An Optimal Likelihood Free Method for Biological Model Selection

Vincent D. Zaballa [1]   Elliot E. Hui [1]

Abstract

Systems biology seeks to create math models of biological systems to reduce inherent biological complexity and provide predictions for applications such as therapeutic development. However, it remains a challenge to determine which math model is correct and how to arrive optimally at the answer. We present an algorithm for automated biological model selection using mathematical models of systems biology and likelihood free inference methods. Our algorithm shows improved performance in arriving at correct models without a priori information over conventional heuristics used in experimental biology and random search. This method shows promise to accelerate biological basic science and drug discovery.

1. Introduction

Biological cellular systems exhibit super exponential scaling in the number of biological states achieved arising from different combinations and sequences of cell regulators, such as messenger proteins and transcription factors (Letsou & Cai, 2016). This complexity impedes our understanding of diseases and development of therapeutics. We focus here on the combinatorial complexity of biology, defined by the vast number of models and their parameters that describe biological systems.

This combinatorial problem in biology is exemplified by promiscuous signaling, which is the phenomenon of multiple protein ligands in a pathway being able to bind to multiple receptors in a competitive manner. The Bone Morphogenetic Protein (BMP) pathway exemplifies this type of signaling with multiple protein ligands, and type I and II receptors present in the pathway, each combining with one another at different rates to form a complex of ligand, type I, and type II receptor to phosphorylate SMAD 1/5/8 to send a downstream gene expression signal. The BMP pathway can be mathematically modeled by mass action kinetics (Antebi et al., 2017) and previous work demonstrated how to optimally infer BMP models' parameters using Likelihood Free Inference (LFI), also known as Simulation Based Inference (SBI), using the SBIDOEMAN algorithm (Zaballa & Hui, 2021). However, since multiple models have been proposed for the BMP pathway (Antebi et al., 2017; Su et al., 2022), there remains ambiguity in determining which model best describes observed experimental data.

We propose to use the previously developed SBIDOEMAN algorithm and a novel method to approximate a model's marginal probability, $p(\mathcal{M}|\boldsymbol{x_o}, \boldsymbol{\theta})$, within Bayesian Model Averaging (BMA) to select a correct model from a set of models proposed. This method, which we call SBIDOEMAN BMA, uses models' prior distributions of parameters, $p(\boldsymbol{\theta})$, to design optimal experiments using a mutual information approximation $I(\boldsymbol{\theta}, \boldsymbol{x}; d)$ between model parameters and data, then determines the posterior distribution of parameters given observed data, $p(\boldsymbol{\theta}|\boldsymbol{x_o})$, by LFI, and finally approximates a marginal likelihood of a biological model given observed data points, $p(\mathcal{M}|\boldsymbol{x_o}, \boldsymbol{\theta})$. This marginal probability is used as a probability measure of a given model, $\mathcal{M}$, and can be used in BMA to determine the next experiment to evaluate and a weighting of possible models.

Previous work for optimal experimental designs in biological systems studied graphical models describing gene regulatory networks, modeled using Bayesian graphs, (Cho et al., 2016) and M-estimators applied to Gaussian Markov Random fields, (Zheng et al., 2018) both of which have closed-form information measures. By contrast, we evaluate methods exclusive to the LFI setting where likelihoods and closed-form information measures are not tractable. Regarding model selection, trained classifiers have been proposed to classify whether data can fit a proposed model or not (Radev et al., 2021). While useful in model selection, this method does not provide a posterior distribution of models' parameters or design optimal experiments. Our method provides an alternative for scientists who have a model of their system that they can simulate but not evaluate its likelihood function, compare models, and design experiments towards the most promising model. Additionally, our method has the potential to be used with biological highthroughput screening (HTS) systems to increase the efficiency of such systems

---

[1]Department of Biomedical Engineering, University of California Irvine, Irvine, CA, USA. Correspondence to: Vincent Zaballa <vzaballa@uci.edu>, Elliot Hui <eehui@uci.edu>.

to discover novel biology and therapeutics.

In summary, the key contributions of this paper are:

- A method to determine the marginal probability of a model given observed data.

- BMA applied to optimal experimental designs to design experiments for a given model.

- An automated algorithm to design and evaluate experiments in biological models that is compatible with HTS of biological systems.

2. Background

2.1. Modeling the BMP pathway

Two mass action kinetics models have been proposed for the BMP pathway. The one-step model in Equation (1) models type I (A) and type II (B) receptors and a ligand (L) forming a trimer complex in a single step (Su et al., 2022)

$$A + B + L \xrightarrow{K} T. \tag{1}$$
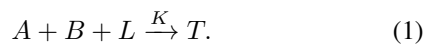
The two-step model in Equations (2) and (3) adds a parameter to model a ligand first binding with a type I receptor before forming a trimeric complex with a type II receptor (Antebi et al., 2017) as follows

$$A + L \xrightarrow{K_1} D \tag{2}$$

$$B + D \xrightarrow{K_2} T. \tag{3}$$

Both models have a complex, T, that phosphrolyates SMAD to send a downstream gene expression signal, S, with a certain efficiency, ϵ as

$$\epsilon T = S. \tag{4}$$

Steady-state signals can be simulated using convex optimization (Su et al., 2022).

2.2. Normalizing Flows

Given a dataset, one may ask what is the probability of a certain data point in the dataset, $p_x(\boldsymbol{x})$, of a variable $\boldsymbol{x}$ with $\mathbb{R}^D$ dimensions. However, this probability density is usually intractable or unknown. Normalizing flows provide a way to answer this question by creating a transformation from a known simple distribution, $p_u(\boldsymbol{u})$, such as a Gaussian distribution, to the data distribution, $p_x(\boldsymbol{x})$, by a series of nonlinear and invertible composition of functions, $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^D$, where $\boldsymbol{f}$ is composed of N functions, $\boldsymbol{f} = \boldsymbol{f}_N \circ \cdots \circ \boldsymbol{f}_1$. We can map from a base distribution to target distribution

using the change-of-variables formula for random variables as

$$p_x(\boldsymbol{x}) = p_u(\boldsymbol{u})|\det \boldsymbol{J}(\boldsymbol{f})(\boldsymbol{u})|^{-1}, \tag{5}$$

where $\boldsymbol{J}(\boldsymbol{f})(\boldsymbol{u})$ is the Jacobian matrix of $\boldsymbol{f}$ evaluated at $\boldsymbol{u}$. See Murphy (2023) for details about normalizing flows.

2.3. Likelihood Free Inference

For models with an implicit or intractable likelihood function, $p(\boldsymbol{x}|\boldsymbol{\theta})$, but whose response may be simulated, we can use LFI methods to approximate the posterior $q(\boldsymbol{\theta}|\boldsymbol{x})$ or likelihood $q(\boldsymbol{x}|\boldsymbol{\theta})$. This can be done by drawing N samples from the prior $p(\boldsymbol{\theta})$ and generating a dataset $\{(\boldsymbol{\theta}_n, \boldsymbol{x}_n)\}_{n=1}^N$ by sampling $\boldsymbol{\theta}_n \sim p(\boldsymbol{\theta})$. Each $(\boldsymbol{\theta}_n, \boldsymbol{x}_n)$ is a joint sample from $p(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})$, and can be used to train a normalizing flow to approximate the posterior $q(\boldsymbol{\theta}|\boldsymbol{x})$ conditioned on an observed $\boldsymbol{x_o}$ (Greenberg et al., 2019; Papamakarios & Murray, 2016) or approximate the likelihood $q(\boldsymbol{x}|\boldsymbol{\theta})$ conditioned on $\boldsymbol{\theta}$. See Papamakarios et al. (2019) for details on applying normalizing flows to LFI.

While LFI provides a method to approximate a model's posterior or likelihood, practical considerations, such as difficulty in rejection sampling in in sequential neural posterior estimate (SNPE) (Greenberg et al., 2019) or prohibitively slow MCMC sampling for sequential neural likelihood estimate (SNLE) (Papamakarios et al., 2018), make LFI methods difficult to implement. In response to this difficulty, recent methods have developed variational methods to approximate the posterior or likelihood (Glöckler et al., 2022; Wiqvist et al., 2021). These methods, referred to here as sequential neural likelihood variational inference (SNLVI), train another normalizing flow, $q_\phi(\boldsymbol{\theta})$, to minimize the divergence from an estimated likelihood, $\phi^* = \mathrm{argmin}_\phi D(q_\phi(\boldsymbol{\theta})||q_\psi(\boldsymbol{x}|\boldsymbol{\theta}))$. We use SNLVI methods to overcome prior practical difficulties in LFI methods.

2.4. Optimal Experimental Design for Implicit Likelihood Model Selection

Optimal experimental designs (OEDs) can be formulated as an optimization (Boyd & Vandenberghe, 2006) or information theoretic problem (MacKay, 1992). Assuming designs are independent of model parameters, we formulate this problem as maximizing the information gain (IG) (Foster et al., 2019), or, the difference in entropy given a proposed design, d, as

$$\mathrm{IG}(\boldsymbol{x}, \boldsymbol{d}) = H[p(\boldsymbol{\theta})] - H[p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{d})]. \tag{6}$$

This objective function can be rewritten as a utility function, $U(\boldsymbol{d})$, that maximizes the mutual information (MI), $I(\boldsymbol{v}; \boldsymbol{y}|\boldsymbol{d})$ between a variable of interest, $\boldsymbol{v}$, and the observed

data, x, at particular design, d. The MI variable of interest, v, can be adapted to the scientific question at hand (Ryan et al., 2016). A gradient-based approach for OEDs was recently proposed for likelihood free models that provides a way to both select a model, $\mathcal{M}$, by BMA and determine its parameters, $p(\theta|\mathcal{M})$ with a minimum number of experiments (Kleinegesse & Gutmann, 2021). Finding designs that optimally discover a model and its parameters can be formulated as the following utility function

$$U(d) = \sum_{\mathcal{M}} \int p(\mathbf{x}|\theta_{\mathcal{M}}, \mathcal{M}, \mathbf{d}) p(\theta_{\mathcal{M}}, \mathcal{M}) \\ \log\left(\frac{p(\theta_{\mathcal{M}}, \mathcal{M}|\mathbf{x}, \mathbf{d})}{p(\theta_{\mathcal{M}}, \mathcal{M})}\right) d\mathbf{x}. \tag{7}$$

We implement equation 7 by simply averaging each model's Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018) MI estimate. The estimated MI is then used as the objective function in Bayesian Optimization using a Gaussian Process (Kleinegesse & Gutmann, 2020).

2.5. Bayesian Model Averaging and the Bayes Factor

The weighting of model probabilities is also known as the Bayes Factor (BF), which we define here as $BF = p(\mathcal{M}_1)/p(\mathcal{M}_0)$, and can be used as a form of model selection where $BF > 10$ is strong evidence for $\mathcal{M}_1$ and $BF < 1/10$ is strong evidence for $\mathcal{M}_0$. We only use the BF for model selection as it uses marginal probabilities that prefer simpler models by the Bayesian Occam's razor effect. Although, this relies on an accurate estimate of the model's marginal probability. See Murphy (2022) for further discussion on various model selection techniques.

2.6. Approximating Model Marginal Probability

To perform model selection, we need an estimate of each model's marginal probability in order to calculate the BF. To do this, we can use a normalizing flow with a Gaussian base distribution $p_u(\mathbf{u})$ that can provide a probability of a model given the posterior parameter distribution and observed data, $p(\mathcal{M}|\boldsymbol{x_o}, \theta, \mathbf{d})$, which is the same as marginal likelihood, $p(\boldsymbol{x_o}|\theta, \mathcal{M}, \mathbf{d})$, when assuming uniform priors over models, $p(\mathcal{M}_i) = 1/|\mathcal{M}|$. This flow is trained by sampling data from the simulator of $\mathcal{M}$ to produce $\boldsymbol{x} \sim p_x(\boldsymbol{x}|\boldsymbol{x_o}, \mathcal{M}, \theta)$ that can be used to train a reverse flow function to a base Gaussian distribution $\boldsymbol{u} = \boldsymbol{f}^{-1}(\boldsymbol{x})$. We propose the following method to approximate the marginal likelihood.

Proposition 2.1. *The marginal likelihood of a model, $\mathcal{M}$, given an observed data vector, $\boldsymbol{x_o}$, and the model's parameters, θ, can be approximated as $p(\boldsymbol{x_o}|\mathcal{M}) \approx 1 - \Phi(\boldsymbol{f}^{-1}(\boldsymbol{x_o}))$, where $\boldsymbol{f}^{-1}$ is the pullback of a trained normalizing flow from the observed data distribution, $p_x(\boldsymbol{x_o})$, to a Gaussian base distribution, $p_u(\boldsymbol{u})$, and Φ is cumulative distribution function of a Gaussian distribution.*

We provide a proof of Proposition 2.1 in Appendix B. Algorithm 1 in Appendix A brings these parts together in SBIDOEMAN BMA to optimally determine models and their parameters.

POLICY	MEDIAN BF	25%	75%
ONE-STEP RANDOM	0.05	0.02	0.17
ONE-STEP EQUI	0.55	0.09	3.72
ONE-STEP SDM BMA	**0.03**	**0.01**	**0.05**
TWO-STEP RANDOM	0.74	0.22	1.28
TWO-STEP EQUI	2.12	0.79	16.11
TWO-STEP SDM BMA	**5.70**	**1.38**	**34.66**

Table 1. Median and interquartile range (IQR) Bayes Factor (BF) values after 5 rounds of experiments for both one-step and two-step datasets compared to random and equidistant experimental design policies. Lower BF is better for the series of one-step models while higher BF is better for the two-step model. For both models, both the median and IQR values are better than competing approaches.
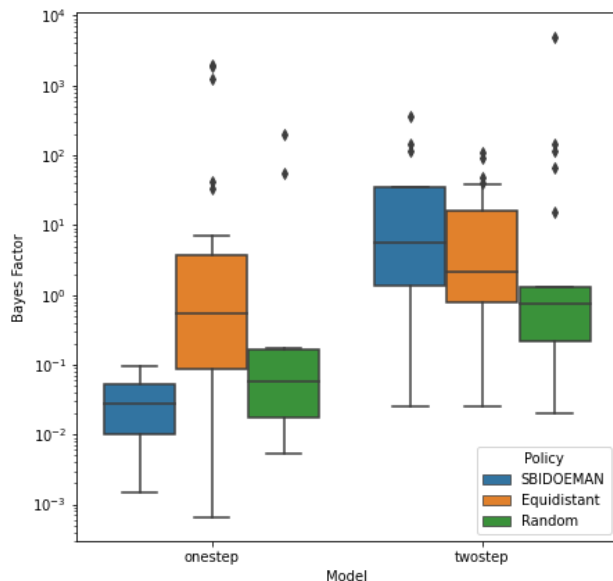

Figure 1. Final Bayes Factor (BF) after 5 design rounds and an ensemble of models. Compared to controls for both models, SBI-DOEMAN BMA performed an order of magnitude better on the one-step model and performed more than two times better than control policies of the two-step model.

3. Results

We evaluated SBIDOEMAN BMA for model selection by evaluating the BF over five rounds of experiments when the one-step BMP pathway was true and when the two-step BMP pathway was true by holding out a single set of parameters for each model, $\theta_T^{\{1,2\}}$. When evaluating performance across designs, we compared to random search, as shown in Figure 2. We also compared final BF to random and equidis-
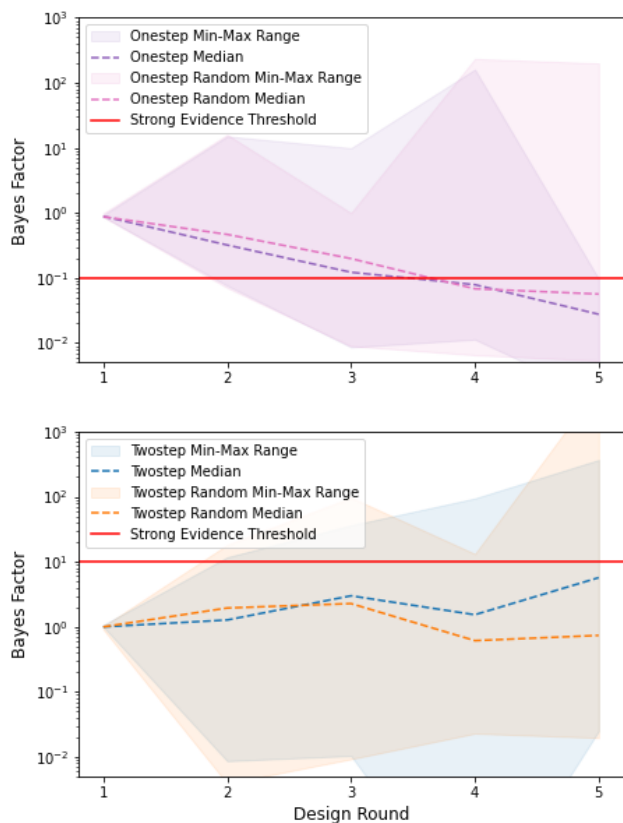
Figure 2. Change in Bayes Factor (BF), $p(\text{twostep})/p(\text{onestep})$, over design round when the one-step (top) and two-step (bottom) models are true. The strong evidence threshold for both models is labeled in red. Top: When the one-step model is true, SBIDOEMAN BF model trends down, indicating the one-step model is true and outperforms random search by the final design. The median BF value for the SBIDOEMAN model strongly suggests the one-step model is true by the fifth round. Bottom: When the two-step model is true the median value of the SBIDOEMAN BF trends upwards, indicating the two-step model is true, and has a median trend that outperforms the competing random search by the last three designs. The two-step model's final value indicates only moderate evidence in favor of the true two-step model.

tant ligand titrations which is a heuristic commonly used in biology to evaluate the response of an assay. Equidistant designs are logarithmically-equal spaced designs across a domain of interest. Here, this would be five equally spaced designs in concentrations from 10^{-3} to 10^{3}ng/ml. Results of the final design comparison are shown in Figure 1 and Table 1.

Examining the change in BF across designs Figure 2, we see that across an ensemble of independent and identically distributed (iid) SBIDOEMAN models that the median performance outperforms random search for both the one-step and two-step models. When looking at the final BF after a budget of 5 designs, as shown in Table 1 and Figure 2, we see that the median performance of SBIDOEMAN BMA

outperforms random and equidistant data, with SBIDOEMAN BMA interquartlie range (IQR) values performing better, or almost better, than competing policy median values. While random search performed as well as SBIDOEMAN BMA in the one-step model, it performs worse in the more complex two-step model, suggesting that principled heuristics and optimal experimental design algorithms are needed for more complex models of biology.

4. Discussion

Robustness and Performance of SBIDOEMAN BMA We demonstrated the ability of SBIDOEMAN BMA to select a true model of the BMP pathway over competing methods, including a standard heuristic in biological systems. Across an ensemble of models, prediction of optimal designs and subsequent posterior evaluation are more efficient. In the process of comparing SBIDOEMAN BMA, we have shown how to estimate a model's marginal probability using normalizing flows. We have also shown that averaging the mutual information estimate between models still results in designs that outperform competing methods in improving the quality of experiments.

Future Work We demonstrated application of SBIDOEMAN BMA to two simple models, the one-step and two-step models, of the BMP pathway, each with two and three parameters, respectively. The efficacy of the SBIDOEMAN algorithm to scale and infer larger implicit likelihoods, such as a BMP model with up to 275 parameters for a BMP two-step model with up to 10 unique ligands, 4 type I receptors, and 3 type II receptors, remains to be seen. Recent innovations in LFI methods, such as SNLVI (Glöckler et al., 2022), show promise to ease such problems, demonstrating robust inference of a neuroscience model with 31 parameters. We also only consider the case without noise or batch effects observed in the model, which true biological systems have. Future work will examine how SBIDOEMAN BMA scales to larger models and how robust it is to noise and batch effects in biological systems.

Computationally, we only used averaging of the mutual information among models to design optimal experiments. Weighting each model's mutual information by its respective marginal probability may lead to improved designs for for the model with more evidence. We used a simple ensemble method to evaluate the performance of iid models using our algorithm, which allows us to measure uncertainty in models' predictions. Mixtures of Experts (MoEs) (Bengio et al., 2013; Shazeer et al., 2017) have also shown promise to improve training and can be combined with ensembling methods to also perform uncertainty quantification (Allingham et al., 2021). These methods could both improve performance and uncertainty quantification in optimal designs for biological models.

Software and Data

We used the hydra configuration manager to track hyperparameters and seeds of experiments (Yadan, 2019). To perform SBI, we used the SBI software library (Tejero-Cantero et al., 2020). The model marginal probability calculation was performed using JAX and Distrax libraries (Bradbury et al., 2018; Babuschkin et al., 2020).

Acknowledgements

References

Allingham, J. U., Wenzel, F., Mariet, Z. E., Mustafa, B., Puigcerver, J., Houlsby, N., Jerfel, G., Fortuin, V., Lakshminarayanan, B., Snoek, J., Tran, D., Ruiz, C. R., and Jenatton, R. Sparse MoEs meet Efficient Ensembles. Oct. 2021. URL https://arxiv.org/abs/2110.03360v1http://arxiv.org/abs/2110.03360.

Antebi, Y. E., Linton, J. M., Klumpe, H., Bintu, B., Gong, M., Su, C., McCardell, R., and Elowitz, M. B. Combinatorial Signal Perception in the BMP Pathway. *Cell*, 170(6):1184–1196.e24, Sept. 2017. ISSN 10974172. doi: 10.1016/j.cell.2017.08.015. URL https://www.sciencedirect.com/science/article/pii/S0092867417309406.

Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Fantacci, C., Godwin, J., Jones, C., Hennigan, T., Hessel, M., Kapturowski, S., Keck, T., Kemaev, I., King, M., Martens, L., Mikulik, V., Norman, T., Quan, J., Papamakarios, G., Ring, R., Ruiz, F., Sanchez, A., Schneider, R., Sezener, E., Spencer, S., Srinivasan, S., Stokowiec, W., and Viola, F. The DeepMind JAX Ecosystem, 2020. URL http://github.com/deepmind.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mutual information neural estimation. In *35th International Conference on Machine Learning, ICML 2018*, volume 2, pp. 864–873, Jan. 2018. ISBN 9781510867963. URL http://arxiv.org/abs/1801.04062.

Bengio, Y., Léonard, N., and Courville, A. Estimating or

Propagating Gradients Through Stochastic Neurons for Conditional Computation. Aug 2013. doi: 10.48550/arxiv.1308.3432. URL https://arxiv.org/abs/1308.3432v1.

Bishop, C. *Pattern Recognition and Machine Learning*. Springer, Jan. 2006. URL https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/.

Boyd, S. and Vandenberghe, L. Convex Optimization. 2006. URL http://www.cambridge.org.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Cho, H., Berger, B., and Peng, J. Reconstructing causal biological networks through active learning. *PLoS ONE*, 11(3), March 2016. ISSN 19326203. doi: 10.1371/journal.pone.0150611. URL /pmc/articles/PMC4773135//pmc/articles/PMC4773135/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4773135/.

Foster, A., Jankowiak, M., O'Meara, M., Teh, Y. W., and Rainforth, T. A unified stochastic gradient approach to designing Bayesian-optimal experiments, Nov. 2019. ISSN 23318422. URL http://arxiv.org/abs/1911.00294.

Glöckler, M., Deistler, M., and Macke, J. H. Variational methods for simulation-based inference. March 2022. doi: 10.48550/arxiv.2203.04176. URL https://arxiv.org/abs/2203.04176v2http://arxiv.org/abs/2203.04176.

Greenberg, D. S., Nonnenmacher, M., and Macke, J. H. Automatic posterior transformation for likelihood-free inference. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pp. 4288–4304. International Machine Learning Society (IMLS), May 2019. ISBN 9781510886988. URL http://arxiv.org/abs/1905.07488.

Kleinegesse, S. and Gutmann, M. U. Bayesian experimental design for implicit models by mutual information neural estimation, Feb. 2020. URL http://arxiv.org/abs/2002.08129.

Kleinegesse, S. and Gutmann, M. U. Gradient-based Bayesian Experimental Design for Implicit Models using Mutual Information

Lower Bounds. May 2021. URL https://arxiv.org/abs/2105.04379v1http://arxiv.org/abs/2105.04379.

Letsou, W. and Cai, L. Noncommutative Biology: Sequential Regulation of Complex Networks. *PLoS Computational Biology*, 12(8):e1005089, Aug 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1005089.

MacKay, D. J. C. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4):590–604, July 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.4.590.

Murphy, K. P. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL probml.ai.

Murphy, K. P. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL probml.ai.

Papamakarios, G. and Murray, I. Fast e-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, number Nips, pp. 1036–1044, May 2016. URL http://arxiv.org/abs/1605.06376.

Papamakarios, G., Sterratt, D. C., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows, May 2018. URL http://arxiv.org/abs/1805.07226.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing Flows for Probabilistic Modeling and Inference. dec 2019. URL http://arxiv.org/abs/1912.02762.

Radev, S. T., D'Alessandro, M., Mertens, U. K., Voss, A., Kothe, U., and Burkner, P. C. Amortized Bayesian Model Comparison With Evidential Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. ISSN 21622388. doi: 10.1109/TNNLS.2021.3124052.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *32nd International Conference on Machine Learning, ICML 2015*, volume 2, pp. 1530–1538. International Machine Learning Society (IMLS), May 2015. ISBN 9781510810587. URL https://arxiv.org/abs/1505.05770v6.

Ryan, E. G., Drovandi, C. C., Mcgree, J. M., and Pettitt, A. N. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*, 84(1):128–154, Apr. 2016. ISSN 17515823. doi: 10.1111/insr.12107. URL http://doi.wiley.com/10.1111/insr.12107.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, Jan. 2017. doi: 10.48550/arxiv.1701.06538. URL https://arxiv.org/abs/1701.06538v1.

Su, C. J., Murugan, A., Linton, J. M., Yeluri, A., Bois, J., Klumpe, H., Langley, M. A., Antebi, Y. E., and Elowitz, M. B. Ligand-receptor promiscuity enables cellular addressing. *Cell Systems*, 0(0), Apr. 2022. ISSN 24054712. doi: 10.1016/j.cels.2022.03.001. URL http://www.cell.com/article/S2405471222001284/fulltexthttp://www.cell.com/article/S2405471222001284/abstracthttps://www.cell.com/cell-systems/abstract/S2405-4712(22)00128-4.

Tejero-Cantero, A., Boelts, J., Deistler, M., Lueckmann, J.-M., Durkan, C., Gonçalves, P. J., Greenberg, D. S., and Macke, J. H. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020. doi: 10.21105/joss.02505. URL https://doi.org/10.21105/joss.02505.

Wiqvist, S., Frellsen, J., and Picchini, U. Sequential Neural Posterior and Likelihood Approximation. Feb. 2021. doi: 10.48550/arxiv.2102.06522. URL https://arxiv.org/abs/2102.06522v2http://arxiv.org/abs/2102.06522.

Yadan, O. Hydra - a framework for elegantly configuring complex applications. Github, 2019. URL https://github.com/facebookresearch/hydra.

Zaballa, V. D. and Hui, E. E. Optimal Design of Experiments for Simulation-Based Inference of Mechanistic Acyclic Biological Networks. nov 2021. doi: 10.48550/arxiv.2111.13612. URL https://arxiv.org/abs/2111.13612v1.

Zheng, S., Pacheco, J., and Fisher, J. W. A robust approach to sequential information theoretic planning. In *35th International Conference on Machine Learning, ICML 2018*, volume 13, pp. 9476–9489. PMLR, July 2018. ISBN 9781510867963. URL http://proceedings.mlr.press/v80/zheng18b.html.

A. Algorithmic implementation

Algorithm 1 Bayesian Model Averaging SBIDOEMAN

1: **Require:** Simulators $f_i(\boldsymbol{d}, \boldsymbol{\theta})$, held-out true parameters $\boldsymbol{\theta_T}$, true simulator f_T, number of simulations for MINEBED N_M, number of acquisitions for Bayesian optimization N_A, number of simulations per LFI round N_S, number of LFI rounds N_R, number of experiments N_E, neural density estimator $q_\phi(\boldsymbol{x}|\boldsymbol{\theta})$, priors over models' parameters $p(\boldsymbol{\theta}_i)$, and models' prior probabilities $p(\mathcal{M}_i)$.

2: **Return:** Models' approximate posterior $p(\boldsymbol{\theta}_i|\boldsymbol{x_o}, \boldsymbol{d}, \mathcal{M}_i)$, models' marginal probabilities $p(\mathcal{M}_i|\boldsymbol{x_o}, \boldsymbol{\theta}_i, \boldsymbol{d})$, and Bayes Factor $BF = p(\mathcal{M}_1)/p(\mathcal{M}_0)$.

3: Initialize a design $\boldsymbol{d_0}$ by random sampling and set $\boldsymbol{d}^* = \boldsymbol{d_0}$

4: Initialize N MINE neural network parameters $\boldsymbol{\psi}_0, \ldots, \boldsymbol{\psi}_N$ where $N = |\mathcal{M}|$

5: Set proposals $\tilde{p}^{\{i\}}(\boldsymbol{\theta}) := p^{\{i\}}(\boldsymbol{\theta})$ for $\mathcal{M}_i \in \{\mathcal{M}\}_{i=1}^N$

6: **for** $j = 1 : N_E$ **do**

7:     **for** $\mathcal{M}_i \in \{\mathcal{M}\}_{i=1}^N$ **do**

8:         **for** $k = 1 : N_A$ **do**

9:             **for** $l = 1 : N_M$ **do**

10:                 $\boldsymbol{\theta}_{k,l}^{\{i\}} \sim \tilde{p}_{k,l}^{\{i\}}(\boldsymbol{\theta})$

11:                 Simulate $\boldsymbol{x}_{k,l}^{\{i\}} \sim f_i(\boldsymbol{d}, \boldsymbol{\theta}_{k,l}^{\{i\}})$

12:                 Optimize MINE parameters ψ_i between simulated data and priors for the model by maximizing the mutual information lower bound $\widehat{I}(\boldsymbol{d}, \psi_i^*)$

13:                 $\widehat{I}(\boldsymbol{d}, \psi^*) = \frac{1}{N} \sum \widehat{I}(\boldsymbol{d}, \psi_i^*)$

14:                 $\boldsymbol{d}^* = \boldsymbol{d}$ if $\widehat{I}(\boldsymbol{d}, \psi^*) > \widehat{I}(\boldsymbol{d}^*, \psi^*)$

15:             **end for**

16:         **end for**

17:     **end for**

18:     Observe simulated experimental condition $\boldsymbol{x_o} = f_T(\boldsymbol{d}^*, \boldsymbol{\theta}_T)$

19:     **for** $\mathcal{M}_i \in \{\mathcal{M}\}_{i=1}^N$ **do**

20:         **for** $k = 1 : N_R$ **do**

21:             **for** $l = 1 : N_S$ **do**

22:                 $\boldsymbol{\theta}_{k,l} \sim \tilde{p}_k^{\{i\}}(\boldsymbol{\theta})$

23:                 Simulate $\boldsymbol{x}_{k,l} \sim f_i(\boldsymbol{x}, \boldsymbol{\theta}_{k,l})$

24:             **end for**

25:             (re-)train $q_\phi^{\{i\}} \leftarrow \underset{\phi}{\mathrm{argmin}} - \frac{1}{N} \sum_{(\boldsymbol{x}_{k,l}^{\{i\}}, \boldsymbol{\theta}_{k,l}^{\{i\}})}^j \log q_\phi^{\{i\}}(\boldsymbol{x}_{k,l}^{\{i\}}|\boldsymbol{\theta}_{k,l}^{\{i\}})$

26:             $\tilde{p}_{k+1}^{\{i\}}(\boldsymbol{\theta}|\boldsymbol{x_o}) \propto p^{\{i\}}(\boldsymbol{x_o}|\boldsymbol{\theta})p_k^{\{i\}}(\boldsymbol{\theta}) \approx q_\phi^{\{i\}}(\boldsymbol{x_o}|\boldsymbol{\theta})p_k^{\{i\}}(\boldsymbol{\theta})$

27:             (re-)train $q_\tau^{\{i\}} \leftarrow \underset{\tau}{\mathrm{argmin}} D_{KL}(q_\tau^{\{i\}}(\boldsymbol{\theta})||\tilde{p}_{k+1}^{\{i\}}(\boldsymbol{\theta}|\boldsymbol{x_0}))$

28:             Set $\tilde{p}_{k+1}^{\{i\}}(\boldsymbol{\theta}) := q_\tau^{\{i\}}(\boldsymbol{\theta})$

29:         **end for**

30:         train $p_x(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{x_o}, \mathcal{M}_i) \leftarrow \underset{\zeta}{\mathrm{argmin}} - \frac{1}{N} \sum_{(\boldsymbol{x}_l, \boldsymbol{\theta}_l)}^j \log p_u(\boldsymbol{f}^{-1}(\boldsymbol{x_o}; \zeta) + \log|\det \boldsymbol{J}(\boldsymbol{f}^{-1})(\boldsymbol{x_o}; \zeta)|$ where $p_u(\boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1})$

31:         $p(\mathcal{M}_i|\boldsymbol{x_o}, \boldsymbol{\theta}, \boldsymbol{d}) = 1 - \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\boldsymbol{f}^{-1}(\boldsymbol{x_o}; \zeta)}{\sqrt{2}}\right)\right)$

32:     **end for**

33:     $BF_j = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}$

34: **end for**

For the choice of hyperparameters, we used $N_M = 5000$, $N_A = 5$, $N_S = 1000$, $N_R = 5$, $N_E = 5$, a SNLE $q_\phi(x|\theta)$ density estimator, starting box uniform priors for $p(\theta)$, and uniform priors for $p(\mathcal{M}_i)$. We evaluated 50 simulations at a time limit of 10 hours. For the one-step model, the random choice had 14 simulations finish, equidistant had 26 simulations finish, and SBIDOEMAN BMA had 15 simulations finish. For the two-step model, random choice had 21 simulations finsih, equidistant had 25 finish, and SBIDOEMAN BMA had 16 finish.

B. Proof of Proposition 2.1

We prove Proposition 2.1 using the following definitions.

Definition B.1. The marginal likelihood is the measure of evidence given a model and is defined as

$$p(\boldsymbol{x_o}|\mathcal{M}) = \int p(\boldsymbol{x_o}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}$$

where the parameters $\boldsymbol{\theta}$ of the model $\mathcal{M}$ are integrated out. If the marginal likelihood is multiplied by the model prior $p(\mathcal{M})$ it then becomes model marginal probability $p(\mathcal{M}|\boldsymbol{x_o})$.

Definition B.2. A normalizing flow is defined as a series of diffeomorphic functions, $\boldsymbol{f}$, that map a base distribution, $p_{\boldsymbol{u}}(\boldsymbol{u})$ to a complex data distribution, $p_{\boldsymbol{x}}(\boldsymbol{x})$. Equation (5) shows the transformation from a base distribution to data distribution. The inverse function $\boldsymbol{f}^{-1}$ will transform observed data $\boldsymbol{x}$ to the base distribution as $\boldsymbol{f}^{-1}(\boldsymbol{x}) = \boldsymbol{u}$, which could be a Gaussian distribution. When training a normalizing flow with fixed base distribution parameters, this can also be seen as a generalized version of the reparameterization trick and a form of variational inference (Rezende & Mohamed, 2015).

Definition B.3. The Cumulative Distribution Function (CDF) of a real-valued random variable X, if X is distributed according to a Gaussian distribution, is defined as

$$\begin{aligned}
F_X(x) &= P(X \le x) \\
&= \Phi(x) \\
&= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du \\
&= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right)
\end{aligned}$$

where $\operatorname{erf}()$ is the error function. A derivation of the Gaussian CDF can be found in Bishop (2006).

Definition B.4. The CDF gives a probability of points smaller than a point but we want to know how many points are larger than the point in the Gaussian distribution, $P(X > x) = 1 - F_X(x)$. This is the complementary CDF, or tail distribution, and is helpful in the context of normalizing flows, where we are interested in determining the probability of simulated point from our model $\boldsymbol{x} \sim p_x(\boldsymbol{x}|\boldsymbol{x_o}, \boldsymbol{\theta}, \mathcal{M})$ being greater than the observed data point $\boldsymbol{x_o}$. Thus, $p(\boldsymbol{x_o}|\mathcal{M}, \boldsymbol{\theta}) \equiv 1 - F_X(\boldsymbol{x_o})$.

Using these definitions, we prove Proposition 2.1 as follows.

Proof. Starting from the definition of the marginal likelihood in Definition B.1, we can approximate the intractable likelihood $p(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M})$ using a normalizing flow trained by sampling $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{1})$, as by variational inference in Definition B.2, and model parameters sampled $\theta \sim q_\tau(\theta)$ where $q_\tau(\theta)$ is the model's posterior given observed data x_o. Thus the marginal likelkihood can be approximated as follows:

$$\begin{aligned}
p(\boldsymbol{x}|\mathcal{M}) &= \int p(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}, \mathcal{M})d\boldsymbol{\theta} && \text{By Definition B.1} \\
&\approx p_x(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M}) && \text{By Definition B.2 and } \boldsymbol{\theta} \sim q_\tau(\boldsymbol{\theta}|\boldsymbol{x_o}, \mathcal{M}).
\end{aligned}$$

The last step can also be seen as a form of variational inference that employs the reparameterization trick for a simulator $f_\mathcal{M}$ as $\boldsymbol{x} \sim p_x(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M}, \boldsymbol{x_o}) \iff \boldsymbol{x} = f_\mathcal{M}(\boldsymbol{d}, \boldsymbol{\theta}), \boldsymbol{\theta} \sim q_\tau(\boldsymbol{\theta}|\boldsymbol{x_o}, \mathcal{M})$, assuming that designs $\boldsymbol{d}$ are known and using the posterior inference of model parameters given a certain observed data point $\boldsymbol{x_o}$.

Using this approximation of the marginal probability distribution, we can find the probability that a point $\boldsymbol{x_o}$ is within this probability density function by converting to a Gaussian base distribution, $p_u(\boldsymbol{u})$, via variational inference. We find models'

marginal probabilities as follows:

$$
\begin{aligned}
p(\boldsymbol{x_o}|\mathcal{M}, \boldsymbol{\theta}) &\equiv 1 - F_X(\boldsymbol{x_o}) && \text{By Definition B.4} \\
&= 1 - \int_{-\infty}^{\boldsymbol{x_o}} p_x(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M}, \boldsymbol{d})d\boldsymbol{x} && \\
&= 1 - \int_{-\infty}^{\boldsymbol{x_o}} p_u(\boldsymbol{f}^{-1}(\boldsymbol{x}))|\det \boldsymbol{J}(\boldsymbol{f}^{-1})(\boldsymbol{x})|d\boldsymbol{u} && \text{By Definition B.2} \\
&\propto 1 - \int_{-\infty}^{\boldsymbol{x_o}} p_u(\boldsymbol{f}^{-1}(\boldsymbol{x}))d\boldsymbol{u} && \\
&= 1 - \Phi(\boldsymbol{f}^{-1}(\boldsymbol{x_o})) && \text{By Definition B.3} \\
&= 1 - \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\boldsymbol{f}^{-1}(\boldsymbol{x_o})}{\sqrt{2}}\right)\right) && \text{By Definition B.3.}
\end{aligned}
$$

We note that this data likelihood becomes the model marginal probability when multiplied by the model's prior, $p(\mathcal{M})$. $\square$

Remark B.5. While we achieved acceptable results by simplifying the change of volume from the data distribution $p_x(x)$ to the base distribution $p_u(u)$ in the third step, it remains to be seen whether there is improved performance in calculating models' marginal probabilities with this value.