Systematic Bayesian Posterior Analysis Facilitates Hypothesis Formation and Guides Investigation of Pancreatic Beta Cell Signaling

Holly A. Huber[a], Senta K. Georgia[b], and Stacey D. Finley[a,c,d]

[a]Department of Biomedical Engineering, University of Southern California, Los Angeles, CA 90089, USA

[b]Department of Stem Cell Biology and Regenerative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.

[c]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA.

[d]Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, CA 90089, USA.

Corresponding Author: Stacey D. Finley, email - sfinley@usc.edu

Keywords: Bayesian inference, Kullback-Leibler divergence, posterior analysis, Suppressor of Cytokine Signaling, pancreatic beta cell, mathematical model, signal transduction

Abstract

Bayesian inference produces a posterior distribution for the parameters and predictions from a mathematical model that can be used to guide the formation of hypotheses, thereby increasing our understanding of a biological process. Previous approaches to such Bayesian hypothesis formation are largely qualitative and time consuming; further, demonstrations of these approaches typically stop at hypothesis formation, leaving the questions they raise unanswered. Here, we introduce a Kullback-Leibler (KL) divergence-based ranking to expedite Bayesian hypothesis formation and investigate the hypotheses it generates, ultimately generating novel, therapeutically-relevant biological insights. We test our approach on a computational model of prolactin-induced JAK2-STAT5 signaling, a pathway that mediates beta cell proliferation. Because diabetes is characterized by a decrease in functional beta cell mass, replenishing this mass by harnessing beta cell proliferation signaling pathways like this one is a viable potential therapeutic strategy. Our KL divergence-based approach selects only a subset of data-informed parameters to search for qualitative evidence of alternative hypotheses, thereby expediting hypothesis formation. Within this subset, we find two plausible ranges for the receptor degradation rate, each characterized by different dynamics of the SOCS protein, which mediates a negative feedback loop in the JAK2-STAT5 pathway. To investigate these dynamics, we refine the posterior of an existing model with additional data. This refined model out-performs the original model when predicting a test data set consisting of protein dynamics that are consistent across ligand and tissue types. Further, it predicts transient SOCS activation with high certainty. We demonstrate that our approach offers a novel quantitative framework for Bayesian hypothesis testing. Further, we establish an improved model of beta cell signaling and determine that the negative feedback controlling the prolactin-induced proliferative response in beta cells is likely transient. This understanding of SOCS dynamics is particularly relevant for harnessing the proliferation signaling response of beta cells.

1. Introduction

The formation and testing of hypotheses is perhaps the primary endeavor of science. Hypotheses emerge by considering uncertainty in scientific study— if we are completely certain in our knowledge, there is no need to form alternative explanatory hypotheses. Thus, uncertainty is necessary to motivate hypothesis formation. Well-designed experimental studies allow one to formally test hypotheses. For complex biological phenomena, such as intracellular protein signaling networks, mathematical models can produce more quantitative, systems-level insight than experiments alone. Such mathematical models must have accurate parameter values in order to generate reliable predictions. However, common methods of estimating model parameter values may not fully capture uncertainty, and thus overlook alternative hypotheses (Hug et al., 2013; Tötsch and Hoffmann, 2020). Bayesian statistics address this limitation by enabling the mathematical characterization of uncertainty in model parameters and subsequent predictions.

Applied to parameter estimation, Bayesian statistics define the distribution of viable parameters given the available experimental data. Formally, this is known as the *posterior distribution*, the distribution of model parameters conditioned on the data. As informative data are added, the set of viable model parameters changes, and the posterior is updated accordingly. For example, new data might increase our certainty about model parameters and thus cause a decrease in the variance of the posterior distribution. On the other hand, conflicting data might decrease certainty about model parameters, and thus increase the variance of the posterior distribution. If approximated correctly, a posterior can reveal uncertainty in the estimated parameter values, and subsequently, uncertainty in the model predictions. Once this model uncertainty is established, the posterior may be used to formulate new hypotheses.

Indeed, previous studies have demonstrated the utility of posteriors for hypothesis formation based on mechanistic models of protein network dynamics. For example, Bayesian data integration has been used to reveal inconsistencies in a classic study on protease self-digest kinetics (Tötsch and Hoffmann, 2020). Bayesian data integration is the process of integrating subsets of data into the posterior inference and observing whether different data constrain the model in contradictory ways (Thijssen et al., 2018). Using this process, the authors were able to form an improved explanatory hypothesis for the observed protease dynamics. In another example, qualitative analyses of the Bayesian posterior revealed alternative mechanisms explaining red blood cell signaling (Hug et al., 2013). To do this, the authors first searched each parameter's posterior for bimodalities. This search revealed two alternative reaction rates that both explained the observed red blood cell signaling response. These

alternative rates were then used to characterize testable hypotheses of protein signaling dynamics.

Though both studies successfully demonstrated the utility of posteriors in hypothesis formation, their methods and scope had some limitations. Both approaches—Bayesian data integration and searching for bimodality in the posterior—rely on largely qualitative posterior analyses of all the model parameters, which limits their scalability (Hug et al., 2013; Thijssen et al., 2018). Furthermore, both papers' scopes were limited to hypothesis formation: the proposed hypotheses remained uninvestigated (Hug et al., 2013; Tötsch and Hoffmann, 2020). Here, we present a systematic posterior analysis approach that facilitates hypothesis formation and, subsequently, investigate the alternative hypotheses it reveals. Our approach uses Kullback-Leibler (KL) divergence to rank parameters by how informed they are by the data, then applies K-means clustering to this ranking to reveal intrinsic subgroups. Subsequently, we select only the most data-informed subgroup for qualitative analysis. Previous work has leveraged KL divergence as a metric of data utility to optimize experimental design (Chaloner and Verdinelli, 1995; Lomeli et al., 2021). Here, by using it to select only a subset of parameters for qualitative analysis, we expedite hypothesis formation. As it is the most data-informed subset, we also prioritize formation of *data-informed* hypotheses in our search.

We demonstrate our approach on a biologically relevant model of pancreatic beta cell signaling. Because a decrease in beta cell mass is a hallmark of Type 2 diabetes, previous research has set out to understand signaling pathways that govern beta cell proliferation to design novel approaches to expand functional beta cell mass (Aguayo-Mazzucato and Bonner-Weir, 2018; Brelje et al., 2004; Fujinaka et al., 2007; Jiao et al., 2013; Millette et al., 2022; Rieck et al., 2009; Rønn et al., 2002; Salazar-Petres and Sferruzzi-Perri, 2022). Pregnancy has been a key model system to understand beta cell proliferation. Observations from both human and rodent pregnancy show that beta cell mass and insulin secretion increase to compensate for increased insulin resistance. These studies revealed prolactin (PRL)-induced janus kinase 2 (JAK2) and signal transducer and activator of transcription 5 (STAT5) signaling as a key regulator of beta cell proliferation (Brelje et al., 2004, 2002). Specifically, the JAK2-STAT5 pathway regulates proliferation by controlling the expression of several proteins. Some of these proteins, such as anti-apoptotic protein B-cell lymphoma-extra large (BCL-xL), influence cell phenotype, while others, such as inhibitor protein Suppressor of Cytokine Signaling (SOCS), form feedback loops with proteins upstream in the pathway to control cell signaling (Fujinaka et al., 2007; Rieck et al., 2009). This is the biological system we model in the present study.

In this work, we model PRL-induced JAK2-STAT5 signaling to (1) test our hypothesis formation approach on a biologically-relevant mathematical model and (2) better understand how to enhance beta cell proliferation. To begin, we improve the posterior inference of our previously established model (Mortlock et al., 2021). Subsequently, we analyze the inferred posterior with our KL divergence-based approach, uncovering two plausible receptor degradation rates that both explain the observed SOCS signaling dynamics controlling beta cell proliferation. We then investigate the proposed hypotheses by refining the posterior estimate with new data and validating with literature evidence. Finally, we apply the improved model to resolve the proposed hypotheses, thereby answering an open question in pancreatic beta cell biology: *what are the SOCS negative feedback dynamics controlling PRL-induced JAK2-STAT5 signaling that drive beta cell proliferation?*

2. Materials and Methods

2.1 Modeling Framework

We modeled PRL-induced intracellular signaling in rodent pancreatic beta cells (**Figure 1**). During pregnancy, PRL induces beta cell proliferation in rodents via the JAK2-STAT5 signaling cascade (Brelje et al., 2004, 2002). Specifically, PRL binds to the prolactin receptor (PRLR), inducing subsequent JAK2 and STAT5 activation (phosphorylation). After activation, STAT5 translocates into the nucleus, where it regulates the transcription of various gene products. STAT5 mediates transcription of proteins that provide positive and negative feedback to the signaling pathway, PRLR and the SOCS inhibitor family, respectively (Rieck et al., 2009). STAT5 also promotes transcription of proteins that characterize a proliferative cell phenotype, such as the anti-apoptotic protein BCL-xL (Fujinaka et al., 2007). Finally, it is important to note that two different isoforms of STAT5, STAT5A and STAT5B, are activated by this cascade. These proteins can form homo- and hetero-dimers, which follow different translocation patterns upon activation (Brelje et al., 2004, 2002).

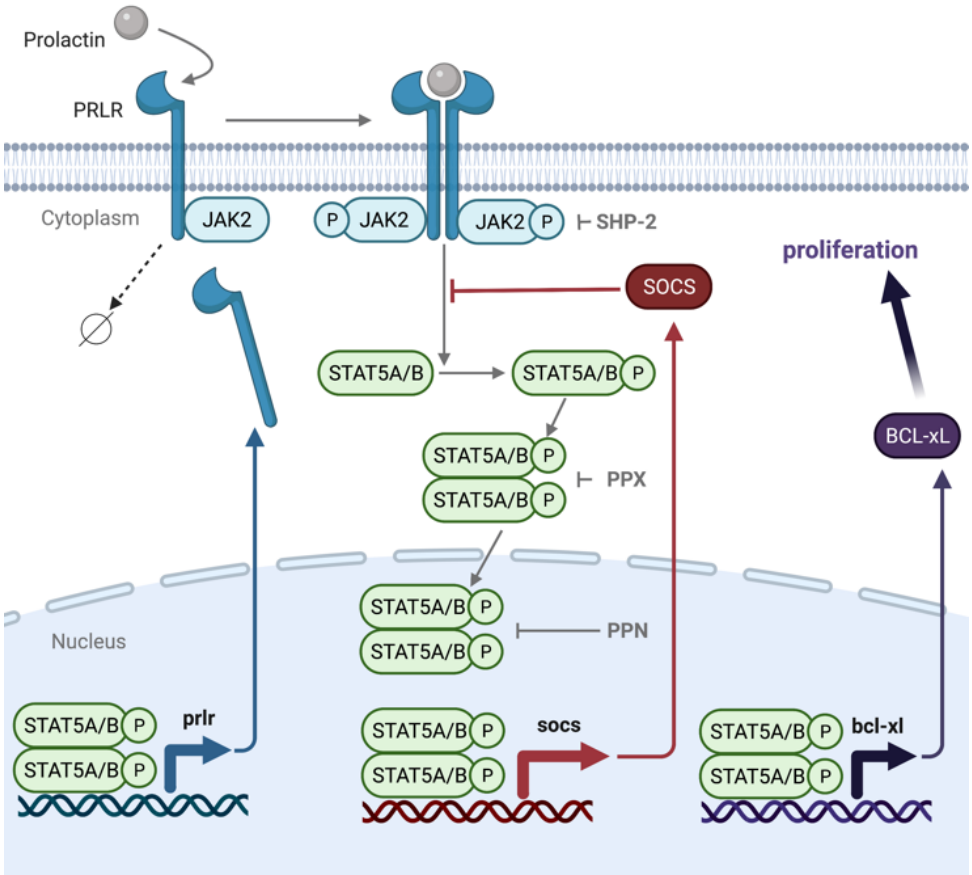

Figure 1. Schematic of the signaling cascade modeled in this work. PRL binds to the PRLR-JAK2 complex, leading to receptor dimerization and JAK2 activation. The activated complex promotes phosphorylation of STAT5A and STAT5B, which dimerize and can translocate to the nucleus. Nuclear STAT5A/B promotes transcription and translation of PRLR, BcL-xL, and SOCS. Feedback from these proteins and the effect of phosphatases (PPX and PPN) are indicated.

To describe this signaling network, we use the ordinary differential equation (ODE) model structure developed by Mortlock et al. This model contains rate equations for each of the described signaling events. In addition, it accounts for degradation of the signaling species, including receptor degradation, as it influences the pathway dynamics. Specifically, the rate of receptor degradation depends on whether PRL is bound or not. Finally, this model also includes rate equations accounting for the effect of nuclear (PPN) and cytoplasmic (PPX) phosphatases observed in the canonical JAK2-STAT5 pathway (Yamada et al., 2003). In total, the model is comprised of 55 species and 60 kinetic parameters, 33 of which we estimate, as in Mortlock et al. These 33 parameters were selected for model fitting based on either sensitivity analysis results or to include all model submodules; Mortlock et al. includes further details. The rate equations are governed by mass action kinetics or Michaelis-Menten kinetics. This model was

most recently used to study the effect of heterogeneity in the intracellular protein concentrations (Simoni et al., 2022). Now, we use it as the basis for this work.

2.2 Training Data

Initially, we use the same training data as in Mortlock et al. to estimate our model parameters. Subsequently, we obtained additional data at extended timepoints and refine our parameter estimates using this data. The original training data for our model is comprised of western blot and immunohistochemical measurements of protein responses in INS-1 cells treated with 200 ng/ml of PRL. Experimental outputs include: (1) phosphorylated JAK2 (pJAK2) up to 6 hours post-PRL stimulation, (2) phosphorylated STAT5A (pSTAT5A) up to 6 hours post-PRL stimulation, (3) phosphorylated STAT5B (pSTAT5B) up to 6 hours post-PRL stimulation, (4) BCL-xL up to 24 hours post-PRL stimulation, (5) the ratio of nuclear to cytoplasmic STAT5A (N:C STAT5A) up to 6 hours post-PRL stimulation, and (6) the ratio of nuclear to cytoplasmic STAT5B (N:C STAT5B) up to 6 hours post-PRL stimulation.

The updated training data for our model includes all the original training data, plus three additional measurements: the relative levels of phosphorylated JAK2, phosphorylated STAT5A, and phosphorylated STAT5B 12 hours post-PRL stimulation. These new timepoints were provided by the authors of the original experimental work (Brelje et al., 2004, 2002). **Supplementary Figure S1** provides a comparison of the original training data to the updated training data.

2.3 Bayesian Posterior Inference

We use the same prior and likelihood distributions as Mortlock et al. to formulate our posterior distribution. **Supplementary Table S1** contains detailed definitions of our prior, likelihood, and proposal distributions. Briefly, the prior distribution is defined as lognormal, centered on the lowest-error parameters found for a model of JAK-STAT signaling in liver cells, with a variance of 2 (Yamada et al., 2003). This variance was assigned based on the variance of kinetic rates pooled across species and pathways (Bar-Even et al., 2011). The likelihood distribution is defined as normal, centered on the data mean with an inverse-gamma distributed variance. The shape of the inverse gamma distribution is defined by two hyperparameters, assigned based on a previously developed posterior estimation procedure (Ghasemi et al., 2011; Makaryan and Finley, 2020). Using these analytical forms for the prior and likelihood distribution, we can approximate the posterior distribution.

We improve upon the posterior approximation procedure from Mortlock et al. by implementing established metrics to monitor the approximation algorithm's convergence. Like Mortlock et al., we use the Metropolis-Hasting algorithm to estimate the posterior. This algorithm generates an aperiodic and irreducible Markov chain that satisfies the detailed balance condition; samples generated from this Markov chain asymptotically mimic samples drawn from the target distribution, here, the posterior distribution(Andrieu et al., 2003). However, while samples converge to the posterior distribution as the number of draws approaches infinity, a well-known challenge when implementing Markov chain methods is ensuring convergence to the posterior distribution with finite samples (Andrieu et al., 2003; Gelman, 2014).

To address this challenge, we compare samples from at least three independent Markov chains via established qualitative (trace plots, autocorrelation plots, and histograms) and quantitative (split potential scale reduction factor, $split - \hat{R}$) diagnostics (Cowles and Carlin, 1996; Gelman, 2014). Only after samples from these three chains are sufficiently similar ($split - \hat{R} < 1.1$) do we conclude convergence and begin analysis of the posterior distribution(Brooks and Gelman, 1998). More details regarding our posterior approximation procedure and the applicability of the $split - \hat{R}$ diagnostic to our posterior approximation problem can be found in the **Supplementary Materials, Section 1**.

2.4 KL Divergence Based Clustering for Prioritizing Data-Informed Parameters

We use Kullback-Leibler (KL) divergence to rank parameters by how informed they are by the data. KL divergence quantifies the difference between two distributions—a larger KL divergence corresponds to a larger difference. We use univariate kernel density estimates and Monte Carlo integration to approximate the KL divergence. Further details on this approximation are included in the **Supplementary Materials, Section 2**.

We calculate the KL divergence from the marginal prior distribution to the marginal posterior distribution. Since the prior is a distribution over the parameters while the posterior is a distribution over the parameters, given the experimental data, if there is a difference between the prior and posterior distribution, it can only be because the experimental data influences the posterior distribution. The more the data informs the posterior, the larger the difference between the prior and posterior, and the larger the KL divergence from the prior to the posterior. In this way, the KL divergence from the marginal prior distribution to the marginal posterior distribution provides a quantitative measure for ranking parameters by how informed they are by the data.

After ranking parameters based on the KL divergence, we apply K-means clustering, an unsupervised learning algorithm that partitions data into *k* disjoint sub-groups, or clusters. The

objective of this clustering is to group similar data points together and uncover underlying patterns. Further details on this objective are included in the **Supplementary Materials, Section 3**. We use scikit learn's *KMeans* function to estimate the optimal group assignment (Pedregosa et al., 2011). We parameterize the *KMeans* function with $k = 2$ clusters. We use 2 clusters because the benefit, that is, increasing similarity, of using more clusters becomes marginal after 2 (**Supplementary Figure S2**). Further details on how we chose 2 clusters is included in the **Supplementary Materials, Section 3**.

2.5 Constructing a Validation Metric from Published Experimental Studies

We use the dynamics of SOCS to test plausible biological hypotheses generated via our posterior analysis. We use SOCS as a test output because it provides negative feedback to the JAK-STAT pathway and thus plays a key role in controlling pathway dynamics. Ideally, we would test our model with measurements of the SOCS concentration in beta cells 0-24 hours following PRL stimulation. Although no direct measurements exist, we were able to find indirect, but nevertheless relevant, evidence pertaining to SOCS dynamics in the literature (Chong et al., 2001; Croker et al., 2008; Endo et al., 2003; Jiao et al., 2013; Ram and Waxman, 1999; Rieck et al., 2009). Subsequently, we synthesized this evidence, to create a validation metric.

First, we determined which member(s) of the SOCS family of proteins likely inhibit PRL-induced signaling in beta cells during pregnancy. To do this, we referenced both quantitative PCR (qPCR) data and luciferase assay data. The qPCR data compared the SOCS transcriptional response of rodents between pregnant and non-pregnant states. The data from two such studies confirms that of the eight SOCS proteins, only CISH and SOCS2 are upregulated during rodent pregnancy (**Supplementary Figure S3A and S3B**) (Jiao et al., 2013; Rieck et al., 2009). These results are confirmed by another study that investigated the binding interactions between CISH and PRLR, our pathway's receptor, via a yeast tri-hybrid system and luciferase assay (Endo et al., 2003). This study found that CISH binds to PRLR and inhibits its activity. Thus, at least CISH has been proven to bind with the receptor in the pathway being modeled here. Based on these three studies, we conclude that either CISH or SOCS2 inhibits PRL-induced JAK2-STAT5 signaling.

Next, we determined the most likely CISH/SOCS2 dynamics from 0-24 hours following ligand stimulation. Although we lacked measurements of this output for PRL-induced JAK2-STAT5 signaling in beta cells, we found consistent CISH/SOCS2 dynamics across other ligands and tissue types: transient SOCS activation, with an early peak (Chong et al., 2001; Ram and Waxman, 1999). Protein dynamics are not necessarily conserved between different ligand

and/or tissues (Miller et al., 2019; Nandagopal et al., 2018). However, this consistency across experimental conditions suggests that it is reasonable to assume these same dynamics describe SOCS behavior in our system. To strengthen this assumption, we specifically sought data from beta cells or ligands related to the hormone, PRL.

First, we found an experiment studying CISH, SOCS1, and SOCS2 in beta cells. In this study, beta cells were incubated with either IFN-γ, IFN-α, IL-1β, or TNF-α, and western blot measurements for each of the three SOCS proteins were recorded at 1, 2, 4, and 24 hours post-incubation (Chong et al., 2001). Strikingly, a reliable difference between SOCS1 dynamics and CISH/SOCS2 dynamics emerged across ligands. While SOCS1 exhibited sustained activation 24 hours following stimulation with no early peak, both SOCS2 and CISH exhibited transient activation 24 hours following incubation, with an early peak. **Supplementary Figure S3C** displays a subset of these results.

We next checked that this difference between SOCS1 and CISH/SOCS2 dynamics across different ligands is mechanistically plausible. We found that CISH and SOCS2 share mechanistic similarities, compared to other SOCS proteins like SOCS1. In general, SOCS proteins have three possible inhibition mechanisms (Croker et al., 2008; Ram and Waxman, 1999). Importantly, SOCS2 and CISH share the same mechanism: competitive inhibition with STAT via receptor phosphotyrosine binding, while SOCS1 does not (Croker et al., 2008; Ram and Waxman, 1999). Thus, it is plausible that CISH/SOCS2 would have similar dynamics due to this mechanistic similarity.

Finally, we sought an experiment confirming these consistent dynamics for a ligand more closely related to PRL. We examined published results for CISH dynamics following stimulation with Growth Hormone (GH). Because both GH and PRL are hormones, and GH activates JAK2-STAT signaling, GH-induced SOCS dynamics provide a biologically comparable system to PRL-induced JAK2-STAT5 signaling. In the study, rats were treated with GH and sacrificed 0.5, 1, 1.5, 4, and 24 hours post stimulation, and the livers were harvested for analysis. CISH mRNA expression was measured at each of these timepoints. Again, CISH displayed transient activation by 24 hours following ligand stimulation with an early peak (Ram and Waxman, 1999).

Thus, based on the consistent CISH/SOCS2 dynamics observed across tissues and ligands related to our system of interest, we designate the following as our target output for model validation: *transient SOCS activation 24 hours following stimulation with PRL, with an early peak.*

2.6 SOCS Shape Classification for Validation Comparison

To compare our model predictions with the target output, we introduced two shape classification criteria that categorized the predicted SOCS dynamics as: (1) displaying an early peak or no early peak and (2) transient or sustained. We apply these two criteria to the distribution of SOCS predictions generated by simulating the model with the parameter sets sampled from the posterior. In total, there are four possible SOCS shapes: no peak and transient activation, early peak and transient activation, no peak and sustained activation, and early peak and sustained activation. **Supplementary Figure S4** displays examples of each of these four shapes, as well as a graphical representation of our classification criterion.

To categorize SOCS time courses as early peak or no early peak, we applied SciPy's *find_peaks* function to the first five hours of SOCS dynamics (Pedregosa et al., 2011). The 'early' window designation of 0-5 hours was selected based on the CISH/SOCS2 western blot measurements (Chong et al., 2001) (**Supplementary Figure S3C).** To ensure our results were robust to changes in this window, we also applied the *find_peaks* function to predictions from 0-3 hours. Changing the window did not change the results (**Supplementary Figure S5).**

Next, to categorize SOCS time courses as transient or sustained, we applied a threshold to the relative SOCS concentration 24 hours following ligand stimulation. A prediction was transient if the SOCS concentration was less than 50% of the maximum SOCS concentration 24 hours post ligand stimulation. We chose the 24 hour time point to be consistent with the western blot data sets, which collected measurements 24 hours post-ligand stimulation (Chong et al., 2001; Ram and Waxman, 1999). The 50% cutoff was derived based on the relative CISH/SOCS2 concentrations collected in the same data sets. To ensure our results were robust to changes in this cutoff, we varied the threshold from 30% to 70%. Changing the threshold did not change the conclusions drawn from the results (**Supplementary Figure S5).**

2.7 Computational Implementation

All code for model analysis is available in the GitHub repository: https://github.com/FinleyLabUSC/KL-Divergence-Bayesian-Hypothesis-Formation.

3. Results

3.1 Calibrated mathematical model recapitulates experimental data

We established a model of pancreatic beta cell signaling dynamics that is consistent with the experimental data. To begin, the model parameters were fitted via Bayesian inference. We then confirmed that the Bayesian inference method converged to the posterior distribution

based on our diagnostic measures (**Supplementary Figure S6A and S7A**). Next, by sampling parameter sets from the posterior distribution, we generated 10,000 predictions for the signaling species' dynamics. These model predictions were subsequently verified against the training data set (**Figure 2**). The majority of training data fall within the 5th to 95th percentile bounds of the model predictions. The only consistent exception is the 4-hour time-point for pJAK2, pSTAT5A, pSTAT5B, and N:C STAT5B. Additions to the model could be considered to better recapitulate this time-point; however, the predicted 5th to 95th percentile bounds include more experimental data points than the previous model fit's bounds (Mortlock et al., 2021). Thus, we concluded that the refined model posterior adequately explained experimental data used for parameter estimation.

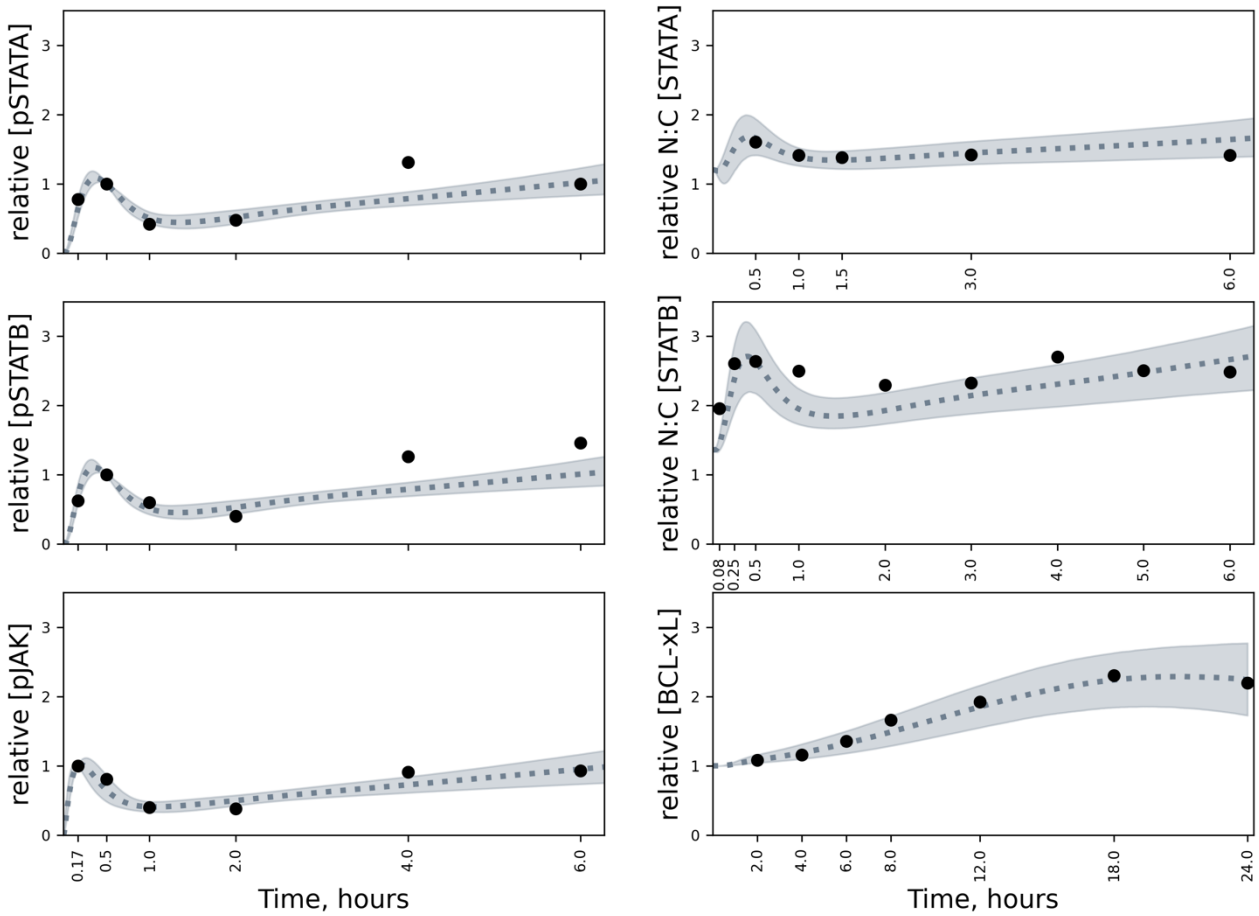

Figure 1. Comparison of inferred model predictions to training data. Predictions were generated by simulating the model with parameter samples from the posterior distribution. Dashed line, median prediction; shading, the 5th to 95th percentiles of predictions; black dots, experimental data mean.

3.2 Systematic Posterior analysis identifies alternative biological hypotheses

Analyses of our inferred posterior distribution identified two plausible PRL:PRLR:SOCS degradation mechanisms driving the signaling response. To quantitatively identify such mechanisms, we first used the KL divergence from the marginal prior distribution to the marginal posterior distribution for each model parameter to rank parameters by how informed they are by the experimental data (**Figure 3A**). We then used K-means to cluster these rankings into groups to determine any underlying patterns within the divergences. Our data was best described by two clusters (**Supplementary Figure S2**). Interestingly, the three parameters comprising the cluster most informed by the data are all related to PRLR degradation— the degradation rate of PRLR mRNA (*k27*), the degradation rate of the PRL:PRLR complex (*deg_ratio*), and the degradation rate of the PRL:PRLR:SOCS complex (*k24*). The SOCS protein is an inhibitor that provides negative feedback on the signaling pathway. Therefore, when the receptor is bound by SOCS, as in the PRL:PRLR:SOCS complex, PRLR is inactive.

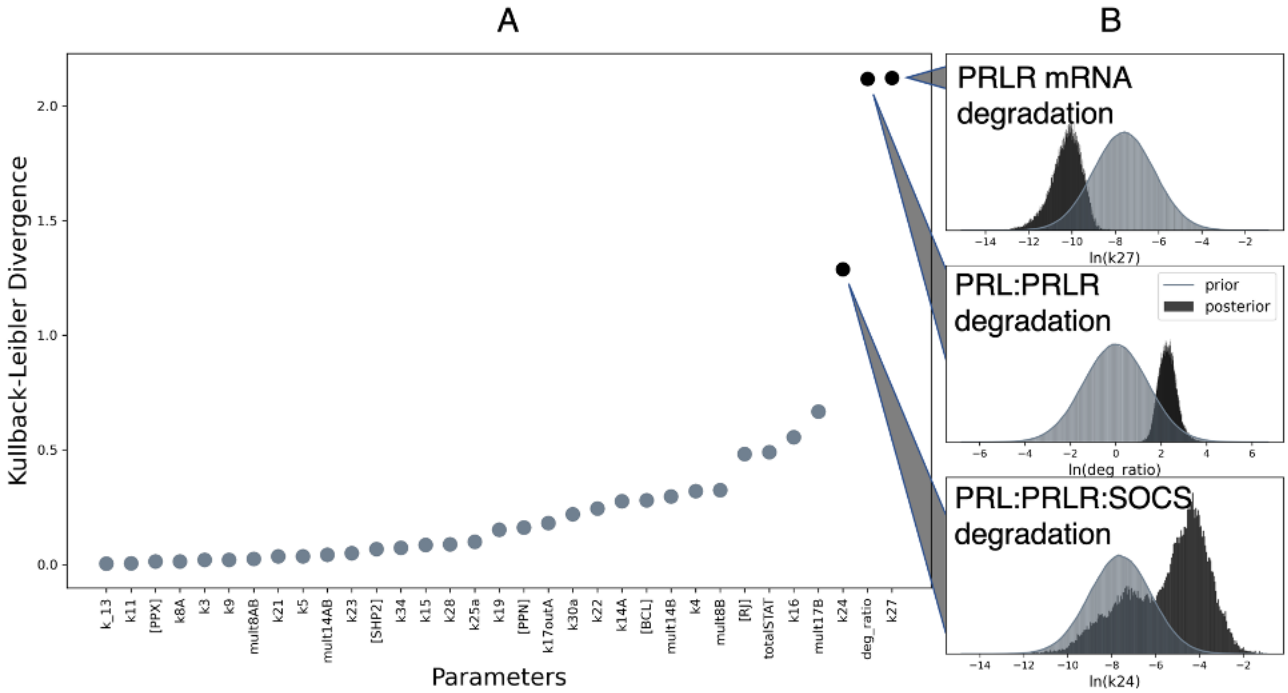

Figure 2. Posterior Analyses. (A) Estimated KL divergences from each parameter's marginal prior distribution to its marginal posterior distribution with significance cutoff. Grey dots, divergences assigned to cluster 1; black dots, divergences assigned to cluster 2. (B) Marginal prior and posterior distributions for parameters with the largest divergence, displayed on a natural-log scale. Grey, prior; black, posterior.

We next searched for qualitative evidence of divergent mechanisms in the marginal distributions of these top three, data-informed parameters (**Figure 3B**). As one would expect, all

three marginal posteriors are observably different from their respective marginal prior distributions. However, only the marginal PRL:PRLR:SOCS complex degradation rate is conceivably bimodal. This bimodality indicates that there are two plausible regimes for the rate at which the SOCS-bound receptor complex is degraded. In summary, using the KL divergence to rank parameters by how informed they are by the data and then searching the most data-informed parameters for bimodalities in their marginal posterior distributions revealed two SOCS-bound receptor degradation regimes that are both plausible given the training data.

Once we uncovered alternative PRL:PRLR:SOCS degradation regimes, we sought to determine if the two possible degradation regimes lead to distinct model predictions. To do this, we first defined two groups of parameter sets from the posterior distribution samples. One group contained 3000 parameter sets with slower PRL:PRLR:SOCS degradation values (**Figure 4A**, blue), while the other group contained 3000 parameter sets with faster PRL:PRLR:SOCS degradation values (**Figure 4A**, orange). We then simulated the model with the parameter sets from either group, comparing the predicted dynamics of the six quantities for which there is experimental data (**Figure 2**). We also considered one additional output, the SOCS protein, given its role in JAK-STAT signaling (see *Methods* section for more detail) and involvement in the receptor degradation rate. Interestingly, there was no clear difference between the predictions for the six quantities that have been measured experimentally due to sampling from the two groups of values for PRL:PRLR:SOCS degradation rate (**Supplementary Figure S8**). However, two clearly distinct SOCS predictions emerged. The predicted SOCS dynamics associated with the slower receptor complex degradation exhibit little to no early peak and transient activation 24 hours after PRL stimulation (**Figure 4B**, blue). On the other hand, the SOCS predictions associated with faster receptor complex degradation exhibit a significant early peak and sustained activation 24 hours after PRL stimulation (**Figure 4B**, orange). Overall, we connected distinct SOCS signaling dynamics to the two plausible ranges of PRL:PRLR:SOCS degradation, thereby producing two hypotheses as to the dynamics of the SOCS protein.

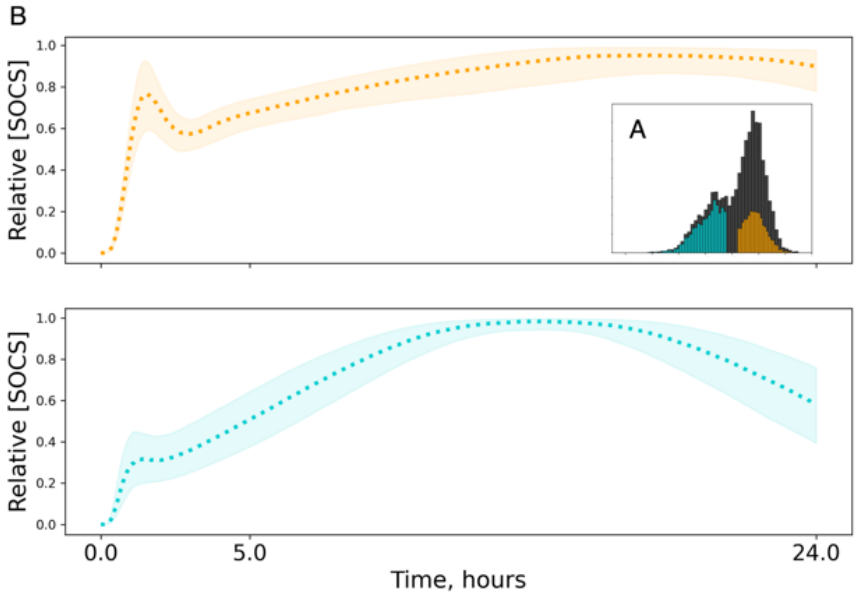

Figure 3. **Predicted SOCS signaling responses associated with the two parameter regimes for the rate of degradation of the PRL:PRLR:SOCS complex.** (A) The posterior distribution indicates either a fast or slow PRL:PRLR:SOCS degradation rate. (B) Simulations produced by sampling from the two parameter regimes resulted in distinct SOCS dynamics in both the short term (0-5 hours) and the long term (up to 24 hours). Dashed line, median prediction; shaded region, inter-quartile range. Orange, predictions sampled from faster degradation regime; blue, predictions sampled from slower degradation regime.

Sampling from the posterior distribution produced a range of distinct predictions for the dynamics of the SOCS protein. Thus, we used the SOCS dynamics to characterize the model predictions. In addition to the two SOCS behaviors described above, the model predicts two more possible behaviors of the SOCS proteins – no early peak with sustained signaling and early peak with transient signaling. Thus, there are a total of four possible SOCS profiles. We used shape classification (see Methods) to determine the predicted frequency of each of these profiles. We applied this classification criterion to 10,000 predictions, generated by randomly sampling 10,000 parameter sets from the posterior. The resulting frequencies are displayed in the left panel of **Figure 5A**. An early peak in SOCS concentration followed by sustained activation emerges as the most probable SOCS profile; this behavior is illustrated in the right panel of **Figure 5A**. We next sought to test the validity of these hypothesized dynamics predicted by the model.

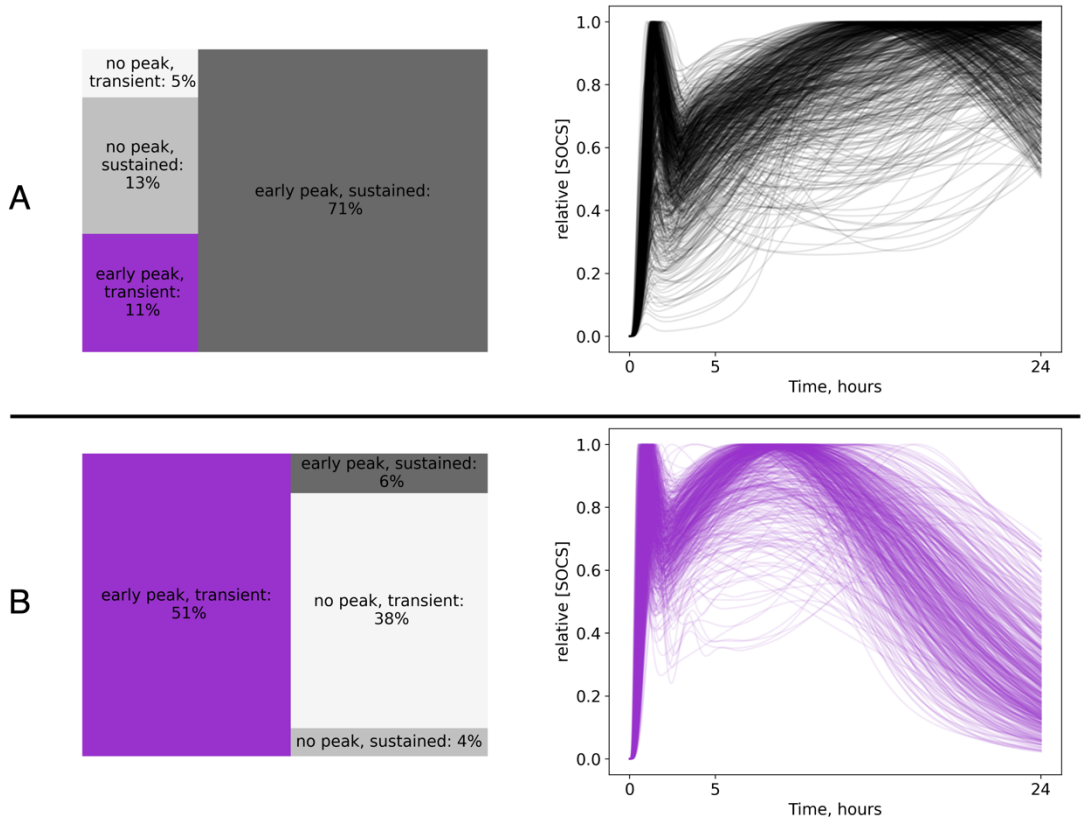

Figure 4. Comparison of the frequency of predicted SOCS dynamics. We determined how many predictions of SOCS dynamics are classified as no peak and transient activation (light gray), early peak and transient activation (dark gray), no peak and sustained activation (medium gray), and early peak and sustained activation (purple; target shape), before and after data integration (left). The most probable SOCS dynamics for each case are also displayed (right). (A) Posterior inferred with the baseline data; (B) Posterior inferred with the extended data.

3.3 Refining posterior produces better predictions and enables hypothesis testing

To confirm or reject our hypothesized SOCS profiles, we first conducted a literature search investigating SOCS dynamics in JAK2-STAT5 pancreatic beta cell signaling. Although there is significant interest in the topic, there exists no direct evidence of SOCS dynamics for our system of interest. However, our literature search did suggest that SOCS likely displays an early peak in its concentration followed by transient activation 24 hours post-stimulation with PRL (**Supplementary Figure S3**). We designated this as the target shape for the next phase of our investigation: model validation. More details on how we deduced this target shape are given in the Methods.

Conducting model validation, we found that our original model is inconsistent with the target shape. Our original model predicted sustained SOCS activation as the most probable behavior, while our validation metric indicates transient SOCS activation is more likely. Thus, we

next we sought to establish a model that better captured the target SOCS behavior found in the literature. To do this, we aimed to better constrain the long-term SOCS dynamics, which requires experimental data at later timepoints. Although we could not find direct measurements of SOCS dynamics following PRL simulation in pancreatic beta cells, we were able to uncover new observations of the JAK2 and STAT5 proteins at later timepoints (personal communication with R.L. Sorenson). We incorporated these new data into our Bayesian inference procedure. We again ensured that (1) our posterior inference algorithm converged via convergence diagnostics (**Supplementary Figure S6B and S7B**) and (2) the updated model predictions recapitulated the training data (**Supplementary Figure S9**). We then validated our updated model against the target shape (**Figure 5B**).

The updated model predictions for SOCS are significantly different from the original predictions. Before integrating the new data for later timepoints, the most probable SOCS dynamics exhibit sustained activation with an early peak, and the literature-confirmed SOCS behavior of transient activation at 24 hours with an early peak accounts for only 11% of the simulations (**Figure 5A**, left). After data integration, more than 50% of the predicted SOCS dynamics exhibit the target SOCS behaviors—transient activation with an early peak (**Figure 5B**, left). Thus, the updated predictions agree with the target behavior better than the original predictions.

We find that when additional measurements for signaling proteins at later timepoints are used to inform the model parameters, different long-term SOCS dynamics are predicted (**Figure 6**). The percentage of the predicted long-term SOCS dynamics following PRL stimulation switched—from sustained (predicted with the original model) to transient activation (predicted with the updated model), as shown in **Figure 6A**. Specifically, transient activation accounts for 90% of predictions, where previously it accounted for only 15% of predictions. Based on this, as well as our literature-informed validation metric, we can be reasonably certain that SOCS displays transient activation in beta cell signaling.

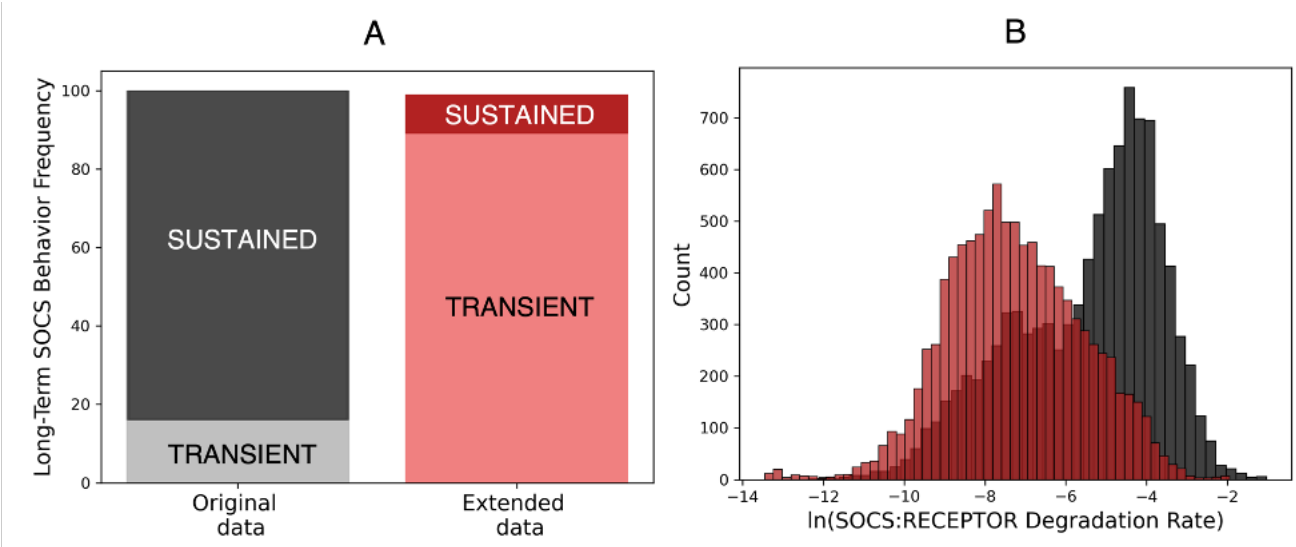

Figure 5. **Comparison between the original and updated models' posterior distribution and predictions.** (A) Bar chart of long-term SOCS behavior classification frequencies before and after data integration. Black, original model (prior to data integration); red, updated model (after data integration). (B) Histograms of the marginal posterior distributions of the PRL:PRLR:SOCS degradation rate before and after integration of data at later time points, displayed on a natural-log scale.

Finally, the change in the SOCS predictions also corresponded with a significant change in the marginal posterior of the PRL:PRLR:SOCS complex degradation rate (**Figure 6B**), as well as smaller changes in the marginal posteriors of other parameters (**Supplementary Figure S10**). In the case of the PRL:PRLR:SOCS complex degradation rate, one mode of the distribution is eliminated after data integration. The mode remaining corresponds to the slower PRL:PRLR:SOCS degradation rate (colored blue in **Figure 4**). Overall, our updated model predicts transient SOCS activation and favors a slower PRL:PRLR:SOCS degradation rate.

4. Discussion

Our work establishes an improved approach to Bayesian hypotheses formation, refines an existing model of pancreatic beta cell signaling, and generates novel biological insights. We introduced KL divergence-based clustering, an approach that facilitates hypothesis generation by prioritizing parameters for qualitative analysis. While other approaches for hypothesis formation using posteriors exist, they are largely qualitative and can be time consuming, as they typically require qualitative evaluation of every model parameter (Hug et al., 2013; Thijssen et al., 2018; Tötsch and Hoffmann, 2020). Our approach streamlines the qualitative evaluation of model parameters by using clustering to select only a subset of parameters for analysis. We prioritize this subset by how informed they are by the data using KL divergences. Prioritizing

parameter estimates that are significantly data-informed for evaluation allows the generation of data-informed hypotheses. Importantly, our approach is versatile. K-means clustering is an unsupervised learning technique (Aldenderfer and Blashfield, 1984; Bible et al., 2013). Our approach is therefore applicable to models with KL divergence patterns unknown *a priori*. Further, KL divergence may be estimated between marginal prior and posterior distributions of any form (Kullback and Leibler, 1951). Our approach is therefore applicable to models with prior and posterior formulations different than those we use here, such as a log-uniform prior distribution.

We tested our approach on a biologically-relevant mathematical model of pancreatic beta signaling and demonstrate that it can be used to generate alternative hypotheses. This model was previously established in Mortlock et al. (Mortlock et al., 2021) and most recently used to investigate cell heterogeneity (Simoni et al., 2022). Before applying our approach, we improved upon the original Bayesian inference procedure presented in Mortlock et al. After verifying convergence of the posterior and verifying that our model adequately explained the available experimental data, we used our KL divergence-based approach to determine if there are multiple biological mechanisms that explain the observed protein dynamics. We found two plausible regimes for the rate of degradation of the SOCS-bound receptor. These regimes characterized unique SOCS dynamics, which formed testable hypotheses of beta cell signaling mechanisms. This result demonstrated how our approach can be used to generate mechanistic hypotheses explaining complex protein signaling networks. Similar studies concluded after proposing hypotheses (Hug et al., 2013; Tötsch and Hoffmann, 2020). In this work, we moved beyond hypothesis formation by testing the proposed SOCS dynamics.

While there has been significant research investigating SOCS' role in the canonical JAK-STAT pathway, as well as the role of SOCS in PRL-mediated JAK2-STAT5 signaling, no observations of SOCS dynamics for our exact system of study exist (Chong et al., 2001; Croker et al., 2008; Jiao et al., 2013; Millette et al., 2022; Ram and Waxman, 1999; Rieck et al., 2009; Rønn et al., 2002; Ye and Driver, 2016). Thus, determining the dynamics of SOCS in response to PRL is an open challenge in pancreatic beta cell biology. To investigate this question, we refined our posterior by incorporating additional data at extended timepoints. We then validated our updated model by ensuring its predictions were consistent with dynamics of the CISH and SOCS2, which are shown to be conserved across experimental conditions (Chong et al., 2001; Croker et al., 2008; Ram and Waxman, 1999). In contrast, the model obtained prior to integrating the new experimental data was not consistent with these target dynamics. This result established the predictive capabilities of our updated model. The model predicts that SOCS

likely displays transient activation after PRL stimulation. Because SOCS provides negative feedback to the JAK-STAT pathway, it plays a crucial role in controlling hormone-mediated beta cell proliferation. Thus, this insight is useful for developing strategies to enhance JAK2-STAT5 signaling and, ultimately, a therapeutic target for Diabetes treatment: beta cell proliferation (Durham et al., 2019).

In addition to determining the most probable SOCS dynamics in response to PRL stimulation, our analysis of the posterior distribution revealed insight about the degradation rate of the SOCS-bound receptor. We investigated how the new data changed the probability of different SOCS-bound receptor degradation rates by comparing the marginal distribution of the SOCS-bound receptor degradation rate before and after data integration. After data integration, the probability of a faster degradation rate decreased, and the slower degradation regime emerged as most plausible. Overall, our analysis of the posterior before and after data integration complements how it has been used in previous works. For example, this simple approach has been used to understand the information in different types of data and to demonstrate model inadequacy (Thijssen et al., 2018; Tötsch and Hoffmann, 2020).

We acknowledge some limitations and future improvements of our work. First, while the clustering results we present here are biologically meaningful, this may not always be the case. Clustering analysis seeks to identify latent structures in data. However, one must consider whether the structure it identifies is “real” or simply imposed on the data by the method (Aldenderfer and Blashfield, 1984; Bible et al., 2013). Here, our top data-informed cluster was deemed meaningful as its members shared a biological mechanism related to receptor degradation. Future work should similarly assess the meaning of clustering results. If no meaningful clusters emerge, one could instead prioritize posterior analysis based solely on KL divergence ranking. Second, while we based our SOCS shape classification criterion on several western blot data sets, there remains an element of ‘modelers choice’ when defining these criteria. If our results were highly sensitive to the criterion definition, this would limit their importance. To address this limitation, we varied both of our classification criteria and ensured our results were conserved. Third, while we conducted a comprehensive literature search and developed a validation for our model with inferable literature evidence, this evidence is less certain than direct observations of SOCS. Future work could focus on collecting western blot measurements of SOCS, particularly from 0-5 hours, to determine whether SOCS displays an early peak in activation. Fourth and final, while rodent models provide valuable insights into human biology, there are certainly limits to the applicability of these insights(Baeyens et al., 2016; Salazar-Petres and Sferruzzi-Perri, 2022). Our mathematical model is calibrated with data

collected in the INS-1 rat cell line (Brelje et al., 2004, 2002; Fujinaka et al., 2007). Though the net effects of pregnancy on beta cells, the expansion of beta cell mass and increased insulin secretory capacity are conserved between rodent and human beta cells, it is unclear how the mechanisms driving these changes in rodent cells map to the mechanisms driving these changes in human cells. Future work can test the applicability of our insights in human beta cell biology.

5. Conclusions

Overall, our work establishes an improved approach to hypothesis formation using the Bayesian posterior, refines an existing model of pancreatic beta cell signaling, and ultimately answers an open question in the field of pancreatic beta cell biology. Our approach to hypothesis formation provides a basis for other investigations of biochemical signaling networks. The refined model provides a quantitative framework for future investigations of beta cell signaling. Finally, the novel insights into pancreatic beta cell biology generated here can potentially be leveraged to develop strategies to increase functional beta cell mass.

Funding Statement

This work was partially supported by a USC Viterbi Graduate School Fellowship to H.A.H.

Acknowledgements

The authors acknowledge members of the Finley research group for constructive feedback.

CRediT authorship contribution statement

Holly A. Huber: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Stacey D. Finley:** Conceptualization, Supervision, Resources, Writing – review & editing.

Declaration of Competing Interest

None

References

Aguayo-Mazzucato, C., Bonner-Weir, S., 2018. Pancreatic β Cell Regeneration as a Possible Therapy for Diabetes. Cell Metab. 27, 57–67. https://doi.org/10.1016/j.cmet.2017.08.007

Aldenderfer, M., Blashfield, R., 1984. Cluster Analysis. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America. https://doi.org/10.4135/9781412983648

Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.I., 2003. An Introduction to MCMC for Machine Learning. Mach. Learn. 50, 5–43. https://doi.org/10.1023/A:1020281327116

Baeyens, L., Hindi, S., Sorenson, R.L., German, M.S., 2016. β-Cell adaptation in pregnancy. Diabetes Obes. Metab. 18 Suppl 1, 63–70. https://doi.org/10.1111/dom.12716

Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D.S., Milo, R., 2011. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. Biochemistry 50, 4402–4410. https://doi.org/10.1021/bi2002289

Bible, J., Datta, Susmita, Datta, Somnath, 2013. Cluster Analysis, in: Informatics for Materials Science and Engineering. Elsevier, pp. 53–70. https://doi.org/10.1016/B978-0-12-394399-6.00004-7

Brelje, T.C., Stout, L.E., Bhagroo, N.V., Sorenson, R.L., 2004. Distinctive roles for prolactin and growth hormone in the activation of signal transducer and activator of transcription 5 in pancreatic islets of langerhans. Endocrinology 145, 4162–4175. https://doi.org/10.1210/en.2004-0201

Brelje, T.C., Svensson, A.M., Stout, L.E., Bhagroo, N.V., Sorenson, R.L., 2002. An immunohistochemical approach to monitor the prolactin-induced activation of the JAK2/STAT5 pathway in pancreatic islets of Langerhans. J. Histochem. Cytochem. Off. J. Histochem. Soc. 50, 365–383. https://doi.org/10.1177/002215540205000308

Brooks, S.P., Gelman, A., 1998. General Methods for Monitoring Convergence of Iterative Simulations. J. Comput. Graph. Stat. 7, 434–455. https://doi.org/10.1080/10618600.1998.10474787

Chaloner, K., Verdinelli, I., 1995. Bayesian Experimental Design: A Review. Stat. Sci. 10. https://doi.org/10.1214/ss/1177009939

Chong, M.M., Thomas, H.E., Kay, T.W., 2001. gamma-Interferon signaling in pancreatic beta-cells is persistent but can be terminated by overexpression of suppressor of cytokine signaling-1. Diabetes 50, 2744–2751. https://doi.org/10.2337/diabetes.50.12.2744

Cowles, M.K., Carlin, B.P., 1996. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. J. Am. Stat. Assoc. 91, 883–904. https://doi.org/10.1080/01621459.1996.10476956

Croker, B.A., Kiu, H., Nicholson, S.E., 2008. SOCS regulation of the JAK/STAT signalling pathway. Semin. Cell Dev. Biol. 19, 414–422. https://doi.org/10.1016/j.semcdb.2008.07.010

Durham, G.A., Williams, J.J.L., Nasim, M.T., Palmer, T.M., 2019. Targeting SOCS Proteins to Control JAK-STAT Signalling in Disease. Trends Pharmacol. Sci. 40, 298–308. https://doi.org/10.1016/j.tips.2019.03.001

Endo, T., Sasaki, A., Minoguchi, M., Joo, A., Yoshimura, A., 2003. CIS1 interacts with the Y532 of the prolactin receptor and suppresses prolactin-dependent STAT5 activation. J. Biochem. (Tokyo) 133, 109–113. https://doi.org/10.1093/jb/mvg004

Fujinaka, Y., Takane, K., Yamashita, H., Vasavada, R.C., 2007. Lactogens promote beta cell survival through JAK2/STAT5 activation and Bcl-XL upregulation. J. Biol. Chem. 282, 30707–30717. https://doi.org/10.1074/jbc.M702607200

Gelman, A., 2014. Bayesian data analysis, Third edition. ed, Chapman & Hall/CRC texts in statistical science. CRC Press, Boca Raton.

Ghasemi, O., Lindsey, M.L., Yang, T., Nguyen, N., Huang, Y., Jin, Y.-F., 2011. Bayesian parameter estimation for nonlinear modelling of biological pathways. BMC Syst. Biol. 5 Suppl 3, S9. https://doi.org/10.1186/1752-0509-5-S3-S9

Hug, S., Raue, A., Hasenauer, J., Bachmann, J., Klingmüller, U., Timmer, J., Theis, F.J., 2013. High-dimensional Bayesian parameter estimation: case study for a model of JAK2/STAT5 signaling. Math. Biosci. 246, 293–304. https://doi.org/10.1016/j.mbs.2013.04.002

Jiao, Y., Rieck, S., Le Lay, J., Kaestner, K.H., 2013. CISH has no non-redundant functions in glucose homeostasis or beta cell proliferation during pregnancy in mice. Diabetologia 56, 2435–2445. https://doi.org/10.1007/s00125-013-3014-x

Kullback, S., Leibler, R.A., 1951. On Information and Sufficiency. Ann. Math. Stat. 22, 79–86. https://doi.org/10.1214/aoms/1177729694

Lomeli, L.M., Iniguez, A., Tata, P., Jena, N., Liu, Z.-Y., Van Etten, R., Lander, A.D., Shahbaba, B., Lowengrub, J.S., Minin, V.N., 2021. Optimal experimental design for mathematical models of haematopoiesis. J. R. Soc. Interface 18, 20200729. https://doi.org/10.1098/rsif.2020.0729

Makaryan, S.Z., Finley, S.D., 2020. Enhancing network activation in natural killer cells: predictions from in silico modeling. Integr. Biol. Quant. Biosci. Nano Macro 12, 109–121. https://doi.org/10.1093/intbio/zyaa008

Miller, D.S.J., Schmierer, B., Hill, C.S., 2019. TGF-β family ligands exhibit distinct signalling dynamics that are driven by receptor localisation. J. Cell Sci. 132, jcs234039. https://doi.org/10.1242/jcs.234039

Millette, K., Rodriguez, K., Sheng, X., Finley, S.D., Georgia, S., 2022. Exogenous Lactogenic Signaling Stimulates Beta Cell Replication In Vivo and In Vitro. Biomolecules 12, 215. https://doi.org/10.3390/biom12020215

Mortlock, R.D., Georgia, S.K., Finley, S.D., 2021. Dynamic Regulation of JAK-STAT Signaling Through the Prolactin Receptor Predicted by Computational Modeling. Cell. Mol. Bioeng. 14, 15–30. https://doi.org/10.1007/s12195-020-00647-8

Nandagopal, N., Santat, L.A., LeBon, L., Sprinzak, D., Bronner, M.E., Elowitz, M.B., 2018. Dynamic Ligand Discrimination in the Notch Signaling Pathway. Cell 172, 869-880.e19. https://doi.org/10.1016/j.cell.2018.01.002

Pedregosa, F., Varoquaux, G., Gramfort, A., Gramfort, V., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://doi.org/10.5555/1953048.2078195

Ram, P.A., Waxman, D.J., 1999. SOCS/CIS protein inhibition of growth hormone-stimulated STAT5 signaling by multiple mechanisms. J. Biol. Chem. 274, 35553–35561. https://doi.org/10.1074/jbc.274.50.35553

Rieck, S., White, P., Schug, J., Fox, A.J., Smirnova, O., Gao, N., Gupta, R.K., Wang, Z.V., Scherer, P.E., Keller, M.P., Attie, A.D., Kaestner, K.H., 2009. The transcriptional response of the islet to pregnancy in mice. Mol. Endocrinol. Baltim. Md 23, 1702–1712. https://doi.org/10.1210/me.2009-0144

Rønn, S.G., Hansen, J.A., Lindberg, K., Karlsen, A.E., Billestrup, N., 2002. The effect of suppressor of cytokine signaling 3 on GH signaling in beta-cells. Mol. Endocrinol. Baltim. Md 16, 2124–2134. https://doi.org/10.1210/me.2002-0082

Salazar-Petres, E.R., Sferruzzi-Perri, A.N., 2022. Pregnancy-induced changes in β-cell function: what are the key players? J. Physiol. 600, 1089–1117. https://doi.org/10.1113/JP281082

Simoni, A.D., Huber, H.A., Georgia, S.K., Finley, S.D., 2022. Phosphatases are predicted to govern prolactin-mediated JAK-STAT signaling in pancreatic beta cells. Integr. Biol. Quant. Biosci. Nano Macro 14, 37–48. https://doi.org/10.1093/intbio/zyac004

Thijssen, B., Dijkstra, T.M.H., Heskes, T., Wessels, L.F.A., 2018. Bayesian data integration for quantifying the contribution of diverse measurements to parameter estimates. Bioinforma. Oxf. Engl. 34, 803–811. https://doi.org/10.1093/bioinformatics/btx666

Tötsch, N., Hoffmann, D., 2020. Bayesian Data Integration Questions Classic Study on Protease Self-Digest Kinetics. ACS Omega 5, 15162–15168. https://doi.org/10.1021/acsomega.0c01109

Yamada, S., Shiono, S., Joo, A., Yoshimura, A., 2003. Control mechanism of JAK/STAT signal transduction pathway. FEBS Lett. 534, 190–196. https://doi.org/10.1016/s0014-5793(02)03842-5

Ye, C., Driver, J.P., 2016. Suppressors of Cytokine Signaling in Sickness and in Health of Pancreatic β-Cells. Front. Immunol. 7, 169. https://doi.org/10.3389/fimmu.2016.00169