Bayesian *Occam's Razor* to Optimize Models for Complex Systems

Chenxi Wang,[1,2,3,†] Jihui Zhao,[1,†] Jingjing Zheng,[1,2,3] Barak Raveh, [4] Xuming He,[5,6*] Liping Sun[1*]

[1]iHuman Institute, ShanghaiTech University, Shanghai 201210, China
[2]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]School of Computer Science and Engineering, The Hebrew University of Jerusalem, 9190416, Israel
[5]School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
[6]Shanghai Engineering Research Center of Intelligent Vision and Imaging, Shanghai 201210, China

[†]These authors contributed equally to this work
[*]To whom correspondence should be addressed E-mail: hexm@shanghaitech.edu.cn; sunlp@shanghaitech.edu.cn.

Abstract

Developing and optimizing models for complex systems poses challenges due to the inherent complexity introduced by multiple types of input information and sources of uncertainty. In this study, we utilize Bayesian formalism to analytically examine the propagation of probability in the modeling process and propose quantitative assessments for it. Upon which, we develop a method for optimizing models for complex systems by (i) minimizing model uncertainty; (ii) maximizing model consistency; and (iii) minimizing model complexity, following the Bayesian *Occam's razor* rationale. We showcase the benefits of this method by optimizing the modeling of the dynamic system of glucose-stimulated insulin secretion in pancreatic β-cells, leading to an optimized model that demonstrates better alignment with experimental observations compared to the non-optimized one. We anticipate that this method will facilitate the construction of accurate, precise, and sufficiently simple models for diverse complex systems. It is implemented in our open-source software *Integrative Modeling Platform* (IMP), ensuring its broad applicability.

MAIN TEXT

Introduction

Models are essential for understanding, modulating, and designing future studies on systems of interest. They are built upon input information (*eg*, physical theories, statistical inference, experimental data, and prior models), and are used to rationalize input information and make predictions. A model (SI Text 1.Glossary) can be defined as a joint distribution over all model variables. Most generally, a modeling process involves three steps: (i) define the model representation that specifies the degrees of freedom whose values are determined by modeling, (ii) define a scoring function for ranking alternative models by their agreements with the input information, and (iii) generate an ensemble of good-scoring models through sampling [1, 2, 3]. Ideally, we want to find all models that satisfy the input information, which allows us to capture the uncertainties in the input information.

Modeling complex systems often necessitates an integrative modeling approach [3, 4]. One example is cell modeling, where a variety of models have been constructed by integrating multiple types of information describing various aspects of the cell. These include spatial-temporal models, ordinary differential equation models, and hierarchical structural models [5, 6, 7]. Metamodeling, a special case of integrative modeling, integrates a collection of input models into a metamodel (*ie*, a model of these models) [8, 9]. Indeed, experimental data can also serve as an input model (*eg*, a single-variable model). Recently, we developed a general metamodeling framework utilizing Bayesian networks (BNs), a well-studied class of Probabilistic Graphical Models (PGMs) [10, 11]. In analogy to representation, scoring, and sampling in a general modeling process, it proceeds through three stages: (i) convert input models into probabilistic surrogate models; (ii) couple surrogate models; and (iii) update surrogate models and input models *via* backpropagation. Metamodeling is of particular importance for modeling, approximating, or optimizing systems of high-complexity, including protein structures [4], engineering designs [12], and policy simulations [13].

Model accuracy (SI Text 1.Glossary) is defined as the degree of conformity of the model to the ground truth. While the ground truth is never known, model accuracy can be estimated by comparing the model variable to an independent reference closely representing the ground truth (*ie*, a gold standard) [14]. Model precision (SI Text 1.Glossary) is the reciprocal of model uncertainty, which can be measured by the spread (*eg*, standard deviation, SD [15]) or the entropy [16] of the model distribution. Model complexity reflects the technical sophistication or quantity of its components. Depending on the specific model representation, model complexity can be measured by the number of variables or the depth of relationships among them. Simpler models should be preferred over unnecessarily complex ones, as suggested by *Occam's Razor* [17, 18]. Given these considerations, the objective of integrative modeling can be formulated as constructing an optimal model, in the sense that it achieves maximal accuracy and precision while maintaining sufficient simplicity, given the accuracy and precision of input information. Upon achieving these objectives, we can expect reliable and efficient interpretation of model results and formulation of model predictions.

Model optimization often involves quantitative assessments of the agreement between model variables and the input information under uncertainty throughout the modeling process [19, 20]. For instance, the representation of a model can be optimized at a certain precision (*ie*, coarse-graining level) to facilitate exhaustive sampling [21]. The scoring function can be optimized by translating frequencies of sparse Nuclear Magnetic Resonance (NMR) data into likelihoods, effectively capturing their uncertainties in protein structure modeling [22]. The sampling process can be optimized to generate an ensemble of models (*eg*, using bagging, boosting, and stacking), facilitating the assessment of model uncertainty [23, 24]. Techniques for model optimization have found extensive applications in various fields, such as predicting climate change impacts [25], risk management [26], and improving the reliability of structure models [27]. However, greater challenges arise in the optimization of models for complex systems due to their inherent high complexity.

Fortunately, Bayesian methods provide a natural probabilistic modeling framework, allowing us to evaluate model variables under uncertainty. Moreover, Bayesian probability theory chooses between alternative results obtained from different models in a manner that embodies *Occam's Razor* [28, 29]. Specifically, this principle favors simpler models due to their tendency to yield lower variability and, consequently, lower model uncertainty. Here, we utilize the intricate case of metamodeling to illustrate model optimization for complex systems, while noting that our method applies to other modeling approaches as well. Leveraging the Bayesian formalism, we establish definitive rules and propose quantitative assessments for the probability propagation from input

models to the metamodel. This is demonstrated through an analytical example involving two Gaussian time-independent input models. Subsequently, we develop a method to optimize metamodeling by (i) minimizing model uncertainty; (ii) maximizing model consistency; and (iii) minimizing model complexity. We apply this method in selecting surrogate models and coupling strategies, as well as in including observations of model variables as an additional data model to resolve conflicts among surrogate models. Finally, we validate the optimized metamodel of the dynamic GSIS system by comparing it with experimental observations, demonstrating better agreements compared to the non-optimized one.

Results

Analytical propagation of probability in metamodeling

Model variables can be classified into two distinct types: (i) target variables, which are of primary interest and inform the answers to questions posed by a model; (ii) nuisance variables, which define the modeling process (SI Text 2.1). We examine how probability propagates for target variables across metamodeling stages.

Probability propagation in the conversion stage

We assume that each input model is provided as a joint distribution spanning its variables (Fig. 1, Input models). The input model distribution is represented by a single model instance along with its associated uncertainty. First, each input model is converted into its corresponding probabilistic surrogate model. The surrogate model (SI Text 1.Glossary) is a potentially simplified input model. It is defined formally by a PGM, which provides both the statistical dependencies between the surrogate model variables (*ie*, surrogate PGM graph) and the joint distribution (*ie*, surrogate PGM distribution) (Fig. 1, Surrogate models) [10]. The conversion involves two steps: (i) select surrogate model variables, often a subset of input model variables and potential additional variables, and (ii) construct the surrogate PGM through various conversion methods such as linear regression fitting [11], Kriging modeling [30], and machine learning methods [31]. The conversion stage is analogous to defining a unified model representation given multiple types of input information, thus specifying the degrees of freedom for the subsequent modeling process. Note that input data can be treated as single-variable models, and can thus be converted into single-variable surrogate models with the same probability distribution.

Next, we analytically propagate the probability of a Gaussian input model to its surrogate model. The input model is defined as:

$$\mathbf{X_{in}} \sim \mathcal{N}(\boldsymbol{\mu}_{in}, \boldsymbol{\Sigma}_{in}), \mathbf{X_{in}} \in \mathbb{R}^n \tag{1}$$

where $\boldsymbol{\mu}_{in}$ and $\boldsymbol{\Sigma}_{in}$ are the mean and covariance matrix of the input model distribution. For example, the input model a is a joint distribution $P(\mathbf{X^a_{in}})$ over a set of two random variables $\mathbf{X^a_{in}} = [X^a_{1,in}, X^a_{2,in}]$ (Fig. 1 Input models). The probability propagation from the input model to the surrogate model follows the chain rule:

$$P(\mathbf{X^a_s}, C_{cv}|\mathbf{X^a_{in}}) \propto P(\mathbf{X^a_s}|\mathbf{X^a_{in}}, C_{cv})P(C_{cv}|\mathbf{X^a_{in}}) \tag{2}$$

where $\mathbf{X^a_s}$ represents surrogate model variables, C_{cv} is the nuisance variable indicating the conversion method. The surrogate model distribution over all conceivable conversion methods $P(\mathbf{X^a_s}, C_{cv}|\mathbf{X^a_{in}})$ is proportional to the surrogate model distribution given the input model and a single conversion method $P(\mathbf{X^a_s}|\mathbf{X^a_{in}}, C_{cv})$, and the distribution of all conversion methods given the input model $P(C_{cv}|\mathbf{X^a_{in}})$. If the conversion method is fixed by the modeler as C^*_{cv} (*ie*, $P(C^*_{cv}|\mathbf{X^a_{in}}) = 1$), then the surrogate PGM distribution is the joint distribution over its surrogate model variables given the input model, $P(\mathbf{X^a_s}|\mathbf{X^a_{in}})$.

With the probability distributions of input models and surrogate models in hand, we can quantify the probability propagation in this stage by (i) changes in model uncertainty as the difference in entropy between the input model $h(\mathbf{X^a_{in}})$ and the surrogate model $h(\mathbf{X^a_s})$; (ii) overlaps between the marginal distributions of aligned variables in the input and surrogate models [32]. The overlap index for a specific variables pair $\eta(X^a_s, X^a_{in})$ is:

$$\eta(X^a_s, X^a_{in}) = \int_{-\infty}^{\infty} \mathbf{min}[p_s(X^a_s), p_{in}(X^a_{in})]dx \tag{3}$$

where $p_{in}(X^a_{in})$ and $p_s(X^a_s)$ are the marginal distributions (*ie*, probabilistic density functions) of single input and surrogate model variables, respectively. The average of overlaps $\bar{\eta}(\mathbf{X^a_s}, \mathbf{X^a_{in}})$, defined as model consistency (SI Text 1.Glossary), reflects the effectiveness of the surrogate models in reproducing the input model. Comparing changes in the uncertainty of surrogate models and consistency between input models and surrogate models can potentially guide us to optimize the conversion stage and select the best surrogate model. Notably, when two candidate surrogate models show comparable uncertainty and consistency with the input model, the optimal surrogate model should be the one with lower complexity (brown PGM in Fig. 1 Surrogate models, consisting of one node and zero edges) compared to the other one (grey PGM in Fig. 1 Surrogate models, consisting of two nodes and one edge).

Probability propagation in the coupling stage

After the conversion of input models into surrogate models, we couple these surrogate models into a metamodel. The metamodel (SI Text 1.Glossary) is defined by a meta PGM, including a meta PGM graph over metamodel variables (*ie*, all surrogate model variables and coupling variables) and a meta PGM distribution spanned over all variables in the meta PGM graph (Fig. 1, Metamodel). The coupling stage is achieved by the construction of a coupling PGM graph (SI Text 1.Glossary) to relate all connecting variables and all the coupling variable(s). It proceeds through three steps: (i) identify connecting variables across surrogate models; (ii) introduce one or more explicit coupling variables to formally relate the connecting variables; (iii) specify conditional probability distributions (CPDs) over the connecting variables and the coupling variable(s). In addition, one might construct the coupling PGM graph by defining CPDs between the connecting variables without introducing explicit coupling variables. For example, two subunits in a protein complex can be coupled directly *via* the binding free energy. The coupling stage is analogous to scoring different pieces of information in other model approaches, such as defining a scoring function combining distance and angle restraints in protein modeling [33], or developing a constraint to integrate macroscopic traffic flow models with microscopic agent-based models [34].

Next, we analytically propagate the probability of two static surrogate models to the metamodel. Assuming the surrogate models are constructed using a single conversion method, the probability propagation from the surrogate models to the metamodel follows the chain rule:

$$P(\mathbf{X_{meta}}, C_{cs}|\mathbf{X_s}, \mathbf{X_{in}}) \propto P(\mathbf{X_s}|\mathbf{X_{meta}}, C_{cs}, \mathbf{X_{in}})P(\mathbf{X_{meta}}|\mathbf{X_{in}}) \tag{4}$$

where $\mathbf{X_{meta}}$ represents metamodel variables, C_{cs} is a nuisance variable indicating the coupling strategy. The meta PGM distribution over all coupling strategies $P(\mathbf{X_{meta}}, C_{cs}|\mathbf{X_s}, \mathbf{X_{in}})$ is proportional to the likelihood $P(\mathbf{X_s}|\mathbf{X_{meta}}, C_{cs}, \mathbf{X_{in}})$ and the prior $P(\mathbf{X_{meta}}|\mathbf{X_{in}})$. The likelihood measures how each surrogate model is satisfied by a metamodel given the coupling strategies and input models. The prior reflects prior knowledge about the statistical dependencies between variables across input models.

Suppose we identify connecting variables from two surrogate models with a simple linear relationship $X^a_{1,s} = kX^b_{1,s}$, where k can be specified by prior knowledge or statistical analysis (SI Text 3.2). We can then introduce an explicit coupling variable Y, whose prior distribution $P(Y)$ is a weighted sum of the marginal distributions of the two connecting variables:

$$P(Y) = \omega_a P(X^a_{1,s}) + \omega_b P(X^b_{1,s}); \frac{\omega_a}{\omega_b} = \frac{\sigma_{X^a_{1,s}}}{k\sigma_{X^b_{1,s}}}; \omega_a + k\omega_b = 1 \tag{5}$$

For such a coupling PGM graph composed of two connecting variables and one coupling variable, there are three possible graph structures: head-to-tail, head-to-head, and tail-to-tail. Note that information about any of the nodes along the PGM chain updates the information of other nodes bidirectionally [10, 35]. We now consider the head-to-tail coupling PGM graph, $X^a_{1,s} \to Y \to X^b_{1,s}$. Forwardly, a new estimation for the probability distribution of $X^a_{1,s}$ updates $X^b_{1,s}$ by propagating information over Y. Backwardly, observing Y updates the estimation of $X^a_{1,s}$. We then specify two Gaussian CPDs for $P(Y|X^a_{1,s})$ and $P(X^b_{1,s}|Y)$ to define the probability propagation along the coupling PGM graph and consequently the coupling strength between two surrogate models (SI Text 4, eq.9-10):

$$P(Y|X^a_{1,s}) \propto \exp\{-\frac{\|Y - \mu^a_{1,s}\|^2}{2(\omega^a_{1,s})^2}\}; P(X^b_{1,s}|Y) \propto \exp\{-\frac{\|X^b_{1,s} - \mu_y\|^2}{2(\omega^b_{1,s})^2}\}$$

where $\mu^a_{1,s}$ is the mean of the marginal distributions of the connecting variables $X^a_{1,s}$, μ_y is the mean of the prior distribution of the coupling variable, $(\omega^a_{1,s})^2$ and $(\omega^b_{1,s})^2$ are the variances of the marginal distributions of the connecting variables $X^a_{1,s}$ and $X^b_{1,s}$, respectively. Note that additional information can be used to inform the prior distribution of the coupling variable $P(Y)$ (*eg*, assign an observation based on external experiments) and the CPDs (*eg*, assign confidence to each surrogate model based on their uncertainties).

In cases where the modeler fixes a single coupling strategy C^*_{cs} as exemplified here (*ie*, $P(C^*_{cs}|\mathbf{X_s}, \mathbf{X_{in}}) = 1$), the meta PGM distribution is then the joint distribution over all metamodel variables, given the surrogate and input model variables, $P(\mathbf{X_{meta}}|\mathbf{X_s}, \mathbf{X_{in}})$ (Fig. 1, Metamodel). It can be factorized as:

$$\begin{aligned} P(\mathbf{X_{meta}}|\mathbf{X_s}, \mathbf{X_{in}}) &= P(X^a_{1,s}, X^b_{1,s}, X^b_{2,s}, Y|\mathbf{X_s}, \mathbf{X_{in}}) \\ &= P(X^a_{1,s}|\mathbf{X_{in}})P(Y|X^a_{1,s}, \mathbf{X_{in}}) \times P(X^b_{1,s}|Y, \mathbf{X_{in}})P(X^b_{2,s}|X^b_{1,s}, \mathbf{X_{in}}) \end{aligned} \tag{6}$$

The probability propagation in this stage can be quantified by (i) comparing the difference in surrogate model entropy before and after coupling, and (ii) examining the average overlap index between the marginal distributions of surrogate model variables before and after coupling, indicated by

$$\bar{\eta}(\mathbf{X_{s,upd}}, \mathbf{X_s}) = \frac{1}{N}\sum_{i=1}^{N} \eta(x^i_{\text{s, upd}}, x_{\text{s}}) \tag{7}$$

where N indicates the number of surrogate model variables. This conveys model consistency and thus reflects the rationalization among pieces of inputs (SI Text 1.Glossary) [36]. Comparing changes in the uncertainty of surrogate models and the consistency among them before and after metamodeling can potentially guide us to optimize the coupling stage and select the best coupling PGM graph. For example, the optimal one should yield a metamodel with lower entropy and higher consistency (*eg*, orange PGM in Fig. 1, Metamodel, head-to-tail) compared to the alternative one (*eg*, grey PGM in Fig. 1, Metamodel, head-to-head).

Probability propagation in the backpropagation stage

After coupling, surrogate models and the corresponding input models are updated by backpropagation. The backpropagation stage is comparable to the process of sampling good-scoring models in a broader modeling context [10]. First, we update the surrogate PGM distributions given the meta PGM distribution obtained in the coupling stage (Fig. 1, Updated surrogate models). Depending on the specific query, this is achieved by either conditioning on or marginalizing out variables in the metamodel using different model inference methods (*eg*, junction tree algorithm[37], continuous Boyen-Koller algorithm [38], reject sampling [39], importance sampling [40]). The updated PGM distribution of each surrogate model Eq.3 is:

$$P(\mathbf{X^i_{s,upd}}|\mathbf{X_{meta}}, \mathbf{X_{in}}) = P(\mathbf{X^i_{s,upd}}|\mathbf{X^{u|u\neq i}_{meta}}, \mathbf{X_{meta}}, \mathbf{X_{in}}) \tag{8}$$

$$\text{or} = \int_{\mathbf{X^{u|u\neq i}_{meta}}} P(\mathbf{X_{s,upd}}|\mathbf{X_{meta}}, \mathbf{X_{in}})dY\, d\mathbf{X^{u|u\neq i}_{meta}} \tag{9}$$

where $\mathbf{X^{u|u\neq i}_{meta}}$ are the variables in the metamodel but not in the surrogate model i. Thus, the surrogate model is updated under the contextualization of all other surrogate models. We can also compute the updated distributions spanned by any subsets of variables in the metamodel. Comparing the efficiencies of different inference algorithms facilitates the optimization of the backpropagation stage, which is particularly important when modeling complex systems. We observed a significantly higher efficiency with the junction tree algorithm compared to the continuous Boyen-Koller algorithm (SI Fig. S5). As a result, we have adopted it for all subsequent model inference analyses. Second, depending on the modeler's interest, one may update the input model based on the updated surrogate model (Fig. 1, Updated input models). For example, the input model of the atomistic protein structure can be updated from the updated surrogate model of the coarse-grained structure using backmapping methods [41]; the input model of a system of ODEs can be updated by linear regression of the time series data sampled from the updated surrogate model [42, 43, 44]. The reconstruction of the updated input model is a separate modeling problem and thus not addressed here.

Next, we compute the updated distribution for surrogate model a, $P(\mathbf{X^a_{s,upd}}|\mathbf{X_{meta}}, \mathbf{X_{in}})$ from the meta PGM distribution of a head-to-tail coupling PGM graph ($X^a_{1,s} \to Y \to X^b_{1,s} \to X^b_{2,s}$) using exact inference [45]:

$$P(\mathbf{X^a_{s,upd}}|\mathbf{X_{meta}}, \mathbf{X_{in}}) = \int\int\int P(X^a_{1,s}, X^b_{1,s}, X^b_{2,s}, Y|\mathbf{X_{meta}}, \mathbf{X_{in}})dY\, dX^b_{1,s}\, dX^b_{2,s} \tag{10}$$

The updated surrogate model is a marginal distribution derived from the joint Gaussian distribution of the metamodel, and thus a joint Gaussian distribution (SI Text 4, eq.23-24). To exemplify, we compute the mean $\boldsymbol{\mu^a_{s,upd}}$ and variance $\boldsymbol{\Sigma^a_{s,upd}}$ of the updated PGM distribution of surrogate model a, which is a single variable model (SI Text 4, eq.25-26):

$$\boldsymbol{\Sigma^a_{s,upd}} = (\sigma^a_{1,s,upd})^2 = \frac{(\sigma^a_{1,s})^2[(\omega_{X^a_{1,s}})^2 + (\omega_{X^b_{1,s}})^2]}{(\sigma^a_{1,s})^2 + (\omega_{X^a_{1,s}})^2 + (\omega_{X^b_{1,s}})^2} \tag{11}$$

We can thus compute changes in its model uncertainty before and after metamodeling using the ratio of model variances:

$$\frac{(\sigma^a_{1,s,upd})^2}{(\sigma^a_{1,s})^2} = \frac{(\omega_{X^a_{1,s}})^2 + (\omega^b_{1,s})^2}{(\sigma_{X^a_{1,s}})^2 + (\omega_{X^a_{1,s}})^2 + (\omega_{X^b_{1,s}})^2} < 1 \tag{12}$$

Plausibly, we analytically demonstrate that the uncertainty of the updated surrogate model a is guaranteed to decrease after metamodeling when considering the head-to-tail coupling PGM graph (Refer to SI Text 4 for the complete probability calculus). Till here, we have offered analytical insights into the probability propagation from input models to the metamodel, and have proposed quantitative assessments for it. These analyses guide us in optimizing models for complex systems by sampling each modeling stage and systematically comparing model entropy, model consistency, and model complexity, as outlined below.

Optimization of metamodeling

We quantify the probability propagation across metamodeling stages using the marginal distributions of aligned variables instead of the model joint distributions, as input models and surrogate models often contain different numbers of variables. Aligned variables (SI Text 1.Glossary) are those found in both the input and surrogate models or within the surrogate model before and after metamodeling. Specifically, we quantify model uncertainty (SI Text 1.Glossary) by summing the entropies of marginal distributions of aligned variables. We compute model consistency (SI Text 1.Glossary) by averaging the overlaps between the marginal distributions of aligned variables. We also consider model complexity as the total degree of the underlying PGM graph, d [46, 47]. Note that other criteria can be employed for quantifying model complexity to accommodate different model representations (*eg*, the number of model variables). Based on these, we address the optimization problem by (i) improving the selection of the surrogate model in the conversion stage, (ii) the selection of the coupling PGM graph in the coupling stage with and without references, (iii) including observations of model variables as an additional data model to resolve conflicts.

Selection of the surrogate model

A good surrogate model should reproduce the variables of interest in the input model by recapitulating their statistical dependencies, while maintaining a reasonable level of complexity. Based on the analysis of probability propagation in the conversion stage, we select the best surrogate model by enumerating and ranking surrogate models by: (i) the uncertainty of each surrogate model i (*ie*, the model entropy $h(\mathbf{X^i_s})$); (ii) the degree of consistency between surrogate and input models (*ie*, the average overlap index for all aligned variables $\bar{\eta}(\mathbf{X^i_s}, \mathbf{X^i_{in}})$); and (iii) the surrogate model complexity (*ie*, total degree of the surrogate PGM graph d). We take the postprandial response (PR) model as the input model [48], a system of ODEs describing changes in insulin and glucose levels after a meal. To estimate its joint distribution, we first assign the uncertainty using the standard error associated with the model variables, derived from the experimental data used to build the model (SI Text 5.1). We construct the surrogate PGM graph (Fig. 2A) by (i) selecting eight surrogate model variables $\mathbf{X^{PR}_s}$ out of twelve input model variables $\mathbf{X^{PR}_{in}}$ and (ii) constructing a dynamic Bayesian network (DBN) by incorporating conditional statistical dependencies derived from the ODEs as edges in the graph [49]. We compute the surrogate PGM distribution by employing probabilistic inference over the surrogate PGM graph *via* the junction tree algorithm.

In principle, we should enumerate the degrees of freedom in the conversion methods to select the optimal surrogate model. This involves the number of surrogate model variables, as well as all conceivable graph structures and CPDs in the surrogate PGM graph. To illustrate, we construct three surrogate models for the postprandial response (PR) input model with different complexities ($d(\mathbf{X^{PR}_{s,1}})$=36; $d(\mathbf{X^{PR}_{s,2}})$=36; $d(\mathbf{X^{PR}_{s,3}})$=42)

(Fig. 2A, SI Table S4). Surrogate model 1 exhibits lower entropy ($h(\mathbf{X^{PR}_{s,1}})$=67.8) and higher consistency ($\bar{\eta}(\mathbf{X^{PR}_{s,1}}, \mathbf{X^{PR}_{in}})$=0.37) compared to surrogate model 2 ($h(\mathbf{X^{PR}_{s,2}})$=68.2, $\bar{\eta}(\mathbf{X^{PR}_{s,2}}, \mathbf{X^{PR}_{in}})$=0.30); and similar model entropy and higher model consistency compared to surrogate model 3 ($h(\mathbf{X^{PR}_{s,3}})$=67.9, $\bar{\eta}(\mathbf{X^{PR}_{s,3}}, \mathbf{X^{PR}_{in}})$=0.25) (Fig. 2B). As expected, variations in entropy and consistency of individual variables or the same variable at different time points may not necessarily align with those of the entire models (Fig. 2CDEF). For example, the plasma insulin level over 420 min in surrogate model 1 shows similar entropy ($h(I^{PR}_{1,s})$=8.5, $h(I^{PR}_{3,s})$=8.5) and significantly higher consistency ($\eta(I^{PR}_{1,s}, I^{PR}_{in})$=0.56, $\eta(I^{PR}_{3,s}, I^{PR}_{in})$=0.03) compared to surrogate model 3 (Fig. 2C). However, the plasma insulin level at 30 min in surrogate model 1 shows a lower variable uncertainty ($\sigma_1 < \sigma_3$) and a similar overlap index ($\eta_1 = \eta_3$) compared to surrogate model 3 (Fig. 2E). Among three candidate surrogate models, the measures of model entropy, consistency, and complexity guide us in selecting the optimal surrogate model, *ie*, the surrogate model 1.

Selection of the coupling PGM graph

A good coupling PGM graph should minimize uncertainty of surrogate models after metamodeling. Furthermore, it should maximize consistency among surrogate models before and after metamodeling, *ie*, changes them minimally while still achieving their statistical dependency restrained by coupling, thus rationalizing them. Based on the analysis of probability propagation in the coupling stage, we select the best coupling PGM graph by enumerating and ranking coupling PGM graphs by: (i) the overlap between the model variable distribution and its reference $\eta(X_{s,upd}, X_{ref})$, if ground truth or gold standard is available; (ii) the updated surrogate model uncertainty (*ie*, the sum of entropy for all updated surrogate models $\sum h(\mathbf{X^{i}_{s,upd}})$); (iii) the consistency among surrogate models (*ie*, the average overlap index for all surrogate model variables before and after metamodeling $\bar{\eta}(\mathbf{X_{s,upd}}, \mathbf{X_s})$); and (iv) the model complexity resulted from specific coupling PGM graphs (*ie*, total degree of the meta PGM graph d). To ensure uniformity in selecting the best coupling PGM graphs with and without references, we introduce the coupling variables along with observation variables. The priors of these observation variables are estimated from the distributions of connecting variables in all coupling PGM graphs, or from a reference if available. Differences in model entropy and model consistency with and without observation variables of coupling variables are shown in SI Text 5.3, SI Fig. S6.

We construct the GSIS metamodel upon three time-dependent Gaussian input models: a system of ODEs describing the postprandial response (PR) [48]; a linear model of a pancreatic cell population (Pa) [11]; and a coarse-grained Brownian dynamics simulation of insulin vesicle exocytosis (VE) [50] (SI Text 5.1, Fig. S1). We construct the coupling PGM graph by identifying six connecting variables, introducing four coupling variables, and specifying seven CPDs over a head-to-tail coupling PGM graph (Fig. 3A). First, we identify the connecting variables using: (i) model variable semantics (*eg*, both S^{VE} in the VE model and S^{Pa}_{cell} in the Pa model describe the rate of insulin secretion of a cell); (ii) prior knowledge about the statistical relations between connecting variables (*eg*, the plasma glucose concentration G^{PR} in the PR model has been estimated to be double the intracellular glucose concentration G^{VE} in the VE model [51]); (iii) statistical analysis (*eg*, S^{PR} in the PR model and S^{Pa}_{is} in the Pa model present a Pearson correlation coefficient [52] of 0.94 in their corresponding surrogate models) (SI Text 5.2, SI Fig. S2). Second, we introduce the coupling variables to formally relate the connecting variables by (i) introducing variables of the same semantics as the connecting variables (*eg*, the coupling variable S^{C}_{pa} relates $S^{PR}$and S^{Pa}_{pa}, all describes the rate of insulin secretion of a pancreas); and (ii) introducing new variables to depict statistical dependencies between the connecting variables (*eg*, the coupling variables G^{C}_{pl} and G^{C}_{cell} relate the glucose concentration in plasma G^{PR} and in the cell G^{VE}). Third, we specify the CPDs between the connecting variables and the coupling variable(s) using (i) statistical dependencies based on prior knowledge (*eg*, three CPDs are specified over the connecting variables G^{VE}_t, G^{PR}_t and the coupling variables $G^{C}_{pl,t}$, $G^{C}_{cell,t}$ in the coupling PGM graph $G^{PR}_t \rightarrow G^{C}_{pl,t} \rightarrow G^{C}_{cell,t} \rightarrow G^{VE}_t$ based on experimental measurements [51]:

$$\begin{aligned} < P(G^{C}_{pl,t}|G^{PR}_t > &=< G^{PR}_t > \\ < P(G^{C}_{cell,t}|G^{C}_{pl,t}) > &= 0.5 < G^{C}_{pl,t} > \\ < P(G^{VE}_t|G^{C}_{cell,t}) > &=< G^{C}_{cell,t} > \end{aligned} \tag{13}$$

where <> denotes the mean of a Gaussian distribution); and (ii) exhaustive sampling (see below, Selection of the coupling PGM graph without reference). We specify CPDs across time slices in the coupling PGM graph by defining both the intra-time-slice statistical dependencies between the coupling variables and the connecting variables, and by the intra-time-slice or inter-time-slice statistical dependencies over variables in individual

surrogate models. For example, CPD of the time-dependent variable $S^{VE}_{t+\Delta t}$ in the coupling graph is defined by the coupling variable at the same time slice $S^{C}_{cell,t+\Delta t}$, the connecting variables at the previous time slice S^{VE}_{t} and other variables in the surrogate VE model, denoted as $P(S^{VE}_{t+\Delta t}|S^{C}_{cell,t+\Delta t}, S^{VE}_{t}, G^{VE}_{t}, \dots)$ (SI Text 5.2, SI Table S5).

Selection of the coupling PGM graph with reference When we have access to the ground truth or the gold standard of certain model variables (*eg*, S^{VE}), our aim is to find a coupling PGM graph that maximizes the overlap $\eta(S^{VE}_{s,upd}, S^{VE}_{ref})$ between the variable S^{VE} and its reference. To illustrate, we construct 121 metamodels with different means and variances of $P(G^{C}_{pl}|G^{PR}, G^{obs}_{pl})$ in one coupling PGM graph structure (Fig. 3A), resulting in varying $\eta(S^{VE}_{s,upd}, S^{VE}_{ref})$ values (Fig. 3B). When the variance σ^2 of $P(G^{C}_{pl}|G^{PR})$ exceeds 10 mM, the overlap is seen to remain low regardless of the mean $< P(G^{C}_{pl}|G^{PR}) >$ (point C in Fig. 3B). For instance, at 300 min, the overlap between the distribution of S^{VE} and its reference is 0.1 (Fig. 3C). When the variance is below 10 mM, a region of substantial overlaps emerges when sampling different means (4.5 mM - 5.5 mM, red region in Fig. 3B). The maximal overlap occurs at a mean of 5.1 mM and a variance of 1 mM (point D in Fig. 3B), which is 0.68 at 300 min (Fig. 3D). Thus, we find the optimal coupling PGM graph with the most reliable estimation for S^{VE} against its reference among the candidate ones (point D in Fig. 3B). Furthermore, we observe a lower model uncertainty ($h_D < h_C$) and a higher model consistency ($\bar{\eta}_D > \bar{\eta}_C$) resulting from the optimized coupling PGM graph compared to the non-optimized one (Fig. 3E).

Selection of the coupling PGM graph without reference In practice, however, we might not have access to the reference of any model variables. Our aim is to find a coupling PGM graph that minimizes the entropy of surrogate models ($\sum h(\mathbf{X^{i}_{s}})$ and $\sum h(\mathbf{X^{i}_{s,upd}})$) and maximizes the consistency among surrogate models ($\bar{\eta}(\mathbf{X_{s,upd}}, \mathbf{X_{s}})$), as well as minimizes the complexity of the metamodel ($d(\mathbf{X_{meta}})$). Ideally, we enumerate the degrees of freedom in the coupling strategies to select the optimal coupling PGM graph. This involves the number of connecting variables, the number of coupling variables, all conceivable graph structures, and CPDs in the coupling graph. To simplify, we showcase these degrees of freedom separately.

First, we select the connecting and coupling variables based on model variable semantics, prior knowledge, and statistical analysis, as already outlined above and in SI Fig. S2. Next, we construct three coupling PGM graph structures using different sets of connecting and coupling variables (Fig. 4A-C). Note that we use CPDs in SI Table S5 to illustrate the selection of coupling PGM graph structures, rather than exhaustively searching for the best set of CPDs for each candidate structure. Among these, G1 exhibits the lowest model entropy (h_1=156.5) and the highest model consistency ($\bar{\eta}_1$=0.63). Hence, we select the optimal coupling PGM graph structure G1, despite its relatively high model complexity (d_1=94). Note that the structure and seven CPDs in the selected coupling PGM graph G1 (Fig. 4A) are identical to those in Fig. 3. Third, we construct a total of 121 metamodels with varying means and variances of $P(G^{C}_{pl}|G^{PR}, G^{obs}_{pl})$ to select its optimal CPD. Noteworthy, we are guaranteed to find one with the maximal model consistency by not changing the surrogate models, *ie*, the model consistency $\bar{\eta}(\mathbf{X_{s,upd}}, \mathbf{X_{s}})$ is 1, indicating that the surrogate models are not effectively coupled or rationalized. To avoid this issue, we implement restraints as follows: (i) we define distributions for the observations of all coupling variables (*eg*, $P(S^{obs}_{pa})$ in Fig. 4A) based on the marginal distributions of two connecting variables before metamodeling (SI Text 5.2, eq.46, Fig. S3);

$$P(X^{obs}) = N(h(a,c), \rho), h(a,c) = \frac{a+c}{2}$$
$$\rho^2 = \frac{c^2+d^2}{2} + \frac{(a-c)^2}{4} \tag{14}$$

(ii) we specify the means and variances of three CPDs, $P(S^{PR}|S^{C}_{pa})$, $P(G^{VE}|G^{C}_{cell})$ and $P(S^{Pa}|S^{C}_{cell})$ in the coupling PGM graph based on prior knowledge about their statistical dependencies. Under these restraints, we sample the CPDs introduced between the coupling variables, their observations and the connecting variables (*eg*, the mean of $P(G^{C}_{pl}|G^{PR}, G^{obs}_{pl})$ equals to $w_1 P(G^{PR}) + w_2 P(G^{obs}_{pl})$ and $w_1 + w_2$=1). As shown in Fig. 4DE, with the variance below 10 mM, the model entropy after metamodeling $\sum h(\mathbf{X^{i}_{s,upd}})$ is lower than that before metamodeling $\sum h(\mathbf{X^{i}_{s}})$ (Fig. 4D, blue region), the model consistency $\bar{\eta}(\mathbf{X_{s,upd}}, \mathbf{X_{s}})$ achieves a maximal region with its mean ranging between 4 mM and 6 mM (red region in Fig. 4E). Thus, among the 121 CPDs in coupling

PGM graph G1, the optimal CPD for $P(G^C_{pl}|G^{PR}, G^{obs}_{pl})$ is found at a mean of 5.1 mM and a variance of 1 mM, resulting in minimal model uncertainty (Fig. 4D, point F) and maximal model consistency (Fig. 4E, point F). Another example of the selection of CPDs is conducted for $P(S^C_{cell}|S^{VE}, S^{obs}_{cell})$ in the coupling PGM graph G1 (SI Fig. S8).

In practice, one should sample and select multiple CPDs simultaneously for each coupling graph structure. Upon the selected coupling PGM graph G1 (Fig. 3, Fig. 4A), we construct a total of 1,296 metamodels with varying means and variations of four CPDs, $P(S^C_{pa}|S^{obs}_{pa}, S^{obs}_{pa})$, $P(S^C_{cell}|S^{VE}, S^{obs}_{cell})$, $P(G^C_{pl}|G^{PR}, G^{obs}_{pl})$, and $P(G^C_{cell}|G^C_{pl}, G^{obs}_{cell})$. We observe that well-constructed CPDs in the coupling PGM graph can minimize the overall surrogate model uncertainty (Fig. 4G). Additionally, the surrogate model consistency is seen to range from 0.45 to 0.73 for varying CPDs, indicating differences in the rationalization among surrogate models (Fig. 4H). Interestingly, we find an approximately negative correlation between model entropy and model consistency for the enumerated CPDs in the coupling PGM graph (Fig. 4I). Indeed, we wish to build a metamodel with the minimal uncertainty of the updated surrogate models and the maximal consistency among surrogate models (*ie*, star in Fig. 4I, parameter values in SI Table S5, plots in SI Fig. S4). For individual surrogate models, we observe a decrease in model entropy of the PR model from 68.1 to 67.8, and almost no changes the Pa model and VE model after metamodeling compared with those before metamodeling (SI Table S7). For the coupling variables, we observe a decrease in SDs of the coupling variables S^C_{pa}, G^C_{pl}, S^C_{cell} and an increase in SDs of the coupling variable G^C_{cell} after metamodeling compared with those before metamodeling (SI Table S7). Thus, among the 1,296 CPDs in coupling PGM graph G1, these changes in model entropy and model consistency before and after metamodeling guide us in selecting the optimal CPDs in the coupling PGM graph, *ie*, star in Fig. 4I.

Including observations for the model system to resolve conflicts

Conflicts between surrogate models [11] can be identified using a metamodel by comparing certain model variable distribution with its reference $\eta(X_{s,upd}, X_{ref})$ if available, model entropy $\sum h(\mathbf{X^i_{s,upd}})$ and model consistency $\bar{\eta}(\mathbf{X_{s,upd}}, \mathbf{X_s})$. We construct 121 surrogate models with varying errors $\overline{err}$ ($< P(G^{PR}_s) >$ - $< P(G^{PR}_{upd,sel}) >$, where $< P(G^{PR}_{upd,sel}) >$ is taken from the selected optimal metamodel in Fig. 4I) and variances σ^2 of G^{PR}. When we have access to the reference of the model variable S^{VE}, changes of the overlap $\eta(S^{VE}_{s,upd}, S^{VE}_{ref})$ are seen to mostly align with changes in the absolute error of G^{PR} (Fig. 5A). Furthermore, we observe a higher model entropy when the variance of G^{PR} is high even with low errors of G^{PR} (red region in Fig. 5B); a low model consistency when the error is high even with low variances (blue region in Fig. 5C). Thus, model conflicts are identified with relatively large error and variance of G^{PR} (Fig. 5A-C, point Y) in contrast to the low error and variance of G^{PR} (Fig. 5A-C, point X).

To optimize the metamodel with conflicting surrogate models, we can (i) include additional observations of certain model variables from new experimental measurements; and (ii) remove certain surrogate models that are judged not well constructed. Here, we include an observation for the intracellular glucose concentration G_{cell} as a data model G^D_{cell} into the GSIS metamodel (Fig. 5D, SI Fig. S7). As expected, this results in an optimized metamodel, with a similar model uncertainty ($h(\mathbf{X_{s,upd,Y}})$=156.4, $h(\mathbf{X_{s,upd,data}})$=156.1), a higher model consistency ($\bar{\eta}(\mathbf{X_{s,upd,Y}}, \mathbf{X_s})$=0.53, $\bar{\eta}(\mathbf{X_{s,upd,data}}, \mathbf{X_s})$=0.61, Fig. 5E), and a slightly higher model complexity ($d(\mathbf{X_{s,upd,Y}}, \mathbf{X_s})$=94, $d(\mathbf{X_{s,upd,data}}, \mathbf{X_s})$=96) compared to the metamodel with conflicts. Moreover, we extract distributions of four variables at 200 min in different surrogate models before and after metamodeling: the intracellular glucose concentration G^{VE}, the insulin secretion rate of one β-cell S^{VE}, the pancreatic insulin secretion rate S^{Pa}_{Pa}, and the plasma insulin concentration I^{PR}. In the optimized metamodel (red solid curve in Fig. 5F), distributions of three variables (G^{VE}, S^{VE}, S^{Pa}_{Pa}) exhibit lower uncertainties after metamodeling and higher overlaps before and after metamodeling, while opposite changes are observed for I^{PR} (Fig. 5F). Till here, the metamodel is further optimized by resolving conflicts by including observations of model variables as an additional data model.

Comparison of optimized and non-optimized metamodel

We optimize the GSIS metamodel following the Bayesian *Occam's razor* rationale as outlined above. The optimized metamodel (star in Fig. 4I, model complexity $d = 94$) outperforms the non-optimized one (constructed using the non-optimized surrogate model 2 in Fig. 2 and the non-optimized CPD $P(G^C_{pl}|G^{PR}, G^{obs}_{pl})$ in the red region in Fig. 4H, model complexity $d = 94$) in three aspects. First, for the PR input model, the uncertainty of

the surrogate model $\sum h(\mathbf{X_s^i})$ decreases from 68.2 to 67.8, while the consistency between the input and surrogate model $\eta(\mathbf{X_s^i}, \mathbf{X_{in}^i})$ increases from 0.30 to 0.37 (Fig. 2B). Second, the surrogate model entropy $\sum h(\mathbf{X_{s,upd}^i})$ decreases from 159.6 to 156.7, while the consistency among the surrogate models $\bar{\eta}(\mathbf{X_{s,upd}}, \mathbf{X_s})$ increases from 0.52 to 0.63 (Fig. 4A). Third, the mean fold change of I^{PR} between normal and T2D subjects predicted by the optimized metamodel (1.64) demonstrates an overall improved agreement with the experimental measurements (1.74) [53] compared to the non-optimized metamodel (1.24) (Fig. 6). Therefore, we have found the optimal metamodel among the candidate ones and validated it using external experimental measurements.

Discussion

Modeling complex systems inherently involves high levels of complexity due to the integration of numerous input information and sources of uncertainty. Taking metamodeling as an example, we present the probability calculus using an analytical example to establish definitive rules for the probability propagation from input models to the metamodel and propose quantitative assessments for it (Fig. 1). Based on these, we have developed a method to optimize models for complex systems by: (i) minimizing model uncertainty, (ii) maximizing consistency between input and surrogate models to ensure the reproduction effectiveness in the conversion stage, as well as among surrogate models to rationalize them in the coupling stage; and (iii) minimizing the model complexity, following the Bayesian *Occam's Razor* rationale. We showcase its benefits in the selection of surrogate models and coupling PGM graphs (Fig. 2-4), as well as including observations of model variables as an additional data model to resolve conflicts among surrogate models (Fig. 5). This approach demonstrates better alignment with experimental observations (Fig. 6) and facilitates the construction of an optimal model that achieves maximal accuracy and precision for complex systems with sufficient simplicity, given the accuracy and precision of input information.

Moreover, facilitated by our method, the iterative nature of various integrative modeling approaches provides a dynamic framework for continuous optimization. If the optimal model is not found or additional input information becomes available, we revisit the construction stages iteratively, aiming to continuously improve its accuracy and precision while ensuring simplicity [54, 11]. For example, in metamodeling, the iterative selection of the surrogate model can be achieved by assessing the uncertainty, consistency, and complexity when converting the input model into a surrogate model, as well as evaluating the updated surrogate models in the backpropagation stage; the iterative selection of the coupling PGM graph can be achieved by assessing the uncertainty of surrogate models and the consistency among them in the backpropagation stage while considering that model complexity should be maintained at an acceptable level. New experiments may be conducted to reveal conflicts in the model, leading to potential adjustments in assumptions when constructing a model (*eg*, simplifying statistical dependencies among model variables). The optimized model is expected to provide new insights into the system and guide experiment design, which in turn drives further optimization.

Although this method is appealing, several aspects require further development to achieve model optimization in an automated manner. First, we ought to design standard scoring functions for the evaluation of the modeling process. This could potentially be achieved by combining the weighted sum of model uncertainty, model consistency, and model complexity, along with other criteria of the modeler's interest (*eg*, model accuracy when the ground truth or gold standard is available) [55, 56]. Second, with a standard scoring function in place, we will be able to conduct exhaustive sampling to thoroughly search the space of alternative models and thus avoid specific local minima (*eg*, stochastic sampling [54], grid search [57]). Third, we should extend the algorithms to accommodate other types of distributions of model variables. For instance, several Bayesian filters can be incorporated to allow for inference of non-Gaussian models, including Kalman filters for linear transition functions, extended Kalman filters for nonlinear transition functions, and particle filters for nonlinear non-gaussian processes [58, 59, 60]. We anticipate that our method following the Bayesian *Occam's razor* rationale will find applications in the construction and optimization of modeling for diverse complex systems, thereby further enhancing our interpretation and prediction capabilities.

Methods

The Python package for our method, along with the models, data files, and examples for the present work are available at https://github.com/salilab/metamodeling/tree/optimization. The metamodel was implemented

using our previous metamodeling methods https:/github.com/salilab/metamodeling/. For more details, see SI Appendix

Acknowledgments

General: We express our gratitude to Andrej Sali for his valuable advice and fruitful discussions. We acknowledge helpful comments by Jared Sagendorf, Eli Draizen, and Neelesh Soni. We are grateful to all members of the Pancreatic ß-Cell Consortium for providing the context in which this research was performed. **Funding:** B.R. was supported by a Minerva Center grant on Cell Intelligence. C.W., J.H.Z., J.J.Z., X.H. and L.S. were supported by ShanghaiTech University. We acknowledge the Shanghai Frontiers Science Center for Biomacromolecules and Precision Medicine at ShanghaiTech University, and the High-Performance Computing (HPC) Platform of ShanghaiTech University for computational resources. **Author contributions:** L.S. and X.H designed and supervised the paper, C.W., J.H.Z., J.J.Z performed the research and wrote the manuscript. B.R. and J.S. supervised the project and provided inputs in the manuscript. **Competing interests:** The authors declare no competing financial interest. **Data and materials availability:** The data that support the plots within this paper and other findings of this study are available online.

References

1. Hancock, M. *et al.* Integration of software tools for integrative modeling of biomolecular systems. *Biophysical Journal* **121**, 31a (2022).
2. Vajda, S., Hall, D. R. & Kozakov, D. Sampling and scoring: A marriage made in heaven. *Proteins: Structure, Function, and Bioinformatics* **81**, 1874–1884 (2013).
3. Rout, M. P. & Sali, A. Principles for integrative structural biology studies. *Cell* **177**, 1384–1403, DOI: 10.1016/j.cell.2019.05.016 (2019).
4. Karaca, E. & Bonvin, A. M. Advances in integrative modeling of biomolecular complexes. *Methods* **59**, 372–381 (2013).
5. Ghaemi, Z., Peterson, J. R., Gruebele, M. & Luthey-Schulten, Z. An in-silico human cell model reveals the influence of spatial organization on rna splicing. *PLoS computational biology* **16**, e1007717 (2020).
6. Macklin, D. N. *et al.* Simultaneous cross-evaluation of heterogeneous e. coli datasets via mechanistic simulation. *Science* **369**, eaav3751 (2020).
7. Ma, J. *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nature methods* **15**, 290–298 (2018).
8. Vangheluwe, H. & de Lara, J. Meta-models are models too. In *Proceedings of the Winter Simulation Conference*, vol. 1, 597–605 vol.1, DOI: 10.1109/WSC.2002.1172936 (2002).
9. Syriani, E. *et al.* Atompm: A web-based modeling environment. In *Joint proceedings of MODELS'13 Invited Talks, Demonstration Session, Poster Session, and ACM Student Research Competition co-located with the 16th International Conference on Model Driven Engineering Languages and Systems (MODELS 2013): September 29-October 4, 2013, Miami, USA*, 21–25 (2013).
10. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).
11. Raveh, B. *et al.* Bayesian metamodeling of complex biological systems across varying representations. *Proceedings of the National Academy of Sciences* **118**, DOI: 10.1073/pnas.2104559118 (2021).
12. Wang, G. G. & Shan, S. Review of metamodeling techniques in support of engineering design optimization. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 4255, 415–426 (2006).
13. Zhong, H. *et al.* Metamodeling for policy simulations with multivariate outcomes. *Medical Decision Making* **42**, 872–884 (2022).
14. Sali, A. From integrative structural biology to cell biology. *Journal of Biological Chemistry* **296**, 100743, DOI: 10.1016/j.jbc.2021.100743 (2021).
15. Draper, D. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **57**, 45–70 (1995).
16. Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379–423 (1948).
17. Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. Occam's razor. *Information processing letters* **24**, 377–380 (1987).
18. Wolpert, D. H. The relationship between occam's razor and convergent guessing. *Complex Systems* **4**, 319–368 (1990).
19. Monsalve-Bravo, G. M. *et al.* Analysis of sloppiness in model simulations: Unveiling parameter uncertainty when mathematical models are fitted to data. *Science advances* **8**, eabm5952 (2022).
20. Thompson, K. M., Burmaster, D. E. & Crouch3, E. A. Monte carlo techniques for quantitative uncertainty analysis in public health risk assessments. *Risk Analysis* **12**, 53–63 (1992).
21. Viswanath, S. & Sali, A. Optimizing model representation for integrative structure determination of macromolecular assemblies. *Proceedings of the National Academy of Sciences* **116**, 540–545 (2019).

22. Rieping, W., Nilges, M. & Habeck, M. Isd: a software package for bayesian nmr structure calculation. *Bioinformatics* **24**, 1104–1105 (2008).

23. Zhou, Z.-H. *Ensemble methods: foundations and algorithms* (CRC press, 2012).

24. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* **55**, 119–139 (1997).

25. Change, I. Climate change 2007: The physical science basis. *Agenda* **6**, 333 (2007).

26. Sharpe, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance* **19**, 425–442 (1964).

27. Frangopol, D. M. & Maute, K. Reliability-based optimization of civil and aerospace structural systems. In *Engineering design reliability handbook*, 559–590 (CRC Press, 2004).

28. Gull, S. F. Bayesian inductive inference and maximum entropy. In *Maximum-entropy and bayesian methods in science and engineering: foundations*, 53–74 (Springer, 1988).

29. Jeffreys, H. *The theory of probability* (OuP Oxford, 1998).

30. Woodard, R. & Stein, M. L. Interpolation of spatial data: Some theory for kriging. *Technometrics* **42**, 436, DOI: 10.2307/1270967 (2000).

31. Deng, H., Shao, W., Ma, Y. & Wei, Z. Bayesian metamodeling for computer experiments using the gaussian kriging models. *Quality and Reliability Engineering International* **28**, 455–466, DOI: 10.1002/qre.1259 (2011).

32. Inman, H. F. & Bradley, E. L. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods* **18**, 3851–3874, DOI: 10.1080/03610928908830127 (1989).

33. Bonomi, M. *et al.* Bayesian weighing of electron cryo-microscopy data for integrative structural modeling. *Structure* **27**, 175–188 (2019).

34. Erdelyi, J.-F., Amblard, F., Gaudou, B., Kaddoum, E. & Verstaevel, N. Exploration of model coupling strategies in a hybrid agent-based traffic simulation. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, 153–167 (Springer, 2021).

35. Puga, J. L., Krzywinski, M. & Altman, N. Bayesian networks. *Nature Methods* **12**, 799–800, DOI: 10.1038/nmeth.3550 (2015).

36. Pastore, M. & Calcagnì, A. Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology* **10**, DOI: 10.3389/fpsyg.2019.01089 (2019).

37. Dawid, A. P., Kjærulff, U. & Lauritzen, S. L. *Hybrid propagation in junction trees*, 85–97 (Springer Berlin Heidelberg, 1995).

38. Boyen, X. & Koller, D. Approximate learning of dynamic models. *Advances in neural information processing systems* **11** (1998).

39. Rubinstein, R. Y. & Kroese, D. P. *Simulation and the Monte Carlo Method* (Wiley, 2016).

40. Djuric, P. M. & Bugallo, M. F. Particle filtering for multivariate state-space models. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, DOI: 10.1109/acssc.2012.6489028 (IEEE, 2012).

41. Lombardi, L. E., Martí, M. A. & Capece, L. Cg2aa: backmapping protein coarse-grained structures. *Bioinformatics* **32**, 1235–1237, DOI: 10.1093/bioinformatics/btv740 (2015).

42. Eisenhammer, T., Hübler, A., Packard, N. & Kelso, J. A. S. Modeling experimental time series with ordinary differential equations. *Biological Cybernetics* **65**, 107–112, DOI: 10.1007/bf00202385 (1991).

43. Liang, H. & Wu, H. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association* **103**, 1570–1583 (2008).

44. Chen, S., Shojaie, A. & Witten, D. M. Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association* **112**, 1697–1707, DOI: 10.1080/01621459.2016.1229197 (2017).

45. Cowell, R. G., Dawid, P., Lauritzen, S. L. & Spiegelhalter, D. J. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks* (Springer Science & Business Media, 2007).

46. Arnborg, S., Corneil, D. G. & Proskurowski, A. Complexity of finding embeddings in ak-tree. *SIAM Journal on Algebraic Discrete Methods* **8**, 277–284 (1987).

47. Chandrasekaran, V., Srebro, N. & Harsha, P. Complexity of inference in graphical models. *arXiv preprint arXiv:1206.3240* (2012).

48. Dalla Man, C., Rizza, R. A. & Cobelli, C. Meal simulation model of the glucose-insulin system. *IEEE Transactions on Biomedical Engineering* **54**, 1740–1749, DOI: 10.1109/tbme.2007.893506 (2007).

49. Dagum, P., Galper, A. & Horvitz, E. *Dynamic Network Models for Forecasting*, 41–48 (Elsevier, 1992).

50. White, K. L. *et al.* Visualizing subcellular rearrangements in intact cells using soft x-ray tomography. *Science Advances* **6**, DOI: 10.1126/sciadv.abc8262 (2020).

51. Fehr, M., Lalonde, S., Lager, I., Wolff, M. W. & Frommer, W. B. Imaging of the dynamics of glucose uptake in the cytosol of cos-7 cells by fluorescent nanosensors. *Journal of Biological Chemistry* **278**, 19127–19133, DOI: 10.1074/jbc.m301333200 (2003).

52. Kinory, S., Freedman, D., Pisani, R. & Purves, R. Statistics. *Technometrics* **21**, 273, DOI: 10.2307/1268533 (1979).

53. Yalamanchi, A. *et al.* lt;pgt;assessment of insulin sensitivity and secretion in patients with fibrocalculous pancreatic diabeteslt;/pgt;. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* **Volume 12**, 779–788, DOI: 10.2147/dmso.s204254 (2019).

54. Viswanath, S., Chemmama, I. E., Cimermancic, P. & Sali, A. Assessing exhaustiveness of stochastic sampling for integrative modeling of macromolecular structures. *Biophysical journal* **113**, 2344–2353 (2017).

55. Hoege, M., Guthke, A. & Nowak, W. Bayesian model weighting: The many faces of model averaging. *Water* **12**, 309 (2020).

56. Burnham, K. P. & Anderson, D. R. *Model selection and multimodel inference: a practical information-theoretic approach* (Springer, 2002).

57. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *Journal of machine learning research* **13** (2012).

58. Julier, S. J. & Uhlmann, J. K. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE* **92**, 401–422 (2004).

59. Carpenter, J., Clifford, P. & Fearnhead, P. Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation* **146**, 2–7 (1999).

60. Julier, S. J. & Uhlmann, J. K. New extension of the kalman filter to nonlinear systems. In *Signal processing, sensor fusion, and target recognition VI*, vol. 3068, 182–193 (Spie, 1997).

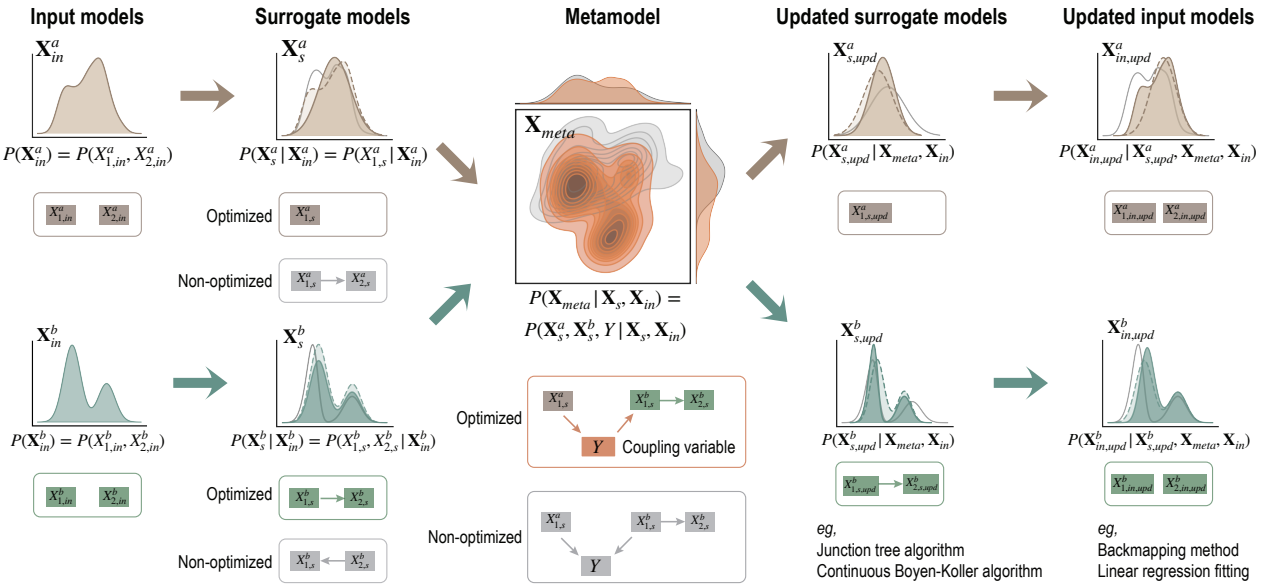
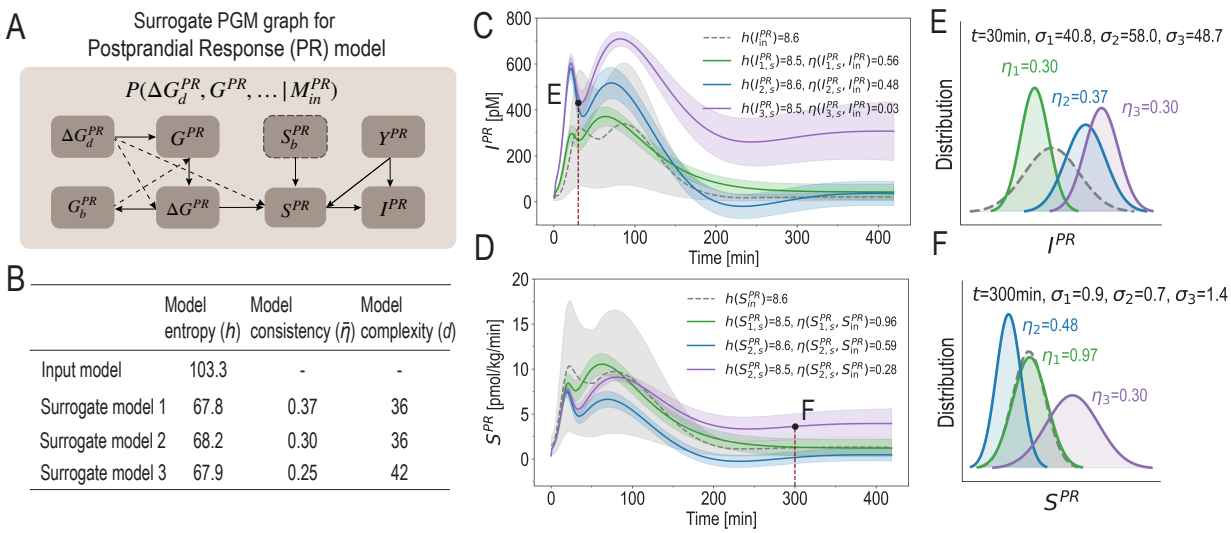
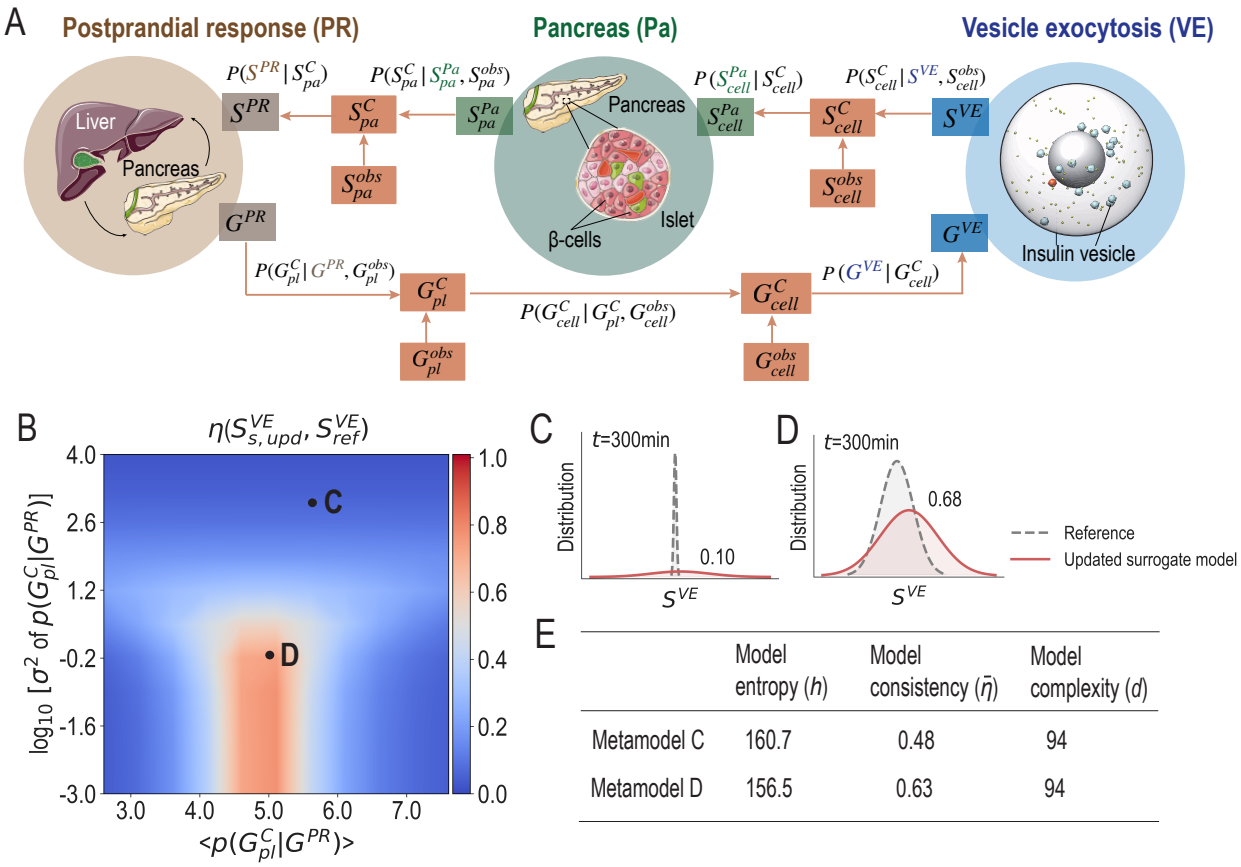


Fig. 1. Overview of probability propagation in metamodeling. Input models a (khaki) with its distribution $P(\mathbf{X}^a_{in})$ and b (green) with its distribution $P(\mathbf{X}^b_{in})$, each consisting of two variables. Surrogate model a comprises one variable, with its distribution $P(\mathbf{X}^a_s|\mathbf{X}^a_{in})$; surrogate model b comprises two variables, with its distribution $P(\mathbf{X}^a_s|\mathbf{X}^a_{in})$. The metamodel constructed using the head-to-tail coupling PGM graph, with its distribution $P(\mathbf{X}_{meta}|\mathbf{X}_s, \mathbf{X}_{in})$. Updated surrogate models a and b obtained using different algorithms for model inference during backpropagation, with their distributions $P(\mathbf{X}^a_{s,upd}|\mathbf{X}_{meta}, \mathbf{X}_{in})$, $P(\mathbf{X}^b_{s,upd}|\mathbf{X}_{meta}, \mathbf{X}_{in})$, respectively. Updated input models a and b obtained using different methods during backpropagation, with their distributions $P(\mathbf{X}^a_{in,upd}|\mathbf{X}^a_{s,upd}, \mathbf{X}_{meta}, \mathbf{X}_{in})$, $P(\mathbf{X}^b_{in,upd}|\mathbf{X}^b_{s,upd}, \mathbf{X}_{meta}, \mathbf{X}_{in})$, respectively. Thick arrows indicate the propagation of uncertainty across metamodeling stages. Colored solid and dashed lines depict the distributions of optimized models between adjacent stages, while grey solid lines depict distributions of non-optimized ones in each stage. Colored model PGM graphs represent optimized models, while grey ones represent non-optimized ones.



	Model entropy (h)	Model consistency ($\bar{\eta}$)	Model complexity (d)
Input model	103.3	-	-
Surrogate model 1	67.8	0.37	36
Surrogate model 2	68.2	0.30	36
Surrogate model 3	67.9	0.25	42

Fig. 2. Probability propagation from input model to surrogate model. (A) Surrogate PGM graph for the Postprandial response (PR) input model. Frames indicate surrogate model variables, while dashed frame indicates time-independent variables, dashed edges indicate additional edges in surrogate model 3. **(B)** Table for model entropy $h(\mathbf{X_s^{PR}})$, model consistency $\bar{\eta}(\mathbf{X_s^{PR}})$ and model complexity d of two computed surrogate models. **(C-D)** Time courses of variables I^{PR} and S^{PR} in the input model (dashed gray line), surrogate model 1 (solid green line), surrogate model 2 (solid blue line), and surrogate model 3 (solid purple line). Dashed or solid lines represent the mean values, while shaded areas represent the standard deviations. $\eta(I_s^{PR}, I_{in}^{PR})$ and $\eta(S_s^{PR}, S_{in}^{PR})$ represent the average overlaps of the marginal distributions of the aligned variables in input and surrogate models over 420 min. **(E-F)** Distributions of I^{PR} and S^{PR} at 30 min. η denotes the overlap of the marginal distributions in input and surrogate models, while σ denotes the standard deviation of the variable marginal distribution.



	Model entropy (h)	Model consistency ($\bar{\eta}$)	Model complexity (d)
Metamodel C	160.7	0.48	94
Metamodel D	156.5	0.63	94

Fig. 3. Optimization of metamodeling by improving the selection of coupling PGM graphs with reference. (A) Schematic illustration of the numerical metamodel. The model PGM graphs of the Postprandial Response (PR), Pancreas (Pa), and Vesicle Exocytosis (VE) models are shown in khaki, green, and blue, respectively. The coupling PGM graphs are shown in orange. **(B)** The distribution overlaps between the updated surrogate model variable $S^{VE}_{s,upd}$ and the reference variable S^{VE}_{ref} as a function of the means (x-axis) and variances (y-axis) of the CPD $P(G^C_{pl}|G^{PR}, G^{obs}_{pl})$ in one selected metamodel in A. **(C-D)** Comparison of the reference and updated surrogate model variable distribution at 100 min and 300 min for two different metamodels in B. **(E)** Table of model entropy, model consistency and model complexity of the metamodels selected in B.

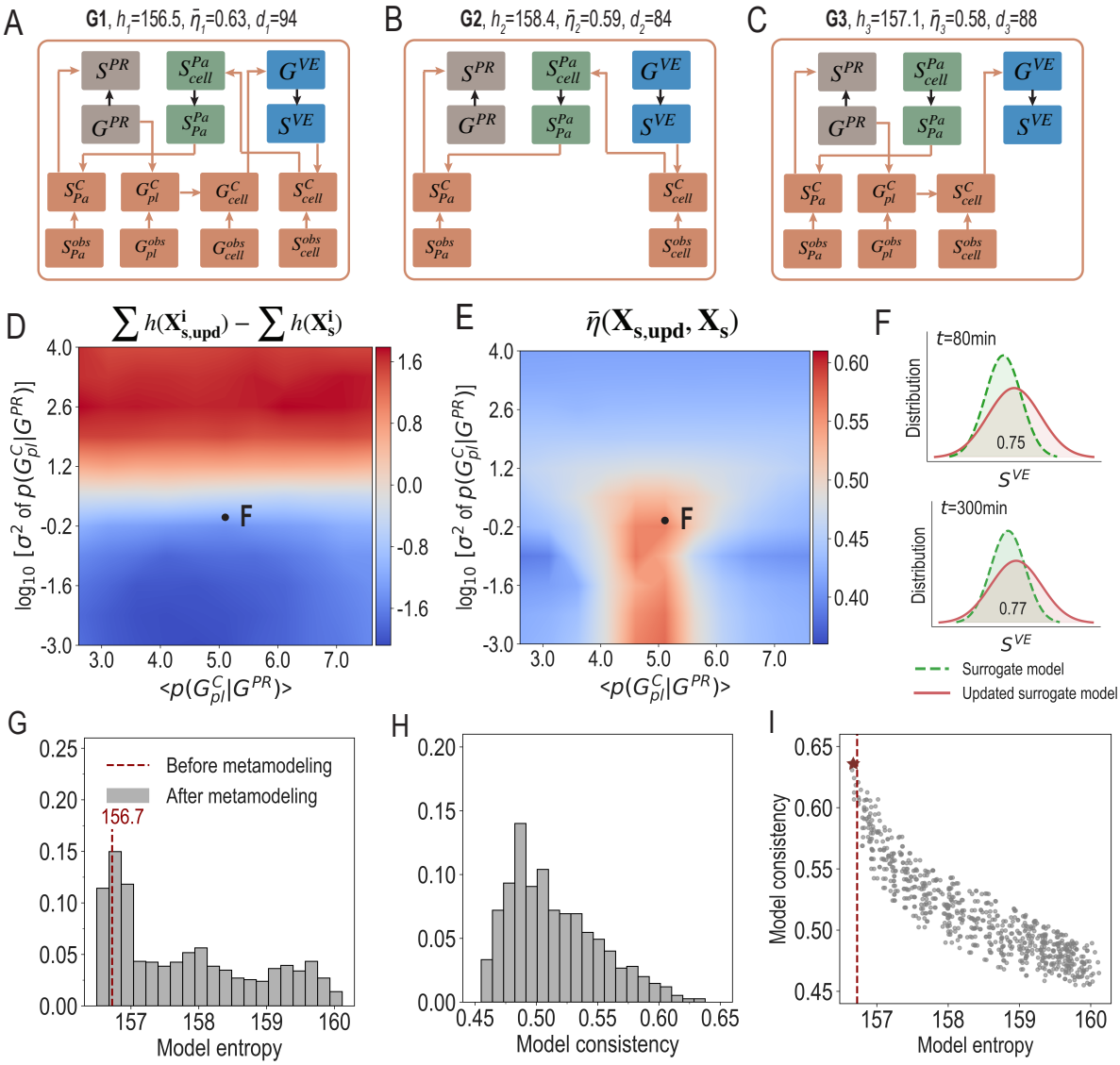
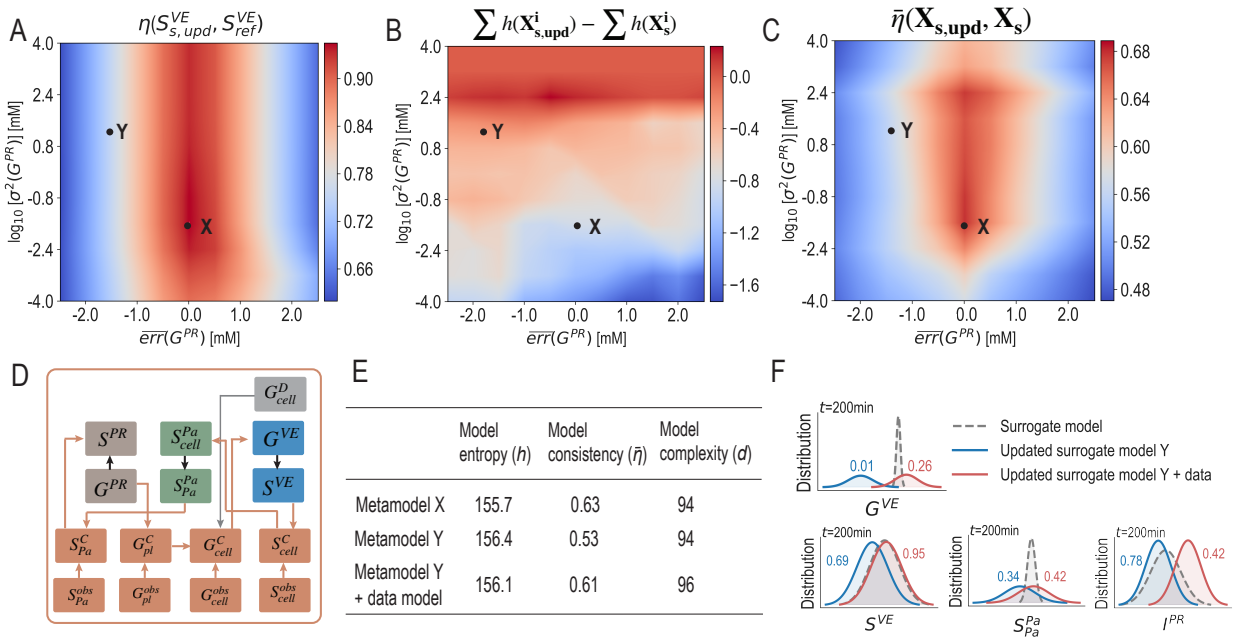


Fig. 4. Optimization of metamodeling by improving the selection of coupling PGM graphs without reference. (A-C) Three coupling PGM graphs. Orange nodes and edges represent coupling variables and CPDs in the coupling PGM graphs. Model entropy and model consistency values are shown in figure titles. **(D)** Difference between model entropy before $\sum h(\mathbf{X}^{\mathbf{i}}_{\mathbf{s}})$ and after $\sum h(\mathbf{X}^{\mathbf{i}}_{\mathbf{s,upd}})$ metamodeling as a function of the means (x-axis) and variances (y-axis) of the CPD $P(G^C_{pl}|G^{PR})$ in coupling PGM graph G1. **(E)** Model consistency $\bar{\eta}(\mathbf{X}_{\mathbf{s,upd}}, \mathbf{X}_{\mathbf{s}})$ as a function of the same sets of means (x-axis) and variances (y-axis) in D. Values in D-E are computed over the metamodel timespan (420 min). **(F)** Comparison of the model variable S^{VE} distribution before and after metamodeling at 80 min and 300 min. **(G)** Distribution of model entropy after metamodeling with varying CPDs of $P(S^C_{pa}|S^{Pa}_{pa}, S^{obs}_{pa})$, $P(S^C_{cell}|S^{VE}, S^{obs}_{cell})$, $P(G^C_{pl}|G^{PR}, G^{obs}_{pl})$, and $P(G^C_{cell}|G^C_{pl}, G^{obs}_{cell})$). The sum of the entropy of three surrogate models before metamodeling is shown as the red dashed line. **(H)** Distribution of model consistency after metamodeling with the same set of CPDs as in G. **(I)** Change of model consistency along with the change of model entropy across all metamodels. The selected coupling PGM graph is shown as the red star.



	Model entropy (h)	Model consistency ($\bar{\eta}$)	Model complexity (d)
Metamodel X	155.7	0.63	94
Metamodel Y	156.4	0.53	94
Metamodel Y + data model	156.1	0.61	96

Fig. 5. Optimize metamodeling by solving model conflicts. (A) The distribution overlaps between the updated surrogate model variable $S^{VE}_{s,upd}$ and the reference variable S^{VE}_{ref} as a function of the errors (x-axis) and logarithm variances (y-axis) of the surrogate model variable G^{PR}. **(B)** Difference of entropy before and after metamodeling as a function of the same sets of errors (x-axis) and logarithm variances (y-axis) in A. **(C)** Model consistency between surrogate and updated surrogate models as a function of the same sets of errors (x-axis) and logarithm variances (y-axis) in A. **(D)** Coupling PGM graph with an additional data model. **(E)** Table of model entropy, model consistency and model complexity of the selected metamodels (point X and point Y) in B-C and the metamodel with additional data model. **(F)** Comparison of surrogate model variable distributions at 200 min before (gray) and after metamodeling with (red) and without (blue) an additional data model.

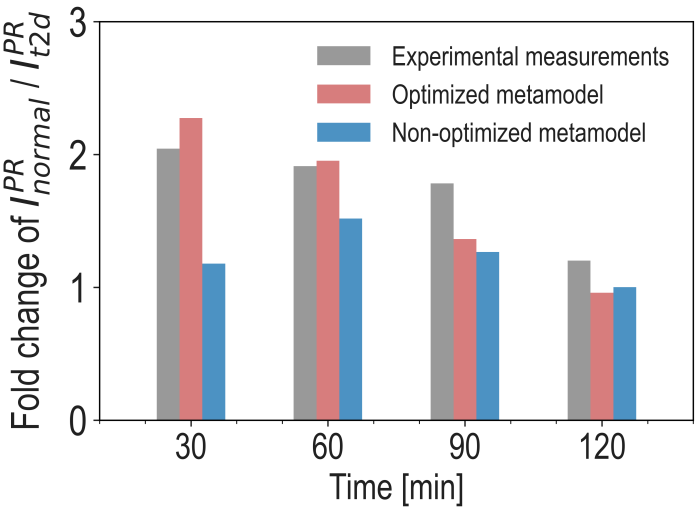


Fig. 6. Comparison of the optimized and non-optimized metamodel. Fold changes of the selected model variable I^{PR} between normal and T2D subjects in the GSIS metamodel. Experimental measurements from 36 normal subjects and 36 T2D subjects are shown in grey. The updated surrogate model variable computed from the optimized and non-optimized metamodels are shown in red and blue, respectively.