

Caltech

Transforming Data Across Environments Despite Structural Non-Identifiability

ACC 2019

Vipul Singhal ^{1,2} Richard M. Murray ¹

July 12, 2019

¹California Institute of Technology

²Genome Institute of Singapore

- Main point: It might not always be necessary for parameters to be (structurally) identifiable for models to do what you want.

- Main point: It might not always be necessary for parameters to be (structurally) identifiable for models to do what you want.
- Set of output-indistinguishable parameters = equivalence class w.r.t parameter identification.

- Main point: It might not always be necessary for parameters to be (structurally) identifiable for models to do what you want.
- Set of output-indistinguishable parameters = equivalence class w.r.t parameter identification.
- Equivalent w.r.t. given modeling objective? When?

Overview

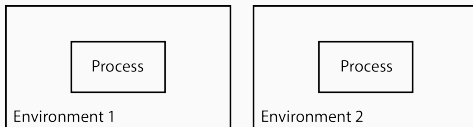
- Main point: It might not always be necessary for parameters to be (structurally) identifiable for models to do what you want.
- Set of output-indistinguishable parameters = equivalence class w.r.t parameter identification.
- Equivalent w.r.t. given modeling objective? When?
- If yes → pick arbitrary point from these sets.

- Main point: It might not always be necessary for parameters to be (structurally) identifiable for models to do what you want.
- Set of output-indistinguishable parameters = equivalence class w.r.t parameter identification.
- Equivalent w.r.t. given modeling objective? When?
- If yes → pick arbitrary point from these sets.
- Examples of problems:

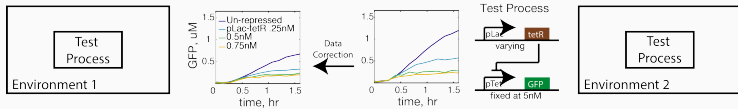
- Main point: It might not always be necessary for parameters to be (structurally) identifiable for models to do what you want.
- Set of output-indistinguishable parameters = equivalence class w.r.t parameter identification.
- Equivalent w.r.t. given modeling objective? When?
- If yes → pick arbitrary point from these sets.
- Examples of problems:
 - batch variability reduction or data transformation

- Main point: It might not always be necessary for parameters to be (structurally) identifiable for models to do what you want.
- Set of output-indistinguishable parameters = equivalence class w.r.t parameter identification.
- Equivalent w.r.t. given modeling objective? When?
- If yes → pick arbitrary point from these sets.
- Examples of problems:
 - batch variability reduction or data transformation
 - characterize subsystem behavior → predict system behavior

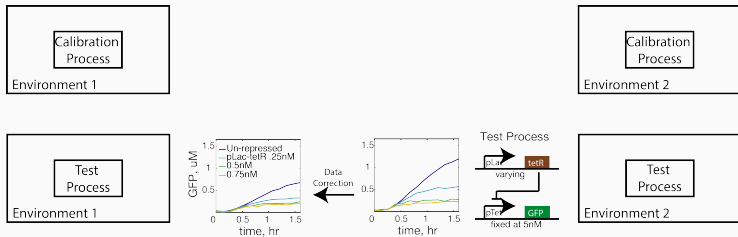
Transforming Process Behavior Across Environments



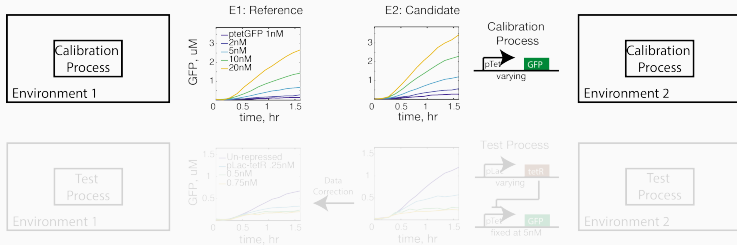
The Data Correction Problem



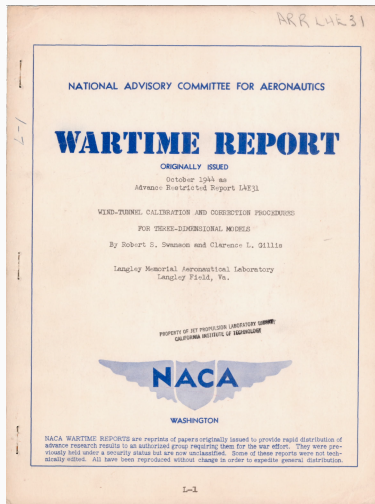
The Data Correction Problem



The Data Correction Problem



Solution: The Calibration-Correction Method

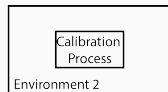
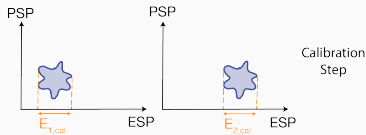
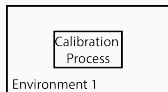


Solution: The Calibration-Correction Method

The usual practice of predicting the flying qualities of airplanes from wind-tunnel tests of relatively small-scale models makes it imperative that the model test results be corrected to free-air conditions. In addition, the large number of wind tunnels in use makes it desirable that a more or less standard calibration and correction procedure be adopted in order to make data from different tunnels as nearly comparable as possible. Not much com-

The Calibration-Correction Method: Calibration Step

Estimate environment specific parameters



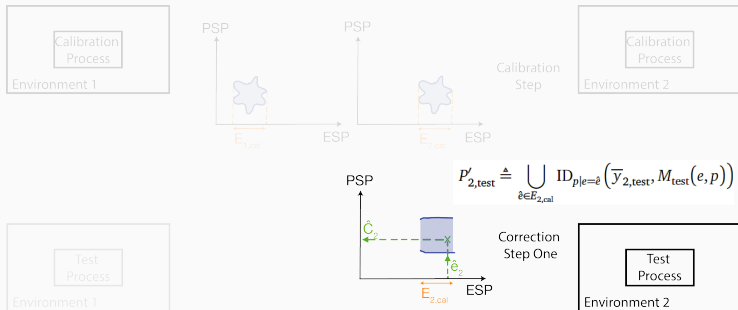
$$\hat{\Theta}_{cal} \triangleq \{ (e_1, e_2, p) \mid \bar{y}_{i,cal} = M_{cal}(e_i, p), i = 1, 2 \}$$

$$E_{i,cal} \triangleq \text{proj}_{e_i} \hat{\Theta}_{cal}, \quad i = 1, 2$$



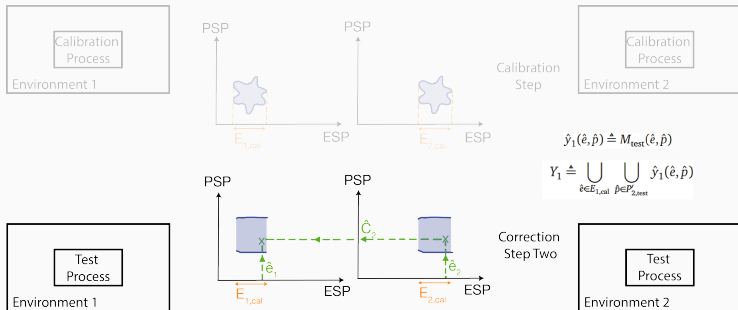
The Calibration-Correction Method: Correction Step 1

Estimate test process parameters in Extract 2



The Calibration-Correction Method: Correction Step 2

Predict test process behavior in Extract 1



Parameter Consistency

Theorem

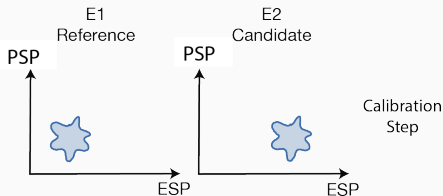
Consider the data correction problem in the model universe. Furthermore, consider the calibration-correction method, and the sets $\tilde{\Theta}_{\text{cal}}$, $E_{1,\text{cal}}$, $E_{2,\text{cal}}$ and $C'_{2,\text{test}}$. Define $\Theta_{i,\text{test}} \triangleq \text{ID}_{\theta}(\bar{y}_{i,\text{test}}, \bar{M}_{\text{test}}(\theta))$ for $i = 1, 2$. Then, the conditions,

$$\tilde{\Theta}_{\text{cal}} \neq \emptyset$$

$$E_{2,\text{cal}} \subseteq \text{proj}_e \Theta_{2,\text{test}}$$

$$E_{1,\text{cal}} \times C'_{2,\text{test}} \subseteq \Theta_{1,\text{test}},$$

are necessary and sufficient for the calibration-correction method to solve the data correction problem.



Parameter Consistency

Theorem

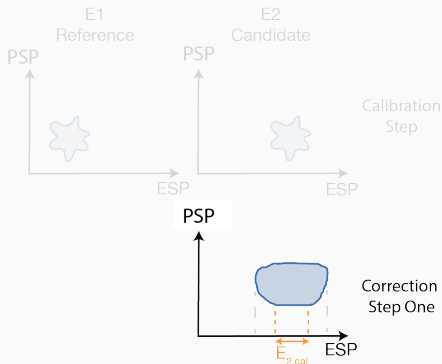
Consider the data correction problem in the model universe. Furthermore, consider the calibration-correction method, and the sets $\tilde{\Theta}_{\text{cal}}$, $E_{1,\text{cal}}$, $E_{2,\text{cal}}$ and $C'_{2,\text{test}}$. Define $\Theta_{i,\text{test}} \triangleq \text{ID}_{\theta}(\bar{y}_{i,\text{test}}, \bar{M}_{\text{test}}(\theta))$ for $i = 1, 2$. Then, the conditions,

$$\tilde{\Theta}_{\text{cal}} \neq \emptyset$$

$$E_{2,\text{cal}} \subseteq \text{proj}_e \Theta_{2,\text{test}}$$

$$E_{1,\text{cal}} \times C'_{2,\text{test}} \subseteq \Theta_{1,\text{test}},$$

are necessary and sufficient for the calibration-correction method to solve the data correction problem.



Parameter Consistency

Theorem

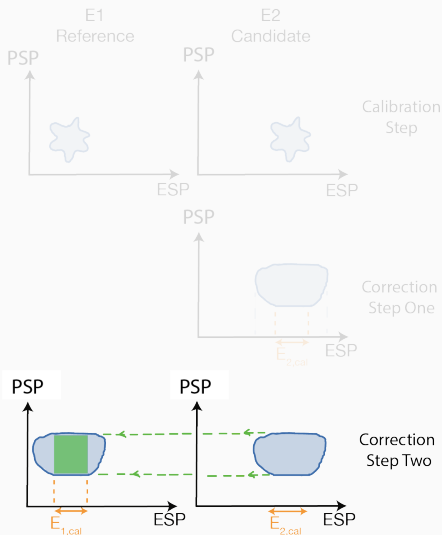
Consider the data correction problem in the model universe. Furthermore, consider the calibration-correction method, and the sets $\tilde{\Theta}_{\text{cal}}$, $E_{1,\text{cal}}$, $E_{2,\text{cal}}$ and $C'_{2,\text{test}}$. Define $\Theta_{i,\text{test}} \triangleq \text{ID}_{\theta}(\bar{y}_{i,\text{test}}, \bar{M}_{\text{test}}(\theta))$ for $i = 1, 2$. Then, the conditions,

$$\tilde{\Theta}_{\text{cal}} \neq \emptyset$$

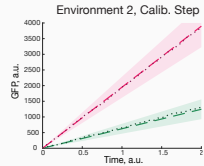
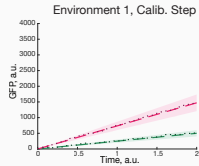
$$E_{2,\text{cal}} \subseteq \text{proj}_e \Theta_{2,\text{test}}$$

$$E_{1,\text{cal}} \times C'_{2,\text{test}} \subseteq \Theta_{1,\text{test}},$$

are necessary and sufficient for the calibration-correction method to solve the data correction problem.



Consider the Test = Calibration Process Case

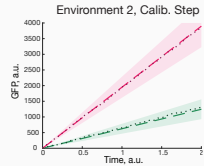
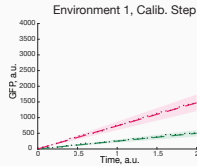


"Test = Calibration" Case



- sim. DNA = 10 a.u.
- sim. DNA = 30 a.u.
- ⋯ exp. DNA = 10 a.u.
- ⋯ exp. DNA = 30 a.u.

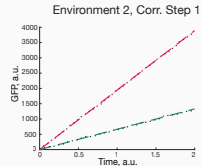
Consider the Test = Calibration Process Case



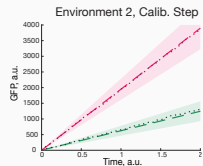
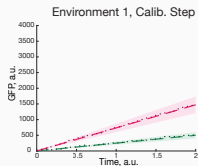
"Test = Calibration" Case



- sim. DNA = 10 a.u.
- sim. DNA = 30 a.u.
- ⋯ exp. DNA = 10 a.u.
- ⋯ exp. DNA = 30 a.u.



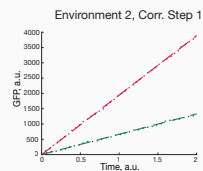
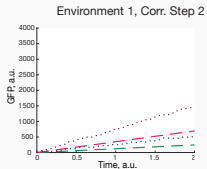
Correcting the Calibration Data Fails



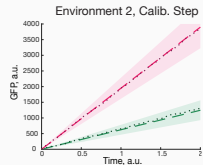
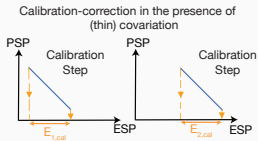
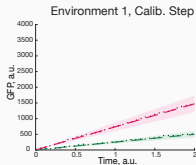
"Test = Calibration" Case



- sim. DNA = 10 a.u.
- sim. DNA = 30 a.u.
- ⋯ exp. DNA = 10 a.u.
- ⋯ exp. DNA = 30 a.u.



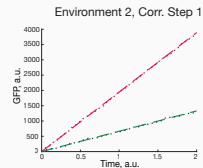
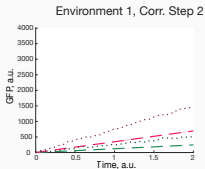
Parameter Covariation Could be a Problem



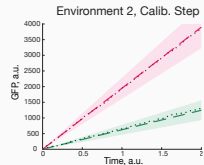
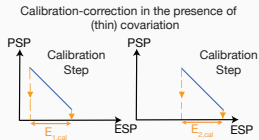
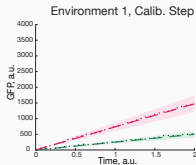
"Test = Calibration" Case

$\frac{p_{Tet}}{\text{varying (10 - 30 a.u.)}} \rightarrow \text{GFP}$

- sim. DNA = 10 a.u.
- sim. DNA = 30 a.u.
- ⋯ exp. DNA = 10 a.u.
- ⋯ exp. DNA = 30 a.u.



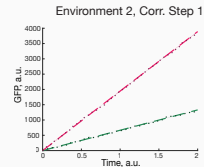
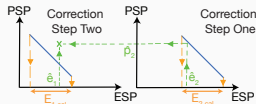
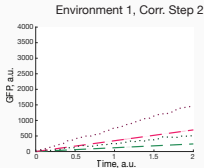
Parameter Covariation Could be a Problem



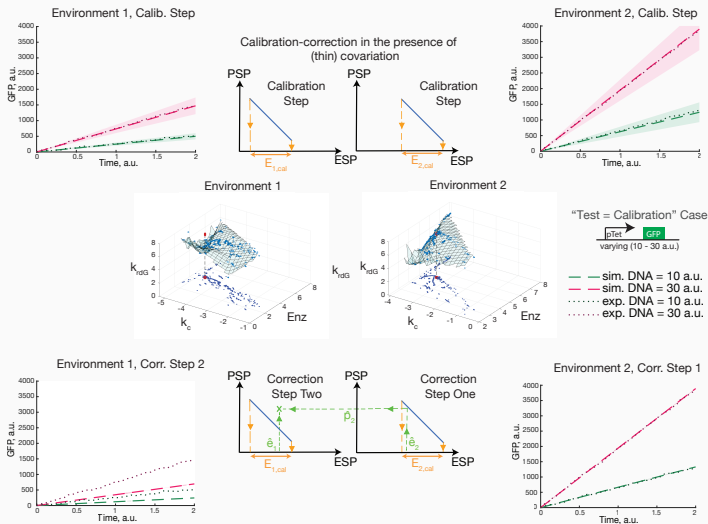
"Test = Calibration" Case



- sim. DNA = 10 a.u.
- sim. DNA = 30 a.u.
- ⋯ exp. DNA = 10 a.u.
- ⋯ exp. DNA = 30 a.u.

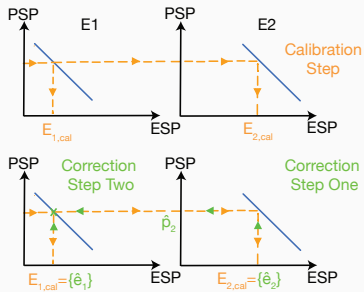


Parameter Covariation Could be a Problem



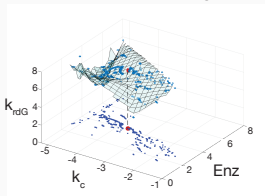
PSP Fixing Helps

Calibration Correction with PSP fixing

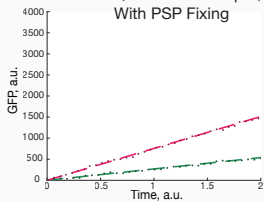


PSP Fixing Helps

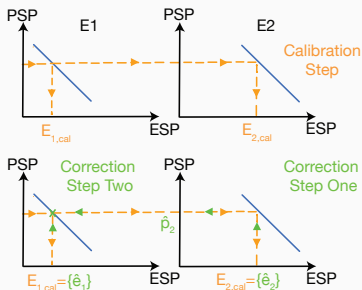
Env. 1, With PSP Fixing



Env 1, Correction Step 2,
With PSP Fixing



Calibration Correction with PSP fixing



Proposition

Consider the 'Test = Calib' case of the DCP. Assume the conditions

$$\begin{aligned}\check{\Theta}_{\text{cal}} &\neq \emptyset, \\ C'_{2,\text{cal}} &\subseteq \text{proj}_c \Theta_{1,\text{cal}}, \\ E_{i,\text{cal}} &\subseteq \text{proj}_e \Theta_{i,\text{cal}}, \quad i = 1, 2,\end{aligned}$$

hold, but the set

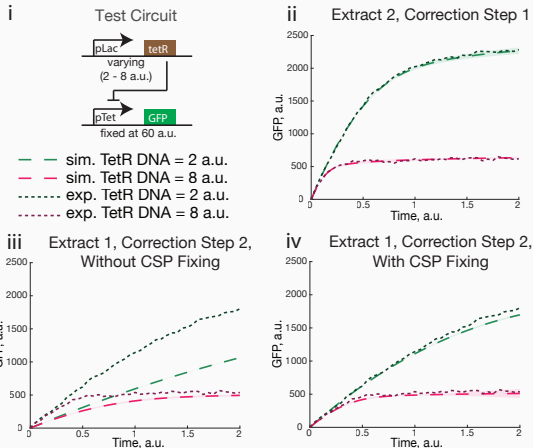
$$\Theta'_{1,\text{cal}} \triangleq \Theta_{1,\text{cal}} \cap (E_{1,\text{cal}} \times \text{proj}_c \Theta_{2,\text{cal}})$$

has covariation with respect to the (e, c) partition. Then, the calibration-correction method fails to solve this problem.

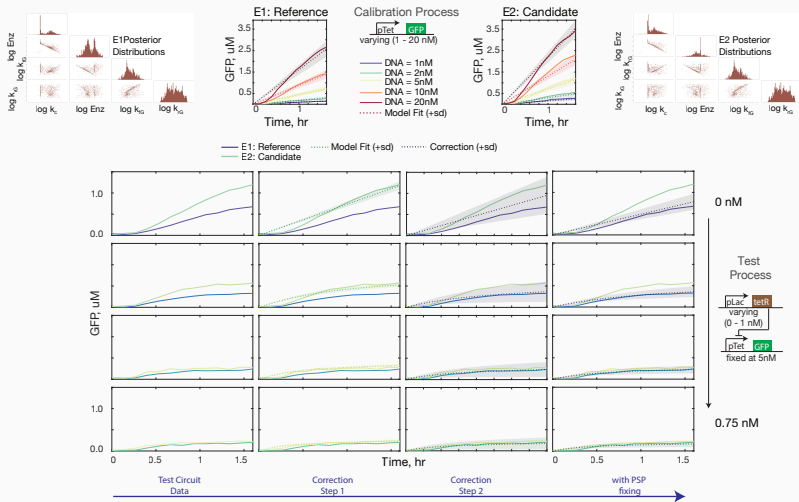
Proposition

Consider the sets $\Theta_{i,\text{cal}} \triangleq \text{ID}_\theta(\bar{y}_{i,\text{cal}}, M_{\text{cal}}(\theta))$ for $i = 1, 2$, and the partition $\theta = (e, c)$. Assume that the $\Theta_{i,\text{cal}}$ have thin covariation in their c coordinates with respect to their e coordinates. Then, the calibration-correction method with CSP fixing is able to solve the DCP for the 'Test = Calib' case.

PSP Fixing Helps for New Test Data Too



PSP Fixing Helps with Experimental Data



Acknowledgements

Thanks to Richard Murray, Sam Clamons, Anandh Swaminathan, William Poole, Wolfgang Halter, Andras Gyorgy, and the Murray lab.

Research supported by:



The Data Correction Problem

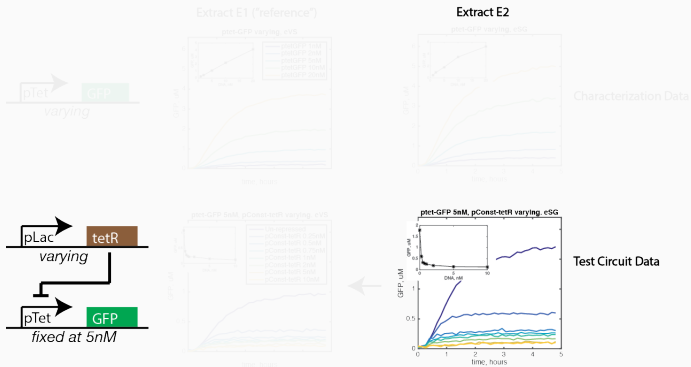


Figure 1: Given: Experimental data ($\bar{y}_{2,t}$) for a circuit in an extract, E2. The subscript t stands for test circuit.

The Data Correction Problem

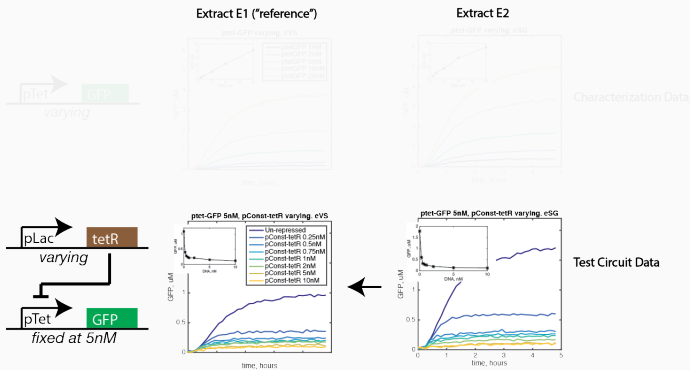


Figure 2: Goal: Transform it into a the behavior $(\bar{y}_{1,t})$ of that experiment in a different extract, E1, which we call the 'reference' extract.

The Data Correction Problem

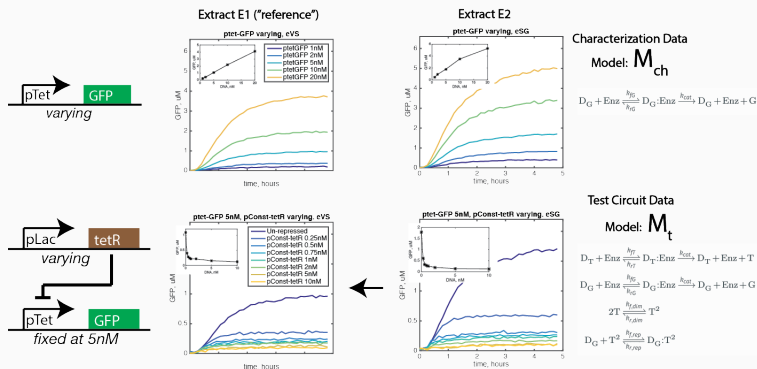


Figure 3: Available tools: Choice of a set of characterization experiments ($\bar{y}_{1,ch}, \bar{y}_{2,ch}$) and corresponding models for both the characterization (M_{ch}) and test (M_t) circuits.

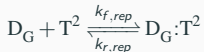
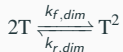
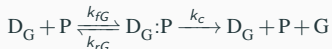
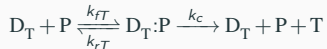
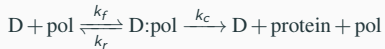
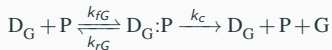
The Data Correction Problem (DCP)

- Have
 - Data for test circuit in E2, the candidate extract, $\bar{y}_{2,t}$
 - Ability to choose characterization experiments, and collect data $\bar{y}_{1,ch}, \bar{y}_{2,ch}$
 - Ability to choose models, M_{ch} and M_t
- Find a method that gives $\hat{y}_{1,t}$, such that $\hat{y}_{1,t} = \bar{y}_{1,t}$

Some Notation

- *Extracts vs Circuits*
- Model parameter vector (θ) coordinates partitioned into Extract Specific Parameters (e) and Circuit Specific Parameters (c), so that $\theta = (e, c)$

ESPs vs CSPs



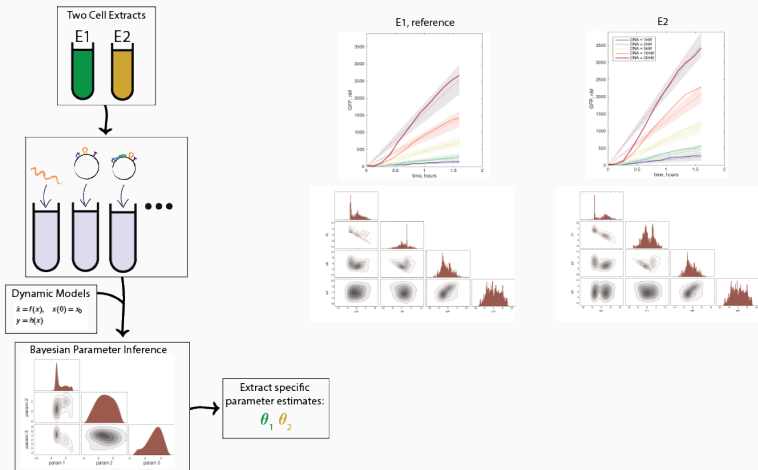
- Environment specific parameters, e : $[P]_{\text{total}}$, k_c
- Process specific parameters, c : everything else

Proposed Solution

The Calibration-Correction Method - Calibration Step



The Calibration-Correction Method - Calibration Step



The Calibration-Correction Method - Calibration Step

- **Calibration Step:** Characterization circuit to estimate extract specific parameters. For $i = \{1, 2\}$,

The Calibration-Correction Method - Calibration Step

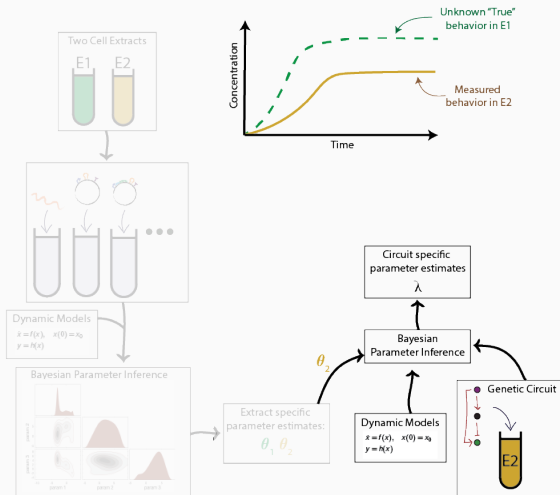
- **Calibration Step:** Characterization circuit to estimate extract specific parameters. For $i = \{1, 2\}$,
 - find $\hat{e}_{i,ch}$ such that $\bar{y}_{i,ch} = M_{ch}(\hat{e}_{i,ch}, c_{i,ch})$ for some $c_{i,ch}$.

The Calibration-Correction Method - Calibration Step

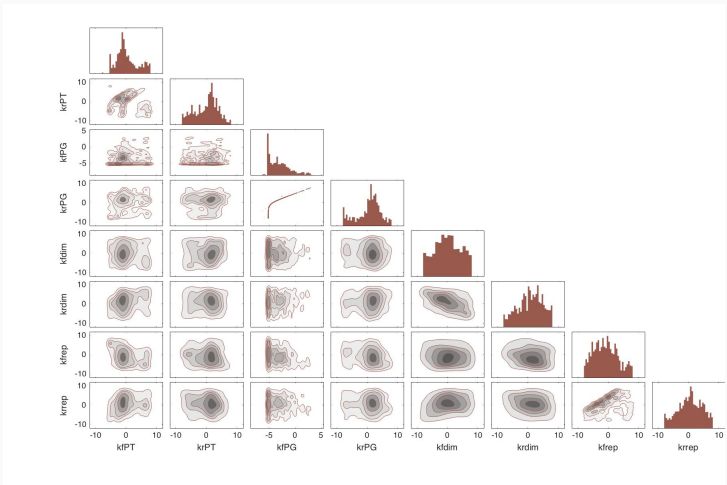
- **Calibration Step:** Characterization circuit to estimate extract specific parameters. For $i = \{1, 2\}$,
 - find $\hat{e}_{i,ch}$ such that $\bar{y}_{i,ch} = M_{ch}(\hat{e}_{i,ch}, c_{i,ch})$ for some $c_{i,ch}$.
 - All such points:

$$E_{i,ch} \triangleq \text{proj}_e \text{ID}_\theta(\bar{y}_{i,ch}, M_{ch}) \quad \forall i \in \{1, 2\}. \quad (1)$$

The Calibration-Correction Method - Correction Step1



The Calibration-Correction Method - Correction Step 1



The Calibration-Correction Method - Correction Step 1

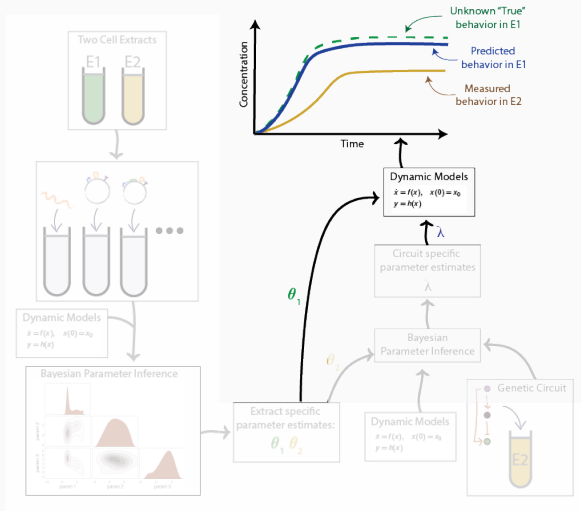
- **Correction Step One:** Test data, Extract 2. $(\bar{y}_{2,t})$ + extract parameters fixed at $\hat{e}_{2,ch} \rightarrow$ identify test parameters $\hat{c}_{2,t}$

The Calibration-Correction Method - Correction Step 1

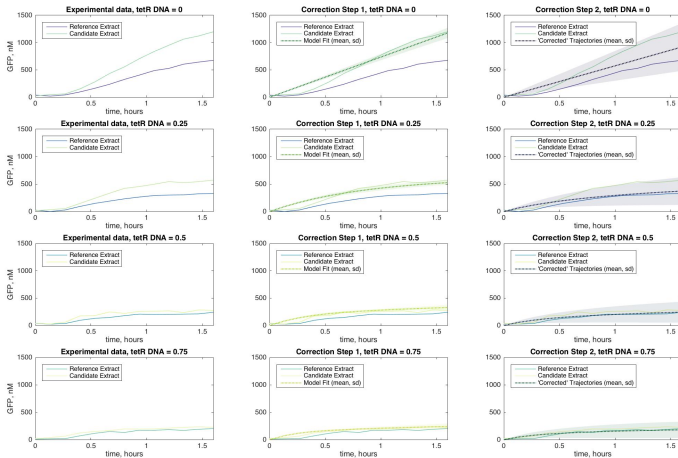
- **Correction Step One:** Test data, Extract 2. $(\bar{y}_{2,t})$ + extract parameters fixed at $\hat{e}_{2,ch} \rightarrow$ identify test parameters $\hat{c}_{2,t}$
 - All such points:

$$C'_{c,t} \triangleq \bigcup_{\hat{e} \in \bar{E}_{2,ch}} \text{ID}_{c|e=\hat{e}} \left(\bar{y}_{2,t}, M_t(e, c) \right)$$

The Calibration-Correction Method - Correction Step 2



The Calibration-Correction Method - Correction Step 2



- **Correction Step Two:** Predict $\hat{y}_{1,t} = M_t(\hat{e}_{1,ch}, \hat{c}_{2,t})$

The Calibration-Correction Method

- **Correction Step Two:** Predict $\hat{y}_{1,t} = M_t(\hat{e}_{1,ch}, \hat{c}_{2,t})$
 - All such trajectories:

$$Y_1 \triangleq \bigcup_{\hat{e} \in E_{1,ch}} \bigcup_{\hat{c} \in C'_{2,t}} \tilde{y}_1(\hat{e}, \hat{c}), \quad (2)$$

where,

$$\tilde{y}_1(\hat{e}, \hat{c}) = M_t(\hat{e}, \hat{c}). \quad (3)$$

Failure Conditions for this Method

- A parameter identification step is attempted when no parameter exists such that the model fits the data
- Correction step two is able to produce a trajectory not equal to the true trajectory

Results

Characterization Experiments Must be at Least as Informative as the Data to be Corrected

Characterization Experiments Must be at Least as Informative as the Data to be Corrected

Define,

$$E_{1,t} \triangleq \text{proj}_e \text{ID}_\theta(\bar{y}_{1,t}, \bar{M}_t),$$

$$E_{2,t} \triangleq \text{proj}_e \text{ID}_\theta(\bar{y}_{2,t}, \bar{M}_t).$$

Characterization Experiments Must be at Least as Informative as the Data to be Corrected

Define,

$$E_{1,t} \triangleq \text{proj}_e \text{ID}_\theta(\bar{y}_{1,t}, \bar{M}_t),$$

$$E_{2,t} \triangleq \text{proj}_e \text{ID}_\theta(\bar{y}_{2,t}, \bar{M}_t).$$

The first two conditions are,

$$E_{1,ch} \subseteq E_{1,t}$$

$$E_{2,ch} \subseteq E_{2,t}$$

Characterization Experiments Must be at Least as Informative as the Data to be Corrected

Define,

$$E_{1,t} \triangleq \text{proj}_e \text{ID}_\theta(\bar{y}_{1,t}, \bar{M}_t),$$

$$E_{2,t} \triangleq \text{proj}_e \text{ID}_\theta(\bar{y}_{2,t}, \bar{M}_t).$$

The first two conditions are,

$$E_{1,ch} \subseteq E_{1,t}$$

$$E_{2,ch} \subseteq E_{2,t}$$

Proof Ideas: If $E_{2,ch} \subseteq E_{2,t}$ does not hold, parameter estimation of the $C_{2,t}$ fails in first correction step. If $E_{1,ch} \subseteq E_{1,t}$ does not hold, then the prediction of a correct trajectory $\hat{y}_{1,t}$ is impossible for any c .

Process Specific Parameter Estimates for the Test Process Must Agree in Both Environments

Process Specific Parameter Estimates for the Test Process Must Agree in Both Environments

Recall that,

$$C'_{2,t} \triangleq \bigcup_{\hat{e} \in E_{2,ch}} \text{ID}_{c|e=\hat{e}} \left(\bar{y}_{2,t}, M_t(e, c) \right),$$
$$C'_{1,t} \triangleq \bigcup_{\hat{e} \in E_{1,ch}} \text{ID}_{c|e=\hat{e}} \left(\bar{y}_{1,t}, M_t(e, c) \right)$$

Process Specific Parameter Estimates for the Test Process Must Agree in Both Environments

Recall that,

$$C'_{2,t} \triangleq \bigcup_{\hat{e} \in E_{2,ch}} \text{ID}_{c|e=\hat{e}} \left(\bar{y}_{2,t}, M_t(e, c) \right),$$
$$C'_{1,t} \triangleq \bigcup_{\hat{e} \in E_{1,ch}} \text{ID}_{c|e=\hat{e}} \left(\bar{y}_{1,t}, M_t(e, c) \right)$$

The third condition is,

$$C'_{2,t} \subseteq C'_{1,t}.$$

Process Specific Parameter Estimates for the Test Process Must Agree in Both Environments

Recall that,

$$C'_{2,t} \triangleq \bigcup_{\hat{e} \in E_{2,ch}} \text{ID}_{c|e=\hat{e}} \left(\bar{y}_{2,t}, M_t(e, c) \right),$$
$$C'_{1,t} \triangleq \bigcup_{\hat{e} \in E_{1,ch}} \text{ID}_{c|e=\hat{e}} \left(\bar{y}_{1,t}, M_t(e, c) \right)$$

The third condition is,

$$C'_{2,t} \subseteq C'_{1,t}.$$

Proof Idea: In the second correction step, the prediction $\hat{y}_{1,t}$ is generated by plugging in an arbitrary point from $C'_{2,t}$ into what would have been $C'_{1,t}$.

The Environment Specific and Process Specific Parameters Should be Independent of Each Other.

The Environment Specific and Process Specific Parameters Should be Independent of Each Other.

Define,

$$\Theta_{1,t} \triangleq \text{ID}_{\theta}(\bar{y}_{1,t}, \bar{M}_t)$$

The Environment Specific and Process Specific Parameters Should be Independent of Each Other.

Define,

$$\Theta_{1,t} \triangleq \text{ID}_{\theta}(\bar{y}_{1,t}, \bar{M}_t)$$

The fourth condition is,

$$E_{1,ch} \times C'_{2,t} \subseteq \Theta_{1,t}$$

The Environment Specific and Process Specific Parameters Should be Independent of Each Other.

Define,

$$\Theta_{1,t} \triangleq \text{ID}_{\theta}(\bar{y}_{1,t}, \bar{M}_t)$$

The fourth condition is,

$$E_{1,ch} \times C'_{2,t} \subseteq \Theta_{1,t}$$

Proof idea: In the second correction step, we pick arbitrary points from the sets $E_{1,ch}$ and $C'_{2,t}$ as a proxy for a point in $\Theta_{1,t}$.

Supplement

Definition (Parameter Identification)

Let the set \mathcal{M}_Y be the set of all pairs $(\bar{y}, M(\theta, u, x_0))$ for which there exists a parameter $\hat{\theta} \in \Omega$ such that $\bar{y} = M(\hat{\theta}, u, x_0)$. Also, let $\mathcal{P}(\Omega)$ be the power set of Ω . We define the *parameter identification of θ* as an operator $ID_\theta : \mathcal{M}_Y \rightarrow \mathcal{P}(\Omega)$, with $ID_\theta(\bar{y}, M) = \{\hat{\theta} \in \Omega \mid \bar{y} = M(\hat{\theta}, u, x_0)\}$

Definition

$M(\theta_A)$ and $M(\theta_B)$ *output-indistinguishable* if,

$$\begin{aligned} &\theta_A, \theta_B \in \Omega, \\ &y(t, \theta_A, u, x_0) = y(t, \theta_B, u, x_0) \quad \forall t \geq 0, \forall u \in \mathcal{U}, \forall x_0 \in \mathcal{X}. \end{aligned} \tag{4}$$

Definition

$M(\theta_A)$ and $M(\theta_B)$ *output-indistinguishable* if,

$$\begin{aligned} &\theta_A, \theta_B \in \Omega, \\ &y(t, \theta_A, u, x_0) = y(t, \theta_B, u, x_0) \quad \forall t \geq 0, \forall u \in \mathcal{U}, \forall x_0 \in \mathcal{X}. \end{aligned} \tag{4}$$

Definition

The i^{th} coordinate of θ_A , denoted $\theta_{A,i}$, is *structurally globally identifiable* (SGI) if, for almost any $\theta_A \in \Omega$, (4) has a unique solution for $\theta_{B,i}$.

Definition

$M(\theta_A)$ and $M(\theta_B)$ *output-indistinguishable* if,

$$\begin{aligned} &\theta_A, \theta_B \in \Omega, \\ &y(t, \theta_A, u, x_0) = y(t, \theta_B, u, x_0) \quad \forall t \geq 0, \forall u \in \mathcal{U}, \forall x_0 \in \mathcal{X}. \end{aligned} \tag{4}$$

Definition

The i^{th} coordinate of θ_A , denoted $\theta_{A,i}$, is *structurally globally identifiable* (SGI) if, for almost any $\theta_A \in \Omega$, (4) has a unique solution for $\theta_{B,i}$.

Definition

$M(\theta)$ SGI if all its parameters θ_i are SGI.

Proof: Necessity of $E_{2,ch} \subseteq E_{2,t}$

- AFSOC $\exists \tilde{e} \in E_{2,ch}$ s. t. $\tilde{e} \notin E_{2,t}$.

Proof: Necessity of $E_{2,ch} \subseteq E_{2,t}$

- AFSOC $\exists \tilde{e} \in E_{2,ch}$ s. t. $\tilde{e} \notin E_{2,t}$.
- This implies $\nexists \tilde{c}$ s.t. $M_t(\tilde{e}, \tilde{c}) = \bar{y}_{2,t}$.

Proof: Necessity of $E_{2,ch} \subseteq E_{2,t}$

- AFSOC $\exists \tilde{e} \in E_{2,ch}$ s. t. $\tilde{e} \notin E_{2,t}$.
- This implies $\nexists \tilde{c}$ s.t. $M_t(\tilde{e}, \tilde{c}) = \bar{y}_{2,t}$.
- \implies parameter ID fails at correction step one (first failure condition).

Proof: Necessity of $E_{1,ch} \times C'_{2,t} \subseteq \Theta_{1,t}$

- Split into three sub conditions

$$E_{1,ch} \subseteq E_{1,t} \triangleq \text{proj}_e \text{ID}_\theta(\bar{y}_{1,t}, \bar{M}_t),$$

$$C'_{2,t} \subseteq C'_{1,t},$$

$$E_{1,ch} \times C'_{2,t} \subseteq \Theta_{1,t},$$

Sufficiency is straightforward

The argument is a simple exercise in checking that the parameters that get picked are in sets such that the correct trajectories get generated.