality across \( n \) iterations (SI Appendix S1.14). Importantly, \( \text{n_eff}/\text{CPU hr} \) must be compared relative to the achieved absolute errors (see further SI Appendix Table S2). Higher acceptance rates are not necessarily desirable. As a rule of thumb, we found that, here, rates >0.45 reduced the effective sample size.
comprehensive analyses across 181 prokaryotic genomes suggest that aspects of network growth (ref. 29 and references therein). We on randomly growing graphs, have been proposed to capture over evolutionary timescales, and a number of mechanisms, based against the overinterpretation of marginal diagnostics (see further properties are in our setting typically not independent and caution (HPD) intervals; we remark at this point that marginal prop-
and using centrality measures such as high probability density summaries may enable model criticism; it is our view that choos-
summarized by a factor of 5 to improve the visualization of differences across models for RS2. Whereas PAP and PAP depart in ODBOX from the data, DD+LNK+PA departs (slightly) in RS from the observed PIN, suggesting that only DDA+PA provides an adequate fit the T. pallidum dataset.

\[ S_k(x) \] under the predictive distribution \( \pi(x|M) \) are typically different from those among \( S_k(x_0) \) if the model is not adequate, revealing model inconsistency in terms of conflicting, codependent summaries. As exemplified in Fig. 2, only a comprehensive set of summaries may enable model criticism; it is our view that choosing comprehensive summaries and discrepancy functions is crucial to ensure the approximation quality of Eq. 6 to the likelihood (12) as well as for model criticism based on posterior densities of summary errors (Eq. 7). To explore model adequacy, we recommend investigating various posterior quantities of \( f_x(x|\theta, M) \) and using centrality measures such as high probability density (HPD) intervals; we remark at this point that marginal properties are in our setting typically not independent and caution against the overinterpretation of marginal diagnostics (see further Fig. 3).

**Example: Criticizing Models of PIN Evolution**

The structure of PINs derives from multiple stochastic processes over evolutionary timescales, and a number of mechanisms, based on randomly growing graphs, have been proposed to capture aspects of network growth (ref. 29 and references therein). We briefly motivate three models of network evolution. Recent comprehensive analyses across 181 prokaryotic genomes suggest that lateral gene transfer probably occurs at a low rate, but that, cumulatively, \( \approx 80\% \) of all genes in a prokaryotic genome are involved in lateral gene transfer (30); model PAP\(^7\) is inspired by this scenario as it proposes network evolution in terms of attachment processes only (Materials and Methods, section 2). At least 40% of genes in prokaryotes appear to be products of gene duplication (31). Model DDA+PA (Materials and Methods, section 2) is designed to quantify the potential role of duplication and divergence in network evolution (12). At least for eukaryotes, the formation or degeneration of functional links between proteins (link turnover) is estimated to occur at a fast rate of \( \approx 10^{-5} \) changed interactions per My per protein pair (20). We extend model DDA+PA into Model DD+LNK+PA (Materials and Methods, section 2), which includes link turnover in terms of preferential loss and gain of protein interactions. Crucially, ABC enables us to account for data incompleteness. Previously, we modeled missing data by randomly sampling proteins from the simulated data (RS1) (12). Here, we examine an alternative model that randomly samples from those proteins that have an interaction in the simulated data (RS2). (For the former, we add “r” to the model acronyms.) Necessarily, all models remain conceptually limited and must be cautiously interpreted; for example, the assumption that the network as a whole evolves at homogeneous rates has been questioned (32).

### Models of Network Evolution Inspired by Horizontal Gene Transfer, Duplication-Divergence, and Link Turnover.

We ask whether the T. pallidum PIN topology is compatible with a number of fundamentally different modes of network evolution, in guise of simplified models. We successfully checked (Materials and Methods, section 5) and applied ABC\(\theta\) (Materials and Methods, section 6) to sample from \( f_x(x|\theta, x_0) \) for the models PAP, DDA+PA, and DD+LNK+PA (under both sampling schemes RS1 and RS2); see Fig. 3 and SI Appendix. Figs. S5 and S6. Based on RS1 (12), all models depart significantly in \( \text{F-KAG} \) as exemplified for PAP\(\alpha\) in Table 3. This motivated us to consider alternative models of missing data, and we found no significant departures in FRAG for any of the considered evolution models under RS2; see Table 3. Turning to the 6 remaining discrepancy functions, we observe (Table 3 and Fig. 3), that only model DDA+PA matches the Treponema pallidum PIN adequately, suggesting that an evolutionary mode of duplication-divergence is most consistent with the T. pallidum PIN dataset. Repeating our analysis based on RS2 for the Helicobacter pylori PIN dataset, we could not substantiate our results further, because all considered models provide an adequate fit to the data. This is surprising, because we expect a similar power of our method on both datasets (SI Appendix. S1.19), and may point to qualitative differences among the two PINs owing to different, underlying experimental protocols.

### Conclusion

The growing complexity of realistic models renders Bayesian model criticism increasingly important and difficult. In this article, we provide Bayesian techniques to comprehensively quantify

<table>
<thead>
<tr>
<th>Model Acronym</th>
<th>( f_{\text{CONN}} )</th>
<th>( f_{\text{FRAG}} )</th>
<th>( f_{\text{F-DIA}} )</th>
<th>( f_{\text{F-EO}} )</th>
<th>( f_{\text{F-PAP}} )</th>
<th>( f_{\text{F-Mi}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAP</td>
<td>([-0.11, 0.11])</td>
<td>([-0.11, 0.15])</td>
<td>([0.28, 0.72])</td>
<td>([-0.16, 0.92])</td>
<td>([-0.012, -0.002])</td>
<td>([-0.20, 0.05])</td>
</tr>
<tr>
<td>PAP</td>
<td>([-0.15, 0.14])</td>
<td>([-0.12, 0.16])</td>
<td>([0.02, 0.61])</td>
<td>([-0.60, 0.91])</td>
<td>([-0.010, -0.003])</td>
<td>([-0.66, 0.01])</td>
</tr>
<tr>
<td>DDA+PA</td>
<td>([-0.24, 0.29])</td>
<td>([-0.34, 0.48])</td>
<td>([-0.12, 0.20])</td>
<td>([-0.48, 0.50])</td>
<td>([-0.002, 0.027])</td>
<td>([-0.66, 0.05])</td>
</tr>
<tr>
<td>DD+LNK+PA</td>
<td>([-0.27, 0.21])</td>
<td>([-0.20, 0.17])</td>
<td>([-0.11, 0.55])</td>
<td>([-0.59, 0.32])</td>
<td>([-0.018, -0.006])</td>
<td>([-1.03, -0.16])</td>
</tr>
</tbody>
</table>

Note that the scales of \( f_k \) correspond to the scales of the summaries, so that small numbers are meaningful. We caution against overinterpretation because the errors are not independent. See Materials and Methods, section 6, for further details.

---

\(^7\) Model acronyms are explained in Materials and Methods, section M2, with underlined characters.

---

Fig. 3. Numerical estimates, obtained from ABC\(\theta\), of the approximate posterior error densities \( f_x(x|\theta, M) \) combined with 50% box plots (black bars) for 2 of 7 summaries to quantify departures of 3 competing models of network evolution to the T. pallidum PIN dataset; PAP\(\alpha\) employs sampling scheme RS1, whereas all others use sampling scheme RS2. ODBOX of PAP\(\alpha\) is miniaturized by a factor of 5 to improve the visualization of differences across models for RS2. Whereas PAP and PAP depart in ODBOX from the data, DD+LNK+PA departs (slightly) in RS from the observed PIN, suggesting that only DDA+PA provides an adequate fit to the T. pallidum dataset.

Table 3. Fifty percent high probability density intervals of \( f_k(x) \), indicating model mismatch relative to the T. pallidum PIN

---

<table>
<thead>
<tr>
<th>Model Acronym</th>
<th>( f_{\text{CONN}} )</th>
<th>( f_{\text{FRAG}} )</th>
<th>( f_{\text{F-DIA}} )</th>
<th>( f_{\text{F-EO}} )</th>
<th>( f_{\text{F-PAP}} )</th>
<th>( f_{\text{F-Mi}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAP</td>
<td>([-0.11, 0.11])</td>
<td>([-0.11, 0.15])</td>
<td>([0.28, 0.72])</td>
<td>([-0.16, 0.92])</td>
<td>([-0.012, -0.002])</td>
<td>([-0.20, 0.05])</td>
</tr>
<tr>
<td>PAP</td>
<td>([-0.15, 0.14])</td>
<td>([-0.12, 0.16])</td>
<td>([0.02, 0.61])</td>
<td>([-0.60, 0.91])</td>
<td>([-0.010, -0.003])</td>
<td>([-0.66, 0.01])</td>
</tr>
<tr>
<td>DDA+PA</td>
<td>([-0.24, 0.29])</td>
<td>([-0.34, 0.48])</td>
<td>([-0.12, 0.20])</td>
<td>([-0.48, 0.50])</td>
<td>([-0.002, 0.027])</td>
<td>([-0.66, 0.05])</td>
</tr>
<tr>
<td>DD+LNK+PA</td>
<td>([-0.27, 0.21])</td>
<td>([-0.20, 0.17])</td>
<td>([-0.11, 0.55])</td>
<td>([-0.59, 0.32])</td>
<td>([-0.018, -0.006])</td>
<td>([-1.03, -0.16])</td>
</tr>
</tbody>
</table>

---

Note that the scales of \( f_k \) correspond to the scales of the summaries, so that small numbers are meaningful. We caution against overinterpretation because the errors are not independent. See Materials and Methods, section 6, for further details.
The PIN topology of the prokaryote _T. pallidum_ provides little support for network evolution dominated by link turnover, or by lateral gene transfer alone. We close by cautioning that it is difficult to convincingly associate a formal framework to our high probability density intervals on multiple diagnostic error terms (6, 7). The presented methods will be useful in the initial stages of model and data exploration (16), and in particular, in efficiently scrutinizing several models by direct, inspection of their summary errors (5), prior to more formal analyses (14).

### 3. Combining Multiple Error Terms.

It is difficult to compare $\xi_{\alpha\beta}(x)$ across $k$ without further transformation, because summaries differ in their sensitivity to changes in $\alpha$ (12) so that the scales of the density estimates vary across summaries and (to a lesser extent) across $i$; see SI Appendix, Fig. S1 In Rejection-ABC (SI Appendix, S1.8), summaries may be precomputed and standardized, but this is not applicable in MCMC. We propose to standardize the variance of each $\xi_{\alpha\beta}(x)$, bearing in mind that this might reduce approximation quality in some cases.

### 4. Details of ABC in PINs.

ABC is similar to the MCMC algorithm proposed in ref. 21, the latter also extends the state space, but includes a scalar $\tau$ (circumventing the need to design an efficient proposal $q(\cdot|x)$). We show that ABC with event samples from $f_{\theta}(i|x, \tau)$ provide convergence results for the smoothing approximation (SI Appendix, S1.5 and S1.11), discuss our non-standard proposal kernel (SI Appendix, S1.12), and provide final details (SI Appendix, S1.13). We do not claim that our smoothing approach based on repeated sampling from $f_0(i|0, M)$ comes at no cost. What we contend is that (i) we obtain improved mixing quality relative to ABC within MCMC, owing to a stabilized, numerical approximation of the likelihood with Eq. (9), and (ii) that we can construct more efficient proposal kernels, a point particularly relevant for ABC in PINs, where the number of error terms easily exceeds the dimensionality of $\theta$. With respect to (i), we compared ABC with ABC-MCMC on the summaries COMP, WRR, ODBC, DIA, CC, TET, FRAG, and $\xi_{\alpha\beta}(x)$ (Table S1), exemplifying that ABC results in much improved acceptance rates and better mixing; this has already been suggested by Becquet and Przeworski (28) when $f_0(x|0, M)$ is available in closed form. As for (ii), we compared ABC with ABC-MCMC on the summaries COMP, WRR, ODBC, DIA, CC, TET, FRAG, and $\xi_{\alpha\beta}(x)$ (Table S1), exemplifying that effectively, repeated sampling may improve the efficiency of standard MCMC methods.

### 5. Testing ABC in PINs.

It was unclear whether our implementation is efficient enough to sample from Eq. 6. First, estimates of Eq. 7 might be inexact because as for technically similar algorithms, and/or the PIN topology, in terms of the chosen summaries, might not be informative enough to evidence discrepancies between the model and the data. Second, given our smoothing approximation based on an adaptively chosen bandwidth $h = h(\theta)$, we might be worried that posterior quantities of $h$ may be unreliable. We have addressed both concerns empirically (SI Appendix, S1.16), comforting that ABC provides useful numerical estimates of $f_0(x|0, M)$ to criticize the models of network evolution considered here, and suggesting that samples $x|0, M$ from the marginal of Eq. 9, obtained by ABC, provide a good approximation to $f_0(x|0, M)$.

### 6. Criticizing Models of Network Evolution.

To contrast models PAP, DDA+PA, and DDA+LNK+PA to the _T. pallidum_ PIN dataset, ABC based on the summaries COMP, WRR, ODBC, DIA, CC, TET, FRAG, and $\xi_{\alpha\beta}(x)$ (Table S1), exemplifying that ABC results in much improved acceptance rates and better mixing; this has already been suggested by Becquet and Przeworski (28) when $f_0(x|0, M)$ is available in closed form. As for (ii), we compared ABC with ABC-MCMC on the summaries COMP, WRR, ODBC, DIA, CC, TET, FRAG, and $\xi_{\alpha\beta}(x)$ (Table S1), exemplifying that effectively, repeated sampling may improve the efficiency of standard MCMC methods.

### ACKNOWLEDGMENTS.

We thank M.P.H. Stumpf for stimulating discussions, T. Hinkley for providing an efficient C++ library to evaluate network summaries, and M. Sternberg for comments on an earlier version of the manuscript. We acknowledge financial support from the Engineering and Physical Sciences Research Council, C.W. by the Danish Cancer Society and the Danish Research Councils, and S.R. by the Wellcome Trust; C.A. by an Advance Research Fellowship from the Engineering and Physical Sciences Research Council, C.W. by the Danish Cancer Society and the Danish Research Councils, and S.R. by the Biotechnology and Biological Sciences Research Council and the Centre for Integrative Systems Biology at Imperial College.


