A systematic framework for biomolecular system identification

Z. A. Tuza[1], L. Bandiera[2,3], D. Gomez-Cabeza[2], G. B. Stan[1] and F. Menolascina[2,3]

***Abstract*— Unwieldy challenges, such as emergent behaviour or lack of compositionality, hamper the rational engineering of synthetic biomolecular systems. The use of mathematical models would address many of these challenges, yet current practices in synthetic biology make obtaining them time and resource intensive. More importantly, in many cases, the process of obtaining data for mathematical models is guided by intuition rather than rigorous modelling requirements.**

To make better use of the available resources, this tutorial proposes an end-to-end framework for biomolecular system identification. Given a biomolecular system inside a biological organism, the framework leverages on system identification techniques to automate the initial proposition of candidate models for the system of interest. Then, statistical methods guide the optimal design of experiments that enable discrimination and calibration of the most plausible model candidates for the underlying system.

We foresee that, by establishing a closed-loop between data-gathering and model identification, the outlined approach will accelerate and standardise computational modelling of biomolecular systems.

I. Introduction

To solve the key challenges in synthetic biology, we need to have the capability to regulate quantities of interest (e.g., biomolecules amounts) inside cells [1], [2]. Such capability can be gained by, for example, designing internal or external feedback loops [3], [4].

Successful feedback control approaches in synthetic biology show that model-based controller design is necessary to achieve the desired closed-loop behaviour [5], [6]. It is also evident that identifying models for biomolecular systems is challenging, and novel system identification approaches are needed to tackle this challenge [1].

This tutorial presents an end-to-end framework where experimental design, data collection and model identification are executed in an automated fashion. Such a framework allows users to fully harness the potential of lab automation and eventually to more efficiently tackle the complexity of biomolecular systems. The key components of such a framework are the algorithmically guided design of experiments and the automated model structure identification from the measurement data [5], [7].

[1]Bioengineering Department, Imperial College London, London, SW7 2AZ, UK. z.tuza@imperial.ac.uk, g.stan@imperial.ac.uk

[2]School of Engineering, Institute for Bioengineering, The University of Edinburgh, Edinburgh, EH9 3DW, UK. lucia.bandiera@ed.ac.uk, david.gomez@ed.ac.uk, filippo.menolascina@ed.ac.uk

[3]SynthSys - Centre for Synthetic and Systems Biology, The University of Edinburgh, Edinburgh, EH9 3BF.

These two key components are often neglected in biomolecular system identification because of the computational demand of the underlying algorithms. However, recent surveys in system identification highlight experiment design as a vital step [8], [9]. An automated model structure identification procedure is also crucial because as the amount of available experimental data increases, building manually-curated models becomes impractical.

It should be emphasised that, while this tutorial focuses on one set of tools and metrics, individual components can be replaced with available alternatives to adapt or improve the performance of the overall framework. Furthermore, the presented set of tools and metrics can be applied individually to gain a better understanding of certain aspects of a biomolecular model.

The paper structure follows the flowchart presented in Fig. 3. Each section, after the preliminaries, introduced in Section II, corresponds to a main step within the general framework of Fig. 3.

II. Preliminaries

A. Model Class

The model class for this tutorial is the class of nonlinear nonnegative systems, which is given as

$$\begin{aligned} \dot{\mathbf{x}} &= f(\mathbf{x}, u, \boldsymbol{\theta}), \quad \mathbf{x}(0) = \mathbf{x}_0(\boldsymbol{\theta}) \in \mathbb{R}^n_+ \\ \mathbf{y} &= h(\mathbf{x}, \boldsymbol{\theta}), \end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n_+$ denotes the states of the model, usually molecular counts or concentrations, hence the nonnegativity requirement of the states. To have nonnegative states, the vector field $f(\mathbf{x}, u, \boldsymbol{\theta})$ has to be essentially nonnegative [10].

There are different levels of modelling details that can be captured in the modelling process, e.g. from Chemical Master Equations to rule-based models [2]; however, for model identification, the class of nonnegative systems has the many tools readily available.

B. Data measurement

The structure of $h(\mathbf{x}, \boldsymbol{\theta})$ depends on the device(s) available for measurements. In most cases, biological signals or genetic activity is observed via fluorescent reporters—proteins that emit light at a specific wavelength after excitation at a different wavelength.

Some of the devices typically used in wetlabs are plate readers (which provide time-series data), chemostats/turbidostats (which correspond to a miniature bioreactor where cellular density and/or growth conditions are kept constant), flow cytometers (which provide snapshot

measurements from hundreds of thousands of cells [11]), and microfluidics combined with video microscopy (i.e. an array of microfluidic chambers observed in a camera's field of view). Each device comes with its own capabilities, which defines $h(\mathbf{x}, \boldsymbol{\theta})$, see e.g. Section VI in [3].

During the model identification process, our goal is to characterize $f(\mathbf{x}, u, \boldsymbol{\theta})$ (and sometimes $h(\mathbf{x}, \boldsymbol{\theta})$ as well) by executing experimental planning techniques introduced in Section VI.

C. Input design

In order to manipulate the state of a biomolecular system, we have two main options: either we can change the initial values of the states, or we can design input signals, i.e. perturbations to the system. The latter one is more suitable for a wide range of experiments because having a tight, known control on the initial condition might not be feasible in a given experimental setup.

In terms of input, there are several types of molecular mechanisms that can be used to change the internal state of a biomolecular system: mostly chemical inducers or blockers; they diffuse through the cell membrane and enhance or reduce gene expression. A downside of these regulatory molecules is that they need to be washed out from the sample to decrease the input signal, which requires a special device setup, e.g. microfluidic chambers, chemostats or turbidostats.

An optimal way to design input signals for (biomolecular) systems is discussed in Section V.

D. The future of biomolecular System identification

There are many open challenges in biomolecular identification, which are stemming from the complexity of the problem and the lack of biosensors. However, the combination of lab automation, precise gene editing using CRISPR, and decreasing the cost of DNA synthesis, combined with systems-and-control-theory-guided experiments opens exciting novel avenues to tackle these challenges.

For example, the state-of-the-art system identification techniques are frequency-domain based and typically use multi-sine input signals [9]. At the cellular level, we are currently ill-equipped for the easy generation of multi-sine perturbations. One promising approach for the generation of dynamic perturbations, including potentially multi-sine perturbations, is optogenetics (where light induced protein conformation change affects gene expression) [12].

III. Automatic generation of dynamic models

A. Estimating model structures from time-series data

Estimating (learning) model structure from time-series or steady-state data has been extensively studied in the literature, e.g. in [13], [14], [15]. Some reports formulate the problem as an integer programming or a multi-objective optimisation problem where one needs to find a Pareto optimal solution. [16].

Recently, another approach has emerged where the model structure estimation problem is formulated as an (undetermined) linear regression problem. If the resulting problem is underdetermined, then it has an infinite number of solutions. However, the problem can be regularised by introducing a penalty on model size/complexity, which then leads to solutions given in terms of sparest model structures [17].

Nonlinear ODE models in biology typically have few right-hand side terms to capture the dynamics. For example, a Lotka-Volterra system, which has two equations and two right-hand terms per equations, see e.g. [15]. For the right choice of parameters, the dynamics, encoded by these four terms in two equations, has a stable limit cycle, an explicitly nonlinear feature. Another typical example is the Lorenz attractor, which has two terms per equations and exhibits chaotic dynamics. Such examples indicate that sparse models are capable of capturing complex dynamical behaviours.

Typically, the dynamics of biomolecular systems can be captured using mass action kinetics, Michaelis–Menten kinetics, or Hill kinetics [2]. Mass action kinetics leads to right-hand-side ODEs where the parameters appear as factors in a linear combination of nonlinear (i.e. product) terms. An appealing feature of mass action kinetics is that it is linear in parameters, so all the parameters can be estimated in the framework presented next.

B. Estimating both ODE structure and parameters

Let us assume that the models under consideration can be written as:

$$\dot{x}_i = \phi(\mathbf{x})\bar{\Theta}_{\cdot,i}, \qquad i = 1, \ldots, n, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^n_+$ is a vector of state variables, $\phi(\mathbf{x}) : \mathbb{R}^n_+ \to \mathbb{R}^m_+$ is a row vector of nonlinear functions (representing the possible kinetics), and the parameter vector for each state are the columns of $\bar{\Theta} \in \mathbb{R}^{m \times n}$. In this paper, we only consider initial value problems, thus $\mathbf{x}(0) = \mathbf{x}_0$ is also given. The solution of the differential equation is denoted by $\mathbf{x}(t)$.

We chose to show the model equation as in (2) to emphasize the fact that each state equation might depend on all other state variables. Furthermore, we assume that we can measure all of the state variables (full state observability). The measurement matrix is then formulated as

$$X_{i,j} = x_j(t_i), \ i = 1, \ldots, p, \ j = 1, \ldots, n, \tag{3}$$

where, t_i is the time of the measurement at sampling time point i, and we collect p samples for each state variable.

The regressor matrix $\mathbf{A}$ in Fig 1. can be computed by evaluating $\phi(\mathbf{x})$ on the measurement data, namely

$$A_{i,j} = \phi_j(X^\top_{i,\cdot}), \quad i = 1, \ldots, p, \quad j = 1, \ldots, m. \tag{4}$$

In (4), we assume that $m \gg p$, i.e. we have many more right-hand-side terms, called "candidate functions", than data point, thus $\mathbf{A}$ is rank deficient.

For example, based on domain knowledge, one can assemble a set of candidate functions as polynomials:

$$\phi(\mathbf{x}) = \begin{bmatrix} x_1 & x_2 & x_1x_2 & x_1^2 & x_2^2 & x_1^2x_2 & x_1x_2^2 \end{bmatrix}. \tag{5}$$

The vector of functions in (5) represents a set of candidate functions for a two-state chemical reaction network with mass action kinetics.

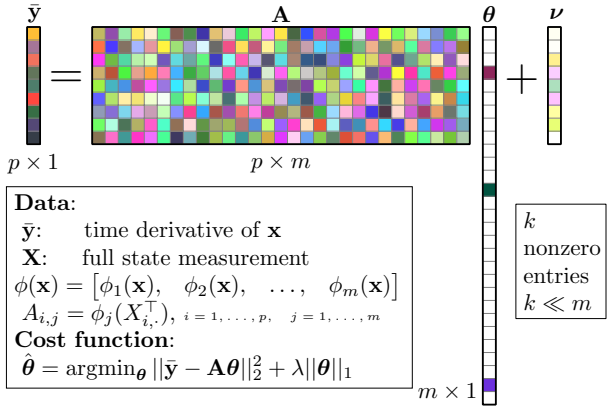

Fig. 1. For each state variable, measurements are stored in $\mathbf{X}$ and the time derivative of $\mathbf{x}$, denoted $\bar{\mathbf{y}}$, is calculated. The set of candidate functions are stored in ϕ, evaluated on $\mathbf{X}$, and stored in $\mathbf{A}$. Solution to the minimisation problem define the returned model structure and correspond to the sparest $\boldsymbol{\theta}$ which fit the linear regression problem at best.

Additionally, we can compute the time derivative of $\mathbf{x}(t)$ as

$$\bar{y}_i(t) = \frac{d}{dt}x_i(t), \quad i = 1, \ldots, n. \tag{6}$$

There are multiple ways to numerically approximate the time derivative from measurement data, see e.g. [18], [19], [20].

Then, for each state, the parameter vector $\boldsymbol{\theta}$ can be estimated as a solution of the minimisation problem described in Fig. 1, and the results collected into $\hat{\bar{\Theta}}$. The estimated model is then given as

$$\dot{\tilde{x}}_i = \phi(\tilde{\mathbf{x}})\hat{\bar{\Theta}}_{\cdot,i}, \qquad i = 1, \ldots, n. \tag{7}$$

It is important to point out that for this model-structure estimation process to be generate biologically relevant models, domain-specific constraints need to be added to the estimation process. These constraints can be used to enforce non-negative states, bounded trajectories or stability [21].

The above presented estimation procedure is implemented in the `ODE composer` toolbox [22].

C. Generating set of candidate models

The next step of the framework needs a set of candidate models. For this reason, the candidate models are generated using the residual error and number of the non-zero candidate functions, see Fig. 2. Ahead of model selection, these candidate model structures are ranked in the preliminary analysis, as discussed in the next section.

IV. Preliminary analyses of the characterized structures

Once plausible network structures have been defined, we move to the next stage in our framework (Fig. 3) and carry out preliminary analyses that aim at assessing parameter identifiability and sensitivity to the measurable outputs. These analyses are performed on all candidate model structures.

Identifiability analyses assess the possibility of assigning unique values to the parameters $\boldsymbol{\theta}$ of a model structure $\mathcal{M}$

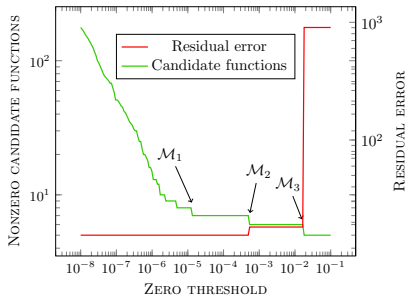

Fig. 2. Generating set of candidate models for the model selection step.

from the system output: a prerequisite for model calibration. Identifiability is distinguished in structural identifiability and practical identifiability.

A. Structural identifiability

Structural identifiability relates to the model structure $\mathcal{M}$ and experimental scheme (denoted by φ): the set of experimental conditions including time-varying stimuli, sampling times, experiment duration and initial conditions. The aim of structural identifiability is to assess the possibility of estimating parameters under ideal conditions (i.e. noise-free and time continuous data). While multiple methods have been developed for structural identifiability [23], [24], [25], the analysis is rarely performed due to the computational cost associated with the required symbolic manipulations, which scales with model non-linearity and complexity. To test local structural identifiability, we use the `STRIKE-GOLDD` toolbox [26], wherein the analysis is framed as a generalisation of observability. Specifically, the software augments the state vector x with the parameter vector $\boldsymbol{\theta}$ and recursively computes a generalised observability-identifiability matrix using the Lie derivatives of the output function $h(\boldsymbol{x}, \boldsymbol{\theta})$. Once the minimum number of Lie derivatives for which the matrix has full rank has been computed, the calculation of each additional derivative is followed by the reassessment of the rank. The computation stops either when the maximum number of Lie derivatives (i.e. the length of the augmented state vector minus one) has been reached or addition of a new derivative does not change the rank. At this stage, the full rank of the observability-identifiability matrix ensures local structural identifiability, while a lower rank denotes lack of identifiability for a specified list of parameters.

B. Practical identifiability

Practical identifiability quantifies the expected uncertainty of the parameter estimates of each model structure, subject to a hypothetical experimental scheme. Using Monte-Carlo sampling, numerous (ideally > 500) noise-corrupted time-series are simulated in response to a predefined experimental profile. After running parameter estimation on each *in silico* replicate, we obtain a set of parameter vectors which are assumed to be contained in an hyper-ellipsoid [27]. Principal Component Analysis (PCA) applied to the 90% interquartile range of the solutions hence details the correlation between

parameters (eccentricity of the hyper-ellipsoid) and uncertainty in their estimates (volume of the hyper-ellipsoid). As described in [27], additional quality measures can be computed from the hyper-ellipsoid for the sake of comparing the expected informative content of experimental schemes.

C. Sensitivity analysis

Through a global sensitivity analysis, the parameters of each model structure are then ranked according to their relative influence on the model predictions $\boldsymbol{y}$. The contribution of each parameter θ to the observables o in an experimental scheme φ is assessed through diverse importance factors derived from the relative sensitivities (see, e.g [28]),

$$s_{\theta}^{\varphi,o} = \frac{\Delta\theta}{\Delta \boldsymbol{y}^{\varphi,o}} \frac{\partial \boldsymbol{y}^{\varphi,o}}{\partial \theta}. \tag{8}$$

Considering that only initial estimates of parameters are available, the ranking is calculated for n_{lhs} parameter vectors $\boldsymbol{\theta} \in \Theta$, obtained through Latin Hypercube sampling of the parameter space Θ, as constrained by available experimental data. In this method, n_{lhs} samples for each parameter are obtained by dividing its range in n_{lhs} non-overlapping intervals and then randomly selecting a value from each interval. Among the importance factors computed to rank parameters, δ^{msqr} can be used to quantify the sensitivity of a specific model structure to each of its parameters. The definition of δ^{msqr} for a specific parameter θ for a set of n_{φ} experiments, n_o observables, and T sampling times is:

$$\delta_{\theta}^{msqr} = \frac{1}{N_D} \sqrt{\sum_{mc=1}^{n_{lhs}} \sum_{\varphi=1}^{n_{\varphi}} \sum_{o=1}^{n_o} \sum_{j=1}^{T} ([s_{\theta}^{\varphi,o}(t_j^{\varphi,o})]_{mc})^2}, \tag{9}$$

with $N_D = n_{lhs} n_{\varphi} n_o T$.

It is important to note that, due to its generality, the above formulation masks the effect of a specific experiment or observable on parametric sensitivities. However, an analogous analysis can be run to gather information on parameters that are more relevant in a given experimental scheme.

We run practical identifiability, sensitivity analysis, and optimal input design (described in the following) using the computational modules implemented in the `AMIGO2` toolbox [29].

V. Optimal experimental design for model discrimination

The availability of a series of model structures $\mathcal{M}_i(\boldsymbol{\theta}_i)$, with $i = 1, 2, ..., P$, demands the solution of model selection: the problem of discriminating the most plausible model among rival candidates. As the model selection is seldom performed due to the high computational and experimental costs it introduces, we address it using sequential Optimal Experimental Design (OED). Indeed, OED provides a suite of statistical methods to guide the definition of experimental schemes (time-varying inputs, measured outputs, sampling times and experiment duration) that balance the trade-off between the incurred costs and the information gained on the biological system.

Adopting a frequentist approach to model discrimination, we design optimal experiments by maximising the divergence of the predicted responses of pairs of models and their associated confidence intervals. To overcome the burden of designing $\binom{2}{P}$ experiments to select among P candidates, we focus on the design of experiments that enhance discrimination of the two most promising model structures according to the goodness of fit ranking on existing datasets. Following [30], we embed OED for model selection in a dynamic optimisation framework and search for the experimental scheme that maximises:

$$\begin{aligned} \boldsymbol{\varphi}_T = \arg\max_{\varphi \in \Phi} \mathop{\mathbb{E}}_{\boldsymbol{\theta}_1 \in \Theta_1, \boldsymbol{\theta}_2 \in \Theta_2} \int_{t_0}^{t_f} [\hat{\boldsymbol{y}}(\varphi, \boldsymbol{\theta}_1, t) - \\ \hat{\boldsymbol{y}}(\varphi, \boldsymbol{\theta}_2, t)]^T \mathcal{Q} [\hat{\boldsymbol{y}}(\varphi, \boldsymbol{\theta}_1, t) - \hat{\boldsymbol{y}}(\varphi, \boldsymbol{\theta}_2, t)] dt, \end{aligned} \tag{10}$$

where $\boldsymbol{\theta}_i$, with $i = 1, 2$ is the parameter vector of model $\mathcal{M}_i$, whose entries are randomly sampled from a normal distribution with mean and standard deviation set by the current parameter estimate value and confidence interval. $\hat{\boldsymbol{y}}(\varphi, \boldsymbol{\theta}_i, t)$ denotes the predicted responses of model $\mathcal{M}_i$, parameterized with $\boldsymbol{\theta}_i$, under the experimental scheme φ. Finally, $\mathcal{Q}$ is a diagonal weighting matrix accounting for magnitude differences in the simulated observables.

To save computational effort, we restrict the design to the decision of control values of the time-varying inputs (i.e. piece-wise constant functions of fixed duration) used to stimulate cells growing in a microfluidic device. This corresponds to addressing an optimal input design (OID) problem. We note that the assumption of a fixed sampling frequency minimises the impact that the continuous approximation of the output functions, which saves computational time, exerts on the loss of the notion of sampling times [31]. Numerically, dynamic optimisation relies on control vector parametrization (CVP), a direct approach that converts the infinite dimensional problem into a finite nonlinear programming problem and proves amenable to large scale ODE models. The time-series $(u_j^*, \boldsymbol{y}_j)$ gathered with the optimally designed experiment are used to augment the available dataset and update the parameter estimates of all model structures. This supports the re-ranking of the models according to their adequacy. The steps described above are repeated until one model $\mathcal{M}$ is considered a winner, as it outranks the best rival by a (subjectively chosen) quantitative threshold ε.

VI. Optimal experimental design for model calibration

The outcome of model selection is a structure $\mathcal{M}$ with an initial set of parameters $\boldsymbol{\theta}$ which we assume to be in an arbitrarily small neighbourhood of the optimal values $\boldsymbol{\theta}^*$. Once the structure $\mathcal{M}$ has been set, our objective is to refine parametric estimates so that $\boldsymbol{\theta} = \boldsymbol{\theta}^*$: this is the goal of Model Calibration.

We treat Model Calibration as a dynamic optimisation problem [27]; our objective is to maximise the amount of information obtained from a *in vivo* experiment. We estimate

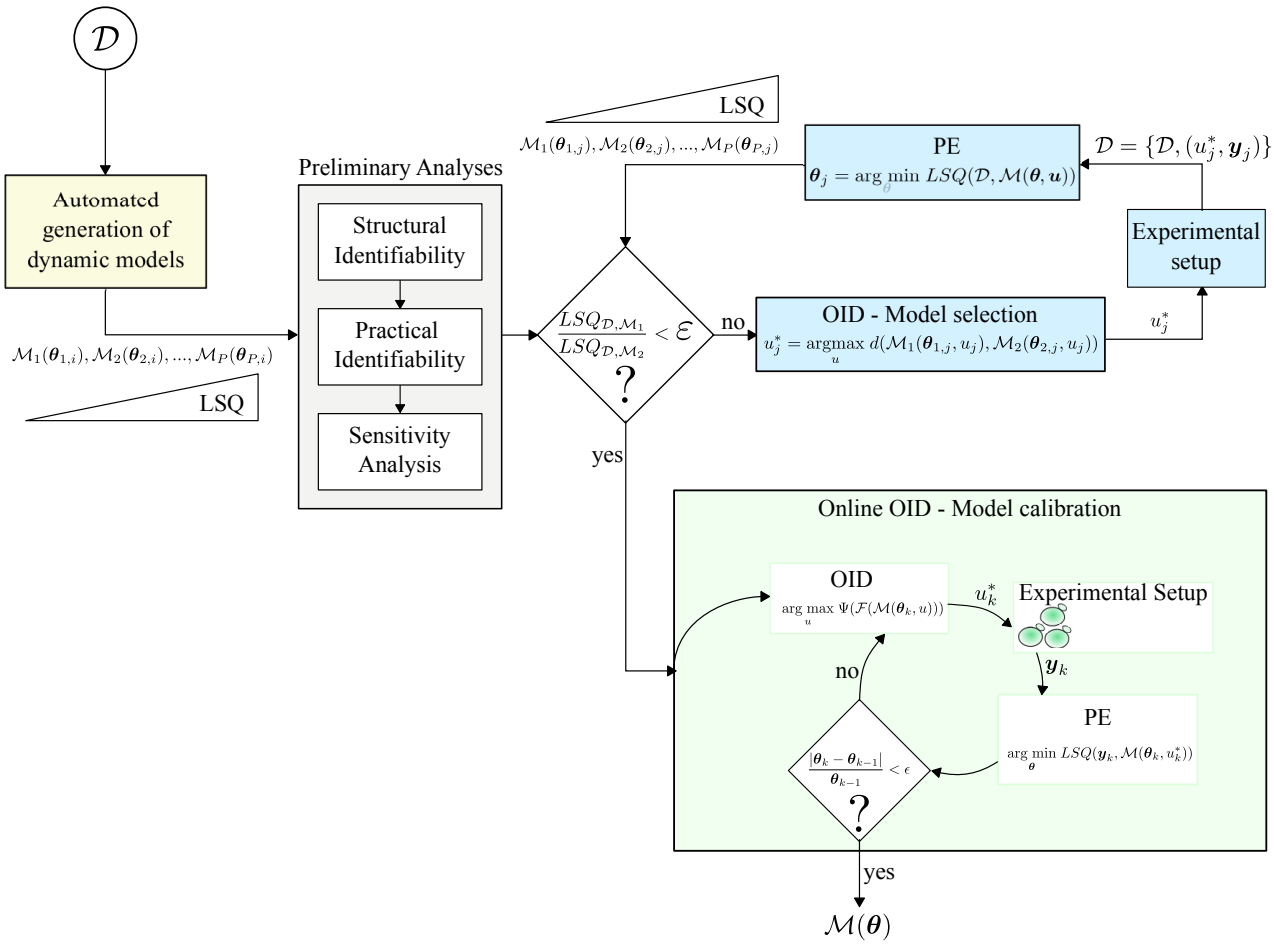

Fig. 3. Flow chart of the framework for systematic modelling of biological systems. Existing time series $\mathcal{D}$ is leveraged upon to postulate a set of model structures. These are ranked according to their plausibility in light of the data. Once preliminary analyses (identifiability and sensitivity) are carried out on each model candidate, the 2 top ranked model structures are better discriminated using Optimal Input Design (OID) for model selection. The experiment is run and the segment of data acquired at iteration j, $(u_j^*, \boldsymbol{y}_j)$, augments the pre-existing dataset. All model structures are re-calibrated on the updated data and re-assessed in light of them. Sequential OID for model selection is repeated until the best model, i.e. the one that outranks all rivals is identified. Hence parameter estimates are refined using on-line OID for model calibration. This loop terminates when the improvement in the accuracy of the parameter estimates meets a user-defined threshold ϵ.

the information yield of an experiment using the Fisher Information Matrix defined as:

$$\mathcal{F} = \sum_{o=1}^{n_o} \sum_{j=1}^{\mathrm{T}} \nabla_\theta y_o(t_j \mid \boldsymbol{\theta})\ \mathbf{Q}_o(t_j)\ \nabla_\theta y_o(t_j \mid \boldsymbol{\theta}), \qquad (11)$$

with: y_o generic observable of the system (e.g. a fluorescent reporter, in total: n_o), $\mathbf{Q}_o(t_j) \in \Omega \subset \mathbb{R}^{m_{oj} \times m_{oj}}$ nonnegative definite symmetric matrix (related to the experimental error), T number of time points. Many measures of information Ψ are defined over $\mathcal{F}$, two of the most popular ones are D- and E-optimality. D-optimality seeks to maximise the $\det(\mathcal{F})$; via the Cramer-Rao bound, we can easily see that this corresponds to minimise the lower bound estimate of the product of the variances associated to the parametric estimates. E-optimality, instead, seeks to maximise the minimum of the eigenvalues of $\mathcal{F}$; i.e. minimise the maximum uncertainty on parametric estimates.

To maximise the information yield, we can optimise either (a) the input time profile or (b) the observables we measure or (c) the sampling times—as well as any combination of the three. For the sake of simplicity here we focus on optimisation of the system's input only; this is also known as Optimal Input Design (OID) and can be formalised, at each iteration k, as the following optimisation problem:

$$u_k^* = \underset{u \in\, \mathcal{U}}{\arg\max}\, \Psi(\mathcal{F}(\mathcal{M}(\boldsymbol{\theta}_k, u))). \qquad (12)$$

At this point we leverage on the cyber-physical nature of our platform [32] to translate the input we have identified solving the optimisation problem u_k^* into a physicochemical stimulus to cells trapped in a microfluidic device and periodically observed via videomicroscopy. The response of the cells $\boldsymbol{y}_k$ is used to update parameter estimates; this is cast again as an optimisation problem whereby, in light on the new data, we seek to find the new parameter estimates $\boldsymbol{\theta}$ that minimise the sum of the squares of the residuals (LSQ) between the model prediction and the experimental data

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\, LSQ(\boldsymbol{y}_k, \mathcal{M}(\boldsymbol{\theta}_k, u_k^*)). \qquad (13)$$

We then iterate, potentially within the same experiment (online Optimal Experimental Design), over this loop: stimulating the cells, quantifying their response, updating our model and redesigning optimal inputs, in light of the new model, until we detect parametric convergence.

VII. CONCLUSIONS

This tutorial has presented a framework for the identification of biomolecular systems using a four-step process (automated model building, preliminary analyses, model selection, and model calibration). The framework has demonstrated a possible integration of different computational tools [21], [26], [29] and a microfluidics platform to provide the users with a data-driven, efficient solution for biomolecular system identification.

VIII. ACKNOWLEDGEMENT

This work was developed within the project "COSY-BIO" (Control Engineering of Biological Systems for Reliable Synthetic Biology): H2020-FETOPEN-2016-2017, project number 766840. This work was also supported by a Royal Academy of Engineering Chair in Emerging Technologies and a U.K. EPSRC Fellowship (EP/M002187/1) to Prof Stan, a Royal Society of Edinburgh-MoST grant, EPSRC grant EP/R035350/1 and EP/S001921/1 to Dr Menolascina, and the EPSRC grant EP/P017134/1-CONDSYC to Dr Bandiera.

REFERENCES

[1] D. Del Vecchio, Y. Qian, R. M. Murray, and E. D. Sontag, "Future systems and control research in synthetic biology," *Annual Reviews in Control*, 2018.

[2] D. D. Vecchio and R. M. Murray, *Biomolecular Feedback Systems*. Princton University Press, 2014.

[3] M. Khammash, M. Di Bernardo, and D. Di Bernardo, "Cybergenetics: Theory and methods for genetic control systems," in *58th IEEE Conference on Decision and Control*, 2019.

[4] F. Blanchini, H. EI-Samad, G. Giordano, and E. D. Sontag, "Control-theoretic methods for biological networks," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, dec 2018.

[5] P. Harrigan, H. Madhani, and H. El-Samad, "Real-time genetic compensation defines the dynamic demands of feedback control." *Cell*, vol. 175, no. 3, pp. 877–886.e10, 2018.

[6] C. Briat, A. Gupta, and M. Khammash, "Antithetic integral feedback ensures robust perfect adaptation in noisy biomolecular networks," *Cell systems*, vol. 2, no. 1, pp. 15–26, 2016.

[7] J. Ruess, F. Parise, A. Milias-Argeitis, M. Khammash, and J. Lygeros, "Iterative experiment design guides the characterization of a light-inducible gene expression circuit," *Proceedings of the National Academy of Sciences*, vol. 112, no. 26, pp. 8148–8153, 2015.

[8] M. Gevers, "Identification for control: From the early achievements to the revival of experiment design," *European journal of control*, vol. 11, no. 4-5, pp. 335–352, 2005.

[9] J. Schoukens and L. Ljung, "Nonlinear system identification: A user-oriented roadmap," *arXiv*, p. 1902.00683v1, 2019.

[10] W. M. Haddad, V. Chellaboina, and Q. Hui, *Nonnegative and Compartmental Dynamical Systems*. Princeton University Press, 2010.

[11] J. Beal, C. Overney, A. Adler, F. Yaman, L. Tiberio, and M. Samineni, "Tasbe flow analytics: A package for calibrated flow cytometry analysis," *ACS Synthetic Biology*, vol. 0, no. 0, p. null, 0, pMID: 31053031. [Online]. Available: https://doi.org/10.1021/acssynbio.8b00533

[12] D. Benzinger and M. Khammash, "Pulsatile inputs achieve tunable attenuation of gene expression variability and graded multi-gene regulation," *Nature communications*, vol. 9, no. 1, p. 3521, 2018.

[13] M. M. Zavlanos, A. A. Julius, S. P. Boyd, and G. J. Pappas, "Inferring stable genetic networks from steady-state data," *Automatica*, vol. 47, no. 6, pp. 1113–1122, 2011.

[14] B. N. Kholodenko, A. Kiyatkin, F. J. Bruggeman, E. Sontag, H. V. Westerhoff, and J. B. Hoek, "Untangling the wires: A strategy to trace functional interactions in signaling and gene networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12 841–12 846, 2002.

[15] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, mar 2016.

[16] I. Otero-Muras and J. Banga, "Automated design framework for synthetic biology exploiting pareto optimality." *ACS Synth Biol*, vol. 6, no. 7, pp. 1180–1193, 2017.

[17] A. Chiuso and G. Pillonetto, "A bayesian approach to sparse dynamic network identification," *Automatica*, vol. 48, no. 8, pp. 1553–1565, aug 2012.

[18] É. Walter, *Numerical Methods and Optimization*. Springer International Publishing, 2014.

[19] R. Chartrand, "Numerical differentiation of noisy, nonsmooth data," *ISRN Applied Mathematics*, vol. 2011, pp. 1–11, 2011.

[20] H. A. Harrington, K. L. Ho, and N. Meshkat, "Differential algebra for model comparison," *arXiv*, 2006.

[21] Z. A. Tuza and G.-B. Stan, "An automatic sparse model estimation method guided by constraints that encode system properties," in *European Control Conference (ECC), Naples Italy*, 2019.

[22] Z. A. Tuza, "Ode composer toolbox," 2019. [Online]. Available: https://github.com/GBS-SynBioLab/ode_composer

[23] J. Karlsson, M. Anguelova, and M. Jirstrand, "An efficient method for structural identifiability analysis of large dynamic systems," *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 941–946, 2012.

[24] G. Bellu, M. P. Saccomani, S. Audoly, and L. D'Angiò, "Daisy: A new software tool to test global identifiability of biological and physiological systems," *Computer methods and programs in biomedicine*, vol. 88, no. 1, pp. 52–61, 2007.

[25] O. Chiş, J. R. Banga, and E. Balsa-Canto, "Genssi: a software toolbox for structural identifiability analysis of biological models," *Bioinformatics*, vol. 27, no. 18, pp. 2610–2611, 2011.

[26] A. F. Villaverde, A. Barreiro, and A. Papachristodoulou, "Structural identifiability of dynamic systems biology models," *PLoS computational biology*, vol. 12, no. 10, p. e1005153, 2016.

[27] E. Balsa-Canto, A. A. Alonso, and J. R. Banga, "Computational procedures for optimal experimental design in biological systems," *IET systems biology*, vol. 2, no. 4, pp. 163–172, 2008.

[28] ——, "An iterative identification procedure for dynamic modeling of biochemical networks," *BMC systems biology*, vol. 4, no. 1, p. 11, 2010.

[29] E. Balsa-Canto, D. Henriques, A. Gábor, and J. R. Banga, "Amigo2, a toolbox for dynamic modeling, optimization and control in systems biology," *Bioinformatics*, vol. 32, no. 21, pp. 3357–3359, 2016.

[30] D. M. Espie, "The use of nonlinear parameter estimation for dynamic chemical reactor modelling." Ph.D. dissertation, Imperial College London (University of London), 1986.

[31] D. Espie and S. Macchietto, "The optimal design of dynamic experiments," *AIChE Journal*, vol. 35, no. 2, pp. 223–229, 1989.

[32] L. Bandiera, Z. Hou, V. Kothamachu, E. Balsa-Canto, P. Swain, and F. Menolascina, "On-line optimal input design increases the efficiency and accuracy of the modelling of an inducible synthetic promoter," *Processes*, vol. 6, no. 9, p. 148, 2018.